Imperial College
London

The Development and use of Bioinformatic Web Applications for Infectious Disease Microbiology

David Michael Aanensen
Imperial College London, Department of Infectious Disease Epidemiology

Thesis submitted for the degree of Doctor of Philosophy of Imperial College

Dedicated to my parents
Susan and Keith Aanensen

ABSTRACT

The ever-increasing generation and submission of DNA sequences, and associated biological data, to publicly available databases demands software for the analysis of the biological meanings held within. The web provides a common platform for the provision of tools enabling the concurrent deposition, visualisation and analysis of data collected by many users from many different locations. Open programmatic web standards allow the development of applications addressing diverse biological questions and, more recently, are providing methods enabling functionality more traditionally associated with desktop software to be provided via the internet.

In this work I detail the development and use of web applications addressing different, but not exclusive, areas of infectious disease microbiology. Firstly, an application utilised by a group of researchers (including myself) to undertake a comparative genetic analysis of the capsular biosynthetic locus from serotypes of the pathogen *Streptococcus pneumoniae* is detailed. Secondly, an application widely used by communities of researchers and public health laboratories for the assignment of microbial isolates to strains via the internet: mlst.net is described. Thirdly, I describe a generic electronic taxonomy application for assigning strains to bacterial species, exemplified using sequences from the viridans group streptococci, the taxonomy of which is notoriously difficult to define. Lastly, I describe the use of web mapping tools for molecular epidemiological databases, such as mlst.net, and further detail their application to the European distribution of genotypes of *Staphylococcus aureus*, to catalogue the global distribution of the emerging amphibian fungal pathogen *Batrachochytrium dendrobatidis* and to provide geocoding tools to encourage users to submit and explore their own data in the geographical context they were collected.

Each application has been designed with generality in mind and through reference to user workflows and biological examples, I demonstrate the extensibility and general applicability of current web development methodologies to enable the provision of applications addressing a diversity of biological questions.

ABSTRACT 3
TABLE OF CONTENTS 4
LIST OF TABLES AND FIGURES 8
TEXT CONVENTIONS 11
ACKNOWLEDGEMENTS 12

Chapter 1: General Introduction 13

1.1 Hypertext Markup Language (HTML) 15
1.2 Cascading Style Sheets (CSS) 16
1.3 Simple Web Framework 17
1.4 Dynamic Web Framework: Server-side Scripting 18
1.5 The Document Object Model (DOM) 21
1.6 JavaScript 22
1.7 Combining JavaScript, CSS and the DOM 22
1.8 Rich Internet Applications: Web 2.0, AJAX and APIs 24
1.9 AJAX Web Framework 24
1.10 Application Programming Interfaces (APIs) 26
1.11 Usability 28

Chapter 2: *Streptococcus pneumoniae* Capsular Polysaccharide (CPS) Web Application 30

2.1 Introduction 31

2.1.1 Capsule biosynthesis in *Streptococcus pneumoniae* 34
2.1.2 Wzx/Wzy-dependant biosynthesis 37
2.1.3 Serotypes 3 and 37: Synthase-dependant pathway 40
2.1.4 Sequencing of 90 *S.pneumoniae cps* loci 41

2.2 Materials and Methods 43

2.2.1 Data handling 43
2.2.2 TribeMCL 43
2.2.3 Gene naming 44
2.2.4 Encoding of raw data 45
2.2.4.1 DNA sequences 45
2.2.4.1.1 Graphical encoding of *cps* loci 45
2.2.4.1.2 *In silico* storage of sequences 45
2.2.4.2 Biochemical repeat-unit structures 46
2.2.4.2.1 Graphical encoding of CPS structures 46
2.2.4.2.2 *In silico* encoding of CPS structures 47
2.2.4.3 Combining sequence and structure representations 47
2.2.4.4 Serology 48
2.2.5 Sequence alignment methods 49
2.2.5.1 Basic Local Alignment Tool (BLAST) 49
2.2.5.2 Clustal 49
2.2.5.3 Needle 50
2.2.6 CPS Web Application 51
2.2.6.1 Database schema 51
2.2.6.2 Presentation and application logic 53
2.2.6.3 Database 53
2.2.6.3.1 Database querying 53

2.2.6.3.2 Homology group querying 55
2.2.6.4 WebACT 56
2.2.6.5 Serotypes 58
2.2.6.6 Structures 59
2.2.6.7 Statistics 60
2.2.7 Serotype clustering 60
2.2.7.1 Dendrogram based on homology group sharing 61
2.2.7.2 Dendrogram incorporating the degree of sequence similarity 62
2.2.8 Interrogating the dendrogram within the CPS Web Application 63
2.2.9 Demonstration of workflow for analysing clusters of serotypes 65

2.3 Results 67

2.3.1 General features of the *dexB-aliA* locus from 90 *S.pneumoniae* serotypes 67
2.3.2 Initial transferases and Wzy linkages 68
2.3.3 *cps* locus clustering 68
2.3.4 Serotypes 44 and 46 are related to serogroup 12 71
2.3.5 Serogroup 11 74

2.4 Discussion 76

2.5 Further work 81

Chapter 3: The Multilocus Sequence Typing Network: mlst.net 84

3.1 Introduction 85

3.2 mlst.net v1 88

3.2.1 mlst.net database schema 89
3.2.2 Curator's interface 90
3.2.3 User's interface 90
3.2.4 Clustering MLST data 94
3.2.4.1 Comparing query strains to the database: Allelic profiles 94
3.2.4.2 Comparing query strains to the database: eBURST 95
3.2.4.3 Comparing query strains to the database: Concatenated sequences 97

3.3 mlst.net v2 98

3.3.1 Interface re-development 99
3.3.2 Integration of Mapping tools for viewing MLST databases 103
3.3.3 Integration of comparative and geographical eBURST 109
3.3.3.1 Geographical eBURST comparisons 111
3.3.3.2 Database query eBURST comparisons 112
3.3.3.3 eBURST menu option 113
3.3.4 Additional mlst.net v2 features 113
3.3.4.1 Publications 113
3.3.4.2 Download data 113
3.3.4.3 Organism-specific tools 114

3.4 Discussion 116

Chapter 4: Assigning Strains to Bacterial Species via the Internet: eMLSA.net 119

4.1 Introduction 120

4.2 Sequencing and cluster analysis 122
4.3 Cluster and species concordance 126
4.4 Using MLSA to identify new species clusters 128
4.5 Single gene trees and identification of interspecies imports 128

4.6 Development of a generic electronic taxonomy web application 129

4.6.1 Method of tree production 129
4.6.2 Method of tree visualisation 132
4.6.3 eMLSA.net database schema 134
4.6.4 Definition of application logic 135

4.7 Walkthrough of eMLSA.net 136

4.7.1 Portal homepage 136
4.7.2 Viridans group sub-site: viridans.eMLSA.net 138
4.7.3 Querying the database with sequences 139
4.7.4 Viewing the species tree: The tree view 140
4.7.5 Collapsing nodes on the tree 143
4.7.6 Viewing individual locus species assignment and trees: The locus view 144
4.7.7 Other options 146

4.8 Discussion 148

Chapter 5: Spatialepidemiology.net: Integrating Web Mapping Tools with Molecular Epidemiological Databases 151

5.1 Introduction 152

5.1.1 Dynamic Mapping within Web Applications 154
5.1.2 Google Maps 154
5.1.3 Google Earth 158
5.1.4 JavaScript libraries 161
5.1.4.1 Prototype 161
5.1.4.2 Script.aculo.us 161
5.1.4.3 Mapstraction 161

5.2 Spatialepidemiology.net 163

5.2.1 Multilocus Sequence Typing Maps: MLST-Maps 164

5.2.1.1 Introduction 164
5.2.1.2 Google Map interface 166
5.2.1.3 Google Earth interface 168

5.2.2 Staphylococcal Reference Laboratory Maps: SRL-Maps 171

5.2.2.1 Introduction 171

5.2.2.2 Data summary 174
5.2.2.3 SRL-Maps Web Application 175
5.2.2.4 Isolate database 175
5.2.2.5 *Spa* type Google Map 176
5.2.2.6 *Spa* type Google Earth 179

5.2.3 *Batrachochytrium dendrobatidis* Maps: *Bd*-Maps 181

5.2.3.1 Introduction 181
5.2.3.1 Data summary 183
5.2.3.2 Handling large datasets using Google Maps 184
5.2.3.3 *Bd*-Maps Web Application 186
5.2.3.4 *Bd* Google Map 187
5.2.3.5 Surveillance 192
5.2.3.6 Other options 194

5.2.4 Geocoding and generating user maps: User-Maps 196

5.2.4.1 Introduction 196
5.2.4.2 User-Maps Web Application 196
5.2.4.3 Map latitude/longitude lookup 196
5.2.4.4 Batch geocoding 197
5.2.4.5 Simple map creation tool 198
5.2.4.6 Advanced Map creation tool 198

5.3 Discussion 199

Chapter 6: General Discussion 204

Bibliography 209
Publications 209
Web references 221

APPENDIX 1: Polysaccharide Biochemistry 224
APPENDIX 2: Relevant Publications 227

LIST OF TABLES AND FIGURES

Table 3.1 Species-specific databases hosted at mlst.net 88

Table 4.1 Loci used for viridans group MLSA scheme 123

Figure 1.1 Simple web server framework 17

Figure 1.2 Dynamic web server framework 18

Figure 1.3 Three-tiered Web Application architecture 19

Figure 1.4 AJAX Web Application architecture 25

Figure 2.1 Makeup of a generic *cps* locus 36

Figure 2.2 Schematic representation of the biosynthesis of CPS by the Wzx/Wzy-dependant pathway 39

Figure 2.3 Graphical representations of the CPS repeat-unit structure and *cps* locus of serotype 9A 48

Figure 2.4 CPS Application database schema 52

Figure 2.5 Database query form 54

Figure 2.6 Database search results including subsequent clustalW alignment and dendrogram visualised using ATV 55

Figure 2.7 Homology group query results 56

Figure 2.8 A three-way ACT comparison: Serotypes 18A vs 18B vs 18C 57

Figure 2.9 Serotype summary detailing serotype 6B 59

Figure 2.10 Clustering of serotype 14 and serogroup 15 based on gene content and sequence-weighted gene content 62

Figure 2.11 Dendrogram of sub-cluster containing serogroup 11, viewed within the CPS application 63

Figure 2.12 ACT comparison based on dendrogram interrogation: Serotypes 11B vs 11C vs 11A vs 11D vs 11F 64

Figure 2.13 Dendrogram based on proportion of shared HGs, adjusted for sequence similarities between the shared HGs 70

Figure 2.14 Comparison of those *cps* loci within cluster 2 assigned to the same subcluster: Serogroup 12 and serotypes 44 and 46 73

Figure 2.15 All serotypes within serogroup 11 fall into cluster 6 75

Figure 3.1 mlst.net v1 database schema 89

Figure 3.2 MLST Workflow: mlst.net v1 92

Figure 3.3 Simple XML format used for Batch Strain Query 93

Figure 3.4 eBURST population snapshots 96

Figure 3.5 Application Logic: mlst.net v1 (page-click-page) versus mlst.net v2 (AJAX) 101

Figure 3.6 mlst.net v2: Initial display 104

Figure 3.7 Sequence Types from Spain displayed within mlst.net v2 105
Figure 3.8 Global distribution of *S.pneumoniae* Sequence Type 81 isolates 106
Figure 3.9 Database query results: Epidemiological view 107
Figure 3.10 Database query results: Map-based view 108
Figure 3.11 Comparative eBURST snapshot: Demonstration 110
Figure 3.12 Comparative eBURST snapshot: Spain vs. South Africa 111
Figure 3.13 Comparative eBURST snapshot: Serotype 23F vs. Serotype 19F 112
Figure 3.14 Is it a pneumococcus? mlst.net v2 115
Figure 4.1 Neighbor-joining tree indicating position of type strains within sequence clusters 125
Figure 4.2 Re-labelling of the neighbour-joining tree from Figure 4.1 127
Figure 4.3 eMLSA.net database schema 134
Figure 4.4 Application logic: eMLSA.net 135
Figure 4.5 eMLSA.net: homepage 137
Figure 4.6 Viridans.emlsa.net: homepage 138
Figure 4.7 Demonstration locus sequences for viridans.emlsa.net 139
Figure 4.8 Sequence clusters and species assignment for viridans.mlst.net demonstration sequences 140
Figure 4.9 Circular tree-type display 142
Figure 4.10 Rectangular tree type display and zooming in to query sequence 143
Figure 4.11 Collapsing the Mitis group branch of the tree to display sub-clustering within the Anginosus and Salivarius groups 144
Figure 4.12 viridans.emlsa.net: The locus view 145
Figure 4.13 Location of foreign *pyk* query sequence within single locus tree 146
Figure 4.14 viridans.emlsa.net: Database query view 147
Figure 4.15 viridans.emlsa.net: Download data view 144
Figure 5.1 Simple Google Maps API example 156
Figure 5.2 Displaying a marker and associated Info Window using the Google Maps API 157
Figure 5.3 The 'Whole Earth browser' Google Earth 159
Figure 5.4 Simple KML example displayed in Google Earth 160
Figure 5.5 Simple map example displayed using Mapstraction and Yahoo Maps 162
Figure 5.6 A single webpage utilising multiple web mapping APIs 163
Figure 5.7 MLST-Maps: Database access page 165
Figure 5.8 The *Staphylococcus aureus* MLST database viewed using Google Maps at MLST-Maps 167

Figure 5.9 Global distribution of *Staphylococcus aureus* Sequence Type 239 and eBURST population snapshot 168

Figure 5.10 *Streptococcus pyogenes* MLST database viewed within Google Earth 169

Figure 5.11 *Streptococcus pyogenes* Sequence Types from the Czech Republic launched from Google Earth and subsequent eBURST snapshot 170

Figure 5.12 SRL-Maps: European distribution of Staphylococcal reference laboratories 176

Figure 5.13 Detailed view of laboratory HU026, Szekszárd, Hungary 177

Figure 5.14 European distribution of *spa* type t008 focussed on SRL HU026 178

Figure 5.15 European distribution of *spa* type t041 focussed on SRL HU026 179

Figure 5.16 European distribution of Staphylococcal reference laboratories viewed using Google Earth 180

Figure 5.17 *Bd*-Maps: Homepage 187

Figure 5.18 *Bd*-Maps: Global thematic overview 188

Figure 5.19 Country centroid markers 189

Figure 5.20 USA State (excepting Alaska and Hawaii) centroid markers 190

Figure 5.21 Zooming the map to view site sample sites in Nevada, USA 191

Figure 5.22 IUCN Red List page for *Rana onca* 192

Figure 5.23 *Bd*-Maps: Surveillance summary 193

Figure 5.24 Surveillance summary for Spain 194

Figure 5.25 Recent *Bd* articles indexed in PubMed, derived from Hubmed.org 195

TEXT CONVENTIONS

Print publications or book chapters listed in the 'Publications' section of the Bibliography are referenced in the text by first author (or first and second if only two authors) and year of publication.

Web references are sequentially numbered and are referred to in the text in the format [W*n*] where 'W' denotes the reference is a website (or is solely available on the web) and '*n*' is the number as listed in the 'Web references' sections of the Bibliography. Web addresses are only included directly within the text if they form part of the developments detailed in this thesis.

The text includes a very small number of example blocks of code. These include line numbers and appear in `'Courier 10-point font'`. If a specific function or part of the code block is referenced to in the text, it too appears in `'Courier 10-point font'`.

ACKNOWLEDGEMENTS

First and foremost I would like to thank Prof. Brian Spratt for giving me the opportunity, guidance and freedom to complete this thesis. I am privileged to find myself in the laboratory of someone with great integrity, humour and foresight and in an environment that is stimulating, enjoyable and challenging.

For the CPS Chapter: Special thanks go to Stephen Bentley for the good times in and out of work. Prof. Peter Reeves for the stimulating conversations, the pints in Blackheath and the opportunity to share in his enthusiasm for polysaccharide genetics, Angeliki Mavroidi (we got there in the end) and James Abbott. For the MLST chapter: Thanks must go to all the MLST curators and data submitters and in particular Dan Godoy, Cynthia Bishop, Rob Willems, Frank Odds, Angela Brueggemann, Karen McGregor, Debra Besson, Jon Thomas, Adrian Whatmore, Gabi Margos and Marie-Elisabeth Bougnoux. For the eMLSA chapter: Cynthia Bishop (again), Mogens Kilian, Bill Hanage and Greg Jordan for thinking along the same lines. For spatialepidemiology.net: Thanks to Prof. Hajo Grundmann for his inexhaustible enthusiasm and all the Staphyloccoccal reference laboratory and EARSS representatives for making the ever more frequent meetings welcoming and enjoyable. In particular, Helena Zemlicková, Jos Monen, Leo Schouls and Carola Schinkel. For *Bd*: Dede Olson, Mat Fisher, Susan Walker and Trent Garner.

To all the friends and colleagues within and outside of the Department of Infectious Disease Epidemiology who have provided a great environment in which to work. In particular, Déirdre Hollingsworth, who always seems to have the right words, Artemis Koukounari for knowing how it feels, Bill Hanage (again), Tom Churcher, Nicole Bilek, Peter White, Andrea Johnson, Derek Huntley and Ed Feil (for many reasons).

For their support and inspiration: thanks to my parents and family, Claire, Kevin and Joe, Mark, Nan and both Albert's Aanensen. To my friends Scott Wallman, Stuart Strickson, Jon and Katy Kirkup, Sam and Hugh Elghali-Evans, Jane Wheeler, Doug Doubleday, Klara Hanincová, Ricardo Dewey and Jennifer Leiter. Special thanks go to a great friend, Paul Vinter and Barry St Michael.

To the memories of Grandad Aanensen, Granny McCullagh,
Stephen Maidment, Klaus Kurtenbach and Steven Dunstan.

Chapter 1: General Introduction

The determination and interpretation of the sequence of nucleotides in a molecule of DNA is fundamental to many areas of biology. The sequencing of single genes, regions of chromosomes that form functional units or operons and the full genome sequences of eukaryotic and prokaryotic organisms is increasingly common, and increasingly easy to achieve. The raw data generated from such projects are basic (consisting of varying numbers of the letters A, C, T and G, each representing a particular nucleotide base) and can be stored and manipulated *in silico*, however, the complex biological meanings inherent in the order of bases belie this underlying simplicity. The development and use of data storage methodologies, computational algorithms and software, statistical techniques and theory aimed at understanding the meaning held within DNA sequences falls within the remit of bioinformatics.

Sequence databases provided by the International Nucleotide Sequence Database Collaboration (INSDC [W1]) include the EMBL Nucleotide sequence database (Emmert et al., 1994), GenBank (Benson et al., 1994) and the DNA Database of Japan (DDBJ) (Tateno et al., 1998). Each mirrors data from the others and, collectively, they provide warehousing for the vast amount of sequence data generated by projects undertaken globally. Since its inception in 1982 the growth in sequence submissions to GenBank has been exponential and now contains over 95 billion nucleotide bases from finished sequencing projects (Benson et al., 2009). The sequences represent over 300,000 species named at the genus level or lower and the rate of data deposition is expected to dramatically increase over the coming years as sequencing technologies continue to evolve.

The INSDC databases provide invaluable resources for the deposition and retrieval of nucleotide sequences and associated bibliographic and biological annotation. Furthermore, each provide methods which allow the detection of similarities between a query sequence and database sequences through interfaces to tools such as the Basic local alignment search tool (BLAST; Altschul et al., 1997).

Whilst sequence similarity searching is a fundamental process in many areas of bioinformatics, the INSDC databases are not designed specifically for the analysis of

discrete systems or to address any particular biological question. They act primarily as a centralised repository whereby annotated sequence data can be searched, extracted and utilised for *ad hoc* analysis. Consequently, a myriad of secondary database systems, often utilising sequence information held in the primary warehouses, have appeared whereby sequence data are post-processed and combined with other biological data sources to address particular questions. For example, the Pfam database (Sonnhammer et al., 1997) contains annotated protein domain families discovered through a combination of experimental evidence and homology with sequences in the primary databases. The scale and scope of these secondary systems cover diverse areas of *in silico* biology, from genomic databases containing intensively curated metabolic pathway and regulatory network details encompassing the cell biology of organisms such as *Escherichia coli* (EcoCyc; Keseler et al., 2005) and *Pseudomonas aeruginosa* (PseudoCyc; Romero and Carp, 2003) to population databases such as the HapMap project which contains data on human genomic variation with the aim of finding genes that affect health, disease and individual responses to medications and environmental factors (Thorisson et al., 2005).

Secondary databases are an essential resource for both computational and experimental biologists and provide specialist information far beyond the scope of the primary data warehouses. Currently, the online Molecular Biology Database Collection, which is a 'database of databases' provided by the journal Nucleic Acids Research, lists over 1170 secondary systems covering a wide spectrum of research topics (Galperin and Cochrane, 2009). Functionally, each one provides differing sets of tools for biological data storage, submission, access, analysis and visualisation. The choice of tools employed and the way in which they are linked together is driven by the questions researchers may have. However, in common to all is access via the internet.

The ubiquitous nature, and ease of access to the internet provides a common platform for software addressing diverse biological questions. Bioinformatic web developers are not restricted by a user's operating system (e.g. Windows, Macintosh or Linux) and furthermore have access to an increasing number of development methodologies allowing the provision of functionality more traditionally associated with desktop software. The same basic development principles, when combined with relevant data, algorithms and visualisations have provided all of the developments mentioned above. Furthermore, the distributed and concurrent use of such tools by diverse research

communities has helped to advance our understanding of the biology inherent within sequence data. These commonly used principles of web development will now be discussed further.

1.1 Hypertext Markup Language (HTML)

The basic language of the web is hypertext markup language (HTML), which provides a means to describe the structure of text-based information in a document by denoting certain text as links, images, paragraphs, lists etc. HTML is written in plain text, which includes a series of "tags" that are surrounded by angle brackets. Web browsers load and interpret HTML documents, displaying elements in the correct format. The basic HTML description for a webpage is as follows:

```
1 <HTML>
2     <HEAD>
3           <TITLE>Example</TITLE>
4     </HEAD>
5     <BODY>
6           <p>Hello example</p>
7     </BODY>
8 </HTML>
```

Every tag is opened and closed and the whole document can be viewed as a tree. The HEAD of a HTML document (Lines 2-4) contains Meta information about the document and the BODY (Lines 5-7) contains the text to be displayed. The HEAD and BODY are both wrapped within a HTML Tag (Lines 1 and 8). When loaded into a web browser, the above HTML would display as:

Hello example

The opening , and closing tags within the BODY of the document denote that the text within forms a paragraph.

Markup, through the use of 'attributes' within an opening tag, can also be used to add style to a document, for example:

```
<p font color="green">Hello example</p>
```

would render in a web browser as:

Hello example

In the early days of the web, a HTML page was static and once displayed was impossible to change without loading a different page. To address the need for more control over elements within a HTML document, the World Wide Web Consortium (W3C) [W2] introduced a series of standards. The W3C is a body that defines open programmatic specifications for web development.

1.2 Cascading Style Sheets (CSS)

The CSS specification was introduced to enable the separation of document content from document presentation [W3]. In our initial simple webpage example, if the HTML were expanded to include five paragraphs, each of which was to be rendered green, every instance of the <p> tag would need to include a 'font color' attribute. CSS allows a single style to be defined for the <p> tag, defined in the HEAD of a document, which when loaded in the browser is implemented (or cascaded) throughout the document. For example, the following HTML:

```
1 <HTML>
2     <HEAD>
3           <TITLE>Example</TITLE>
4                 <STYLE>
5                 p {color: green;}
6                 </STYLE>
7     </HEAD>
8     <BODY>
9           <p>Hello example</p>
10          <p>Hello example1</p>
11          <p>Hello example2</p>
12    </BODY>
13 </HTML>
```

would render in a web browser as:

Hello example

Hello example1

Hello example2

The CSS specification can be used to define all aspects of presentation logic (the look and formatting). The complete set of specifications for a HTML document can be stored in a separate CSS file, which is then referenced within a HTML document as follows:

```
1 <HTML>
2     <HEAD>
3           <TITLE>Example</TITLE>
4           <link rel="stylesheet" href="styles.css" />
5     </HEAD>
6     <BODY>
7           <p>Hello example</p>
8           <p>Hello example1</p>
9           <p>Hello example2</p>
10    </BODY>
11 </HTML>
```

In this way, multiple documents can link to the same stylesheet and the entire presentation logic for a website (a collection of related web pages) can be defined in one place. This not only limits the amount of markup within a single document but also means that the entire look and feel of a website can be changed simply by changing the stylesheet.

1.3 Simple Web Framework

Using HTML and CSS alone allows classic hypertext navigation within static websites, i.e. those that consist entirely of a set of related web pages that do not change when requested by the browser. The architecture for such websites can be seen in Figure 1.1.

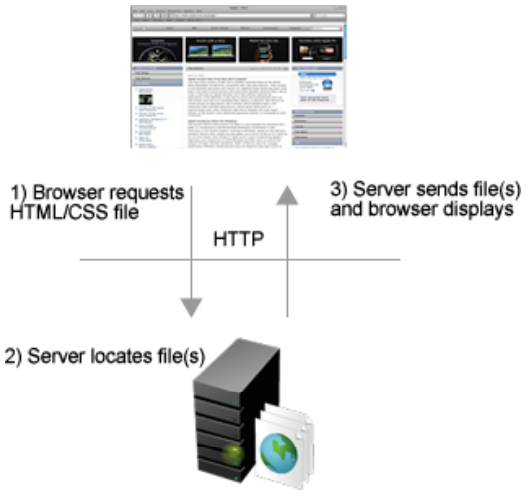


Figure 1.1 Simple web server framework.

When a user enters a web address (a Uniform Resource Locator or URL) into their browser (referred to as a client), a request is directed to the relevant web server via the hypertext transfer protocol (HTTP) [W4] (Figure 1.1(1)) and the default HTML document for that web address is located by the server (Figure 1.1(2)) and sent back to the browser for display, again via HTTP (Figure 1.1(3)). Navigation of a website is

achieved by clicking on links within the webpage. Links can call further HTML documents from the same site or link to documents on other web servers.

In the early days of the web these sites were pervasive (and still are) as they are easy and cheap to develop, and generally suffice for most people's requirements (e.g. simple homepages, simple business presence etc.). However, they offer little functionality and interactivity and content often becomes stagnant. As the web grew, and in particular, as businesses recognised a possible revenue stream, technologies were developed to allow webpages to become more interactive, linking to ever changing data sources and allowing content to become dynamic. This was achieved through the introduction of server-side scripting.

1.4 Dynamic Web Framework: Server-side Scripting

Server-side scripting is a web server technology in which a user's request (Figure 1.1(1)) is fulfilled by running a script directly on the web server to generate dynamic HTML pages. The server architecture for a website using server side scripting can be seen in Figure 1.2.

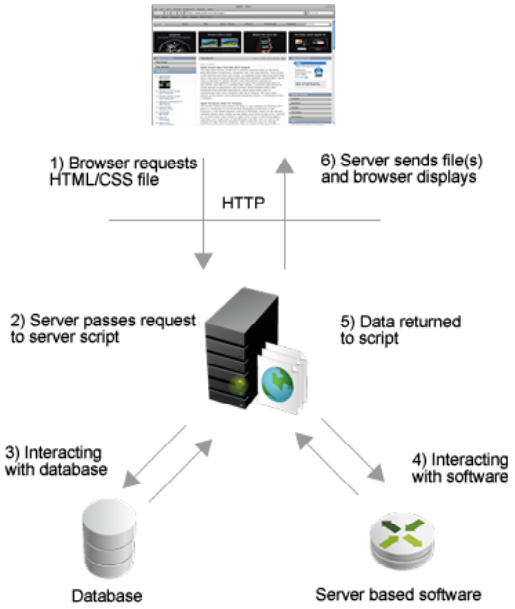


Figure 1.2 Dynamic web server framework.

Server side scripting is often used to interact with databases to, for example, return the results from a form-based request from a HTML page in the browser (Figure 1.2(3)). Other scripting examples include, sending email directly from a web server or using

the user's request to interact with server based software (Figure 1.2(4)). Within bioinformatics, many servers run ClustalW as a web service [W5]. ClustalW is a stand-alone software package installed on the web server and the request from the browser (a set of sequences) is passed, using server-side scripting, to ClustalW and the results are then formatted and passed back to the browser as HTML.

Many scripting languages exist and common examples include PHP, Perl, ASP or ASP.NET, JSP, Ruby etc. The choice of language is generally dependent on the developer's preference and the type of web server used (e.g. Microsoft IIS [W6], Apache [W7] etc.) but most languages provide similar functionality.

Server-side scripting allows application logic to be defined on the server. 'Application Logic' (often also termed business logic) is a non-technical term generally used to describe the functional algorithms that handle information exchange between a database/server software and a user interface. The application logic is translated into code within server-side scripts and defines the functionality inherent within a dynamic web site like that illustrated in Figure 1.2.

In software engineering this type of dynamic web server is commonly referred to as having a 'three-tiered architecture' and describes a client-server application as opposed to a single-tiered architecture as exemplified by most stand-alone, installable pieces of software which reside solely on the client machine. The three tiers are 1) presentation, 2) application processing and 3) data management, each of which are logically separate (and can be located in different locations). We can now overlay these concepts onto our dynamic web server figure as in Figure 1.3.

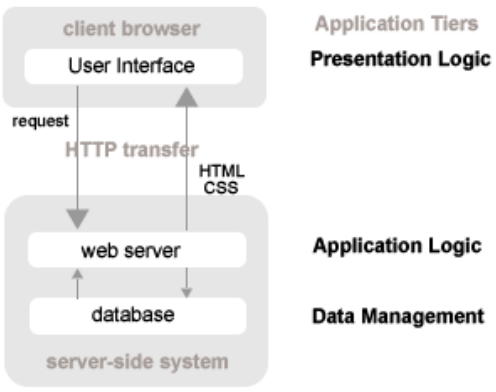


Figure 1.3 Three-tiered Web Application architecture. Adapted from [W15].

Presentation logic is handled by CSS, which is interpreted by the client and cascaded through the HTML file passed from the server. Following a client request (e.g. submitting a form), the Application Logic, coded into the server-side scripts, run on the server and interact with the database, formatting the results into HTML which are passed by the web server back to the client via HTTP.

The vast majority of dynamic websites that currently exist adhere to this three-tiered architecture. For example, a common bioinformatics task is to search for similarities to a nucleotide or protein sequence in a public database, such as GenBank, using the local alignment program BLAST.

The European Bioinformatics Institute (EBI) has a web server running BLAST (WU-BLAST2) [W8]. When entering the web address into a browser, a request is sent to the EBI web server (via HTTP), which returns, to the client (your browser) the relevant HTML that contains all the tags that define the form fields, images, text etc. that you can see. A CSS file is linked to this HTML document (as described) which defines the overall layout and style of the page. When pasting in a DNA/protein sequence and clicking 'Run BLAST', the client sends the sequence you have entered (along with the parameters selected within the form) back to the EBI web server. On the server, scripts receive this data (in this case written in Perl) and perform a BLAST search against the database of public sequences. An intermediate HTML page is displayed telling you that the job is processing (handled by the server-side script), and when complete the results are formatted on the server, into HTML, and sent back to your browser, again via HTTP. The same CSS file that formatted the initial HTML page is used for defining the presentation of the results. Links on the page allow you to delve deeper into the results, again with requests and responses being handled by the server-side scripts.

The process of using websites such as this one can be referred to as 'page-click-page' websites. Navigation is achieved by loading an initial page (which is static once loaded), clicking (a button on a form, a link etc.) to send a request to the server, which sends back a new page (again, static once loaded) and so on.

Most web content is currently of this 'page-click-page' variety. All elements defined in a HTML document when loaded into the browser are displayed in the order that

they are written. To change the order, or to add additional elements, further processing by the server is required.

Recently, a shift in web development philosophy/practices is providing techniques that allow HTML pages, once loaded from the server, to be more dynamic and interactive, often including computational processes undertaken within the browser, that traditionally can only be undertaken by server-side scripting. These new developments build on two additional parts of the HTML development process that warrant further explanation: the Document Object Model and JavaScript.

1.5 The Document Object Model (DOM)

The W3C definition for the DOM is as follows [W9]:

"The DOM is a platform and language-neutral interface that will allow programs and scripts to dynamically access and update the content, structure and style of HTML and XML documents. The document can be further processed and the results of that processing can be incorporated back into the presented page."

XML stands for eXtensible Markup Language and is a general-purpose specification for creating custom markup languages. In HTML, the W3C specification defines what the tags are (which are adhered to within web browsers). With XML, a user can define the TAGS for their own purposes [W10].

The DOM provides an object-oriented representation of the individual elements and content in a web page, along with methods for updating, moving and removing them. In our initial example of a HTML document, the tree like structure contains two main parts, the HEAD and the BODY. The BODY contains those elements that are displayed within the browser and these are the sections that are open to manipulation using the DOM. This manipulation requires scripting to be undertaken directly within the client rather than on the server. All currently available web browsers contain a parser that interprets HTML and converts it into the DOM. Accessing and manipulating the DOM is undertaken using client-side JavaScript.

1.6 JavaScript

JavaScript (officially named ECMAScript-262 [W11]) was the first scripting language to be widely implemented in web browsers. For many years JavaScript was a very lightweight language, which was used to add superficial interactivity to a webpage. A simple example being the 'rollover' a term used to describe any image (static or a button) which when the mouse pointer is 'rolled' over, changes to another image. Similarly to CSS, JavaScript code is placed in the HEAD of a HTML document and can also be placed in an external file which is loaded once (into the browsers JavaScript engine) and available to all pages within a website.

Using the example of the rollover, the action is achieved through a call to a function, placed on the initial image, which when triggered (by the mouse pointer rolling over the image) calls the function, which causes the image to change. The JavaScript source code is interpreted and executed by the browser's JavaScript engine.

Because JavaScript code is executed locally, in the client browser, it can respond to a user's action quickly, making an application feel more responsive, and furthermore, JavaScript can detect user actions that HTML cannot, such as individual keystrokes.

1.7 Combining JavaScript, CSS and the DOM

Returning to our initial HTML document, we can now combine the HTML and CSS with some JavaScript that performs a simple manipulation on the 'Hello example' text:

```
1 <HTML>
2     <HEAD>
3           <TITLE>Example</TITLE>
4           <STYLE>
5           #egText {color: green;}
6           </STYLE>
7           <SCRIPT LANGUAGE="JavaScript">
8           function changeColor(color) {
9           document.getElementById('egText').style.color = color;
10          }
11          </SCRIPT>
12    </HEAD>
13    <BODY>
14          <DIV ID="egText">Hello example</DIV>
15          <A HREF="JavaScript:changeColor('orange')">Orange</A>
16          <A HREF="JavaScript:changeColor('blue')">Blue</A>
17    </BODY>
18 </HTML>
```

Loaded into a browser this would display as:

Hello example

Orange Blue

The text ‘Hello example’, rather than in a paragraph tag as previously, is now in a <DIV> tag (Line 14). The DIV (short for Division) is most often used within DOM-oriented pages to define areas to be amended. Each individual DIV element is given an ID, in this case ‘`egText`’ which is used both to define the CSS style (Line 5), and also as an identifier within the JavaScript function (Line 9). Below the DIV tag are two <A> tags (Lines 15 and 16) - <A> is the HTML ‘Anchor’ tag used to define hyperlinks and the HREF attribute contains the destination of the link. In this case the destination is a JavaScript function called ‘`changeColor`’ (Line 8) and contains one variable (in the first <A> tag ‘orange’). When this link is clicked, the `changeColor()` function is called and, using the DOM the element ‘egText’ is changed to the specified colour (in this case orange). Similarly, the ‘Blue’ link will change the ‘Hello Example’ text to blue.

As with our initial example, CSS and JavaScript can be located in separate documents:

```
1 <HTML>
2     <HEAD>
3           <TITLE>Example</TITLE>
4           <link rel="stylesheet" href="styles.css" />
5           <script type="text/JavaScript" src="scripts.js"></script>
6     </HEAD>
7     <BODY>
8           <DIV ID="egText">Hello example</DIV>
9           <A HREF="JavaScript:changeColor('orange')">Orange</A>
10          <A HREF="JavaScript:changeColor('blue')">Blue</A>
11    </BODY>
12 </HTML>
```

Over the past 2-3 years, JavaScript has matured into a fully formed programming language that contains standard data structures for data manipulation and can be used to achieve tasks previously only possible using server-side scripting languages. These simple principles, a DOM that gives access to manipulation of all parts of a document, CSS for presentation logic and JavaScript for defining complex functions within the client, in conjunction with server-side scripting, have become web standards responsible for what are known as rich internet applications.

1.8 Rich Internet Applications: Web 2.0, AJAX and APIs

The last few years have seen a shift in web development practices bringing together a group of ideas and technologies collectively under the name ‘Web 2.0’ [W12]. The concept of Web 2.0, like many others, doesn’t have a hard boundary but can be summarised as a shift towards using the web as a platform. This trend is being driven by a set of technologies which allow the provision of rich internet applications, executed within a browser, that provide functions/interactivity normally associated with traditional desktop applications.

1.9 AJAX Web Framework

Our basic web site above illustrated a page-click-page interface. Such a site initially loads an individual page containing text, images, forms etc. and each page must be dealt with before navigating to the next. After submitting a form, for example, one waits until the browser executes the server request and refreshes the window with the new page. Certain patterns of behaviour become ingrained when interacting with such sites: ‘only click submit once’, ‘do not click the back button when submitting a form’ etc. Such sites suffer from certain issues. Firstly, there is a significant delay while each new page is loaded which interrupts the ‘flow’ of the application. Secondly, an entire page is loaded on each occasion, often returning substantial portions of HTML that were loaded with the initial page (for example, menus, images etc.) and this can amount to a significant amount of data. This unnecessary download further exacerbates the delay in loading a new page.

Such websites offer a somewhat inferior experience compared to those applications installed on a users machine. Within a traditional desktop application (e.g. Microsoft Excel [W13]), one expects the display (the contents of a spreadsheet) to remain visible and the interface elements (menus etc.) to respond to commands while the underlying processing of these requests occurs in the background without refreshing the entire display.

Asynchronous JavaScript and XML (AJAX), is a set of techniques that allows these methods to be enabled in web applications. With AJAX, web applications loaded into the client browser can retrieve data from the server asynchronously, in the background, without interfering with the display and behaviour of the existing page.

The enabling development allowing this to occur is the *XMLHttpRequest* (XHR) object [W14]. XHR is an open specification integrated within the DOM and is used within JavaScript functions to send an HTTP request directly to a web server, silently, and load the server response data (XML) directly back into JavaScript. Once the XML is loaded into JavaScript it can easily be used to manipulate the currently active HTML document in the browser, via the DOM, without the need to load a new document.

AJAX is not a new technology, rather a set of existing standards combined to provide a framework on which to build rich internet applications. The components of AJAX programming are as follows (adapted from [W15]):

- Presentation using **HTML** and **CSS**
- Dynamic display and interaction using the **DOM**
- Data interchange and manipulation using **XML**
- Asynchronous data retrieval using **XMLHttpRequest**
- **JavaScript** binding everything together

Through using this relatively simple framework a developer can progress beyond page-click-page applications towards a more desktop like web experience. Architecturally, this is achieved by adding an extra layer of application logic within the browser and means we move from a 3-tiered architecture to a *n*-tiered approach as seen in Figure 1.4.

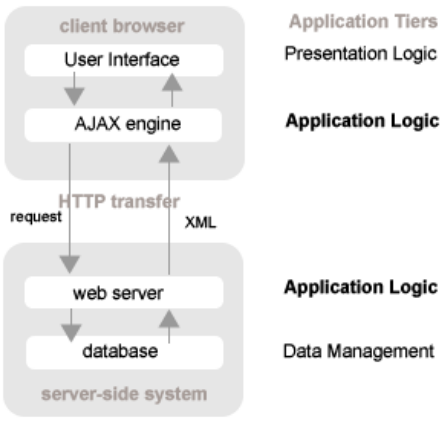


Figure 1.4 AJAX Web Application architecture.

This extra application logic layer, often referred to as an AJAX engine (to denote the part of the browser that deals with JavaScript), handles requests from the user and in the background handles server communication. Every user action that would normally generate a HTTP request takes the form of a call to a JavaScript function within the AJAX engine. Any response that does not require server interaction (e.g. simple data validation, manipulation and navigation) the AJAX engine handles on its own. If data is submitted for processing, loading additional interface code or retrieving new data, the requests are made asynchronously via XHR, usually using XML, without delaying a users interaction with the application.

One of the first examples of an AJAX application was Google Suggest [W16], a Google Search engine interface that offers the user suggestions as a query is typed in. When a user presses each key, the applications AJAX layer queries Googles server for suitably similar search phrases and presents the returned data in a drop-down box. Along with each suggested phrase, the number of results that would be expected for a search with that phrase are listed. At any point the user can select one of the suggestions instead of continuing to type. Because the server is queried with every keypress (using JavaScript to detect, and AJAX to send and return results to the server), the drop-down list updates dynamically as the user types, with no page-refreshes or similar interruptions.

1.10 Application Programming Interfaces (APIs)

In traditional software development, an Application Programming Interface (API) is a set of routines, data structures, object classes and/or protocols provided by libraries and/or operating systems in order to build a software application. Examples of traditional APIs are provided by Microsoft Windows [W17] or Cocoa for Apple's Mac OS X [W18]. The implementation of the API is the underlying functionality a programmer can access (for example Windows libraries). Using APIs, a developer can build on functionality provided by such libraries to give a quicker and richer development cycle.

The move towards the internet as a platform suggested by Web 2.0 has been facilitated by the provision of APIs for services available via HTTP. Rather than providing access to underlying operating systems, web APIs provide programmatic interfaces to a wide range of data types and functionality provided by data owners and

can be integrated by web developers to enhance their own applications. In the same way that traditional software APIs provide developers easy access to the platform they are developing on, web APIs provide web developers with easy access to large resources of data and functionality available in different locations on the web.

Probably the most well known and widely used web API is provided by Google Maps [W19]. The Google Maps API provides a set of functions, which can be used by developers to embed and interact with maps within their webpage. Google Maps is implemented as an AJAX application. When a user drags the map, a JavaScript function is called which sends data (current position and needed location) to the server asynchronously (using XHR) to retrieve the relevant images with which to update the map. All interaction with the maps is handled this way (clicking to zoom etc.) and allows a seamless, powerful user experience not possible without the principles of AJAX. Further examples of popular web APIs are provided by the Amazon API [W20], which allows the results of book searches to be easily embedded in webpages, and the Yahoo Search API (BOSS) [W21], which retrieves web searches.

The availability of web APIs and the increasing innovation in web developments, combining available data, has led to the term 'web mashups'. A mashup (a term originally used to describe the sampling of various existing pieces of music into a hybrid song in hip-hop) is a web application that combines data from one or more sources into a single integrated tool (for a comprehensive list of APIs and applications see [W22]). One of the first examples of such an application was HousingMaps [W23]. Using the API provided by CraigsList [W24] (an online-notice board for real estate listings across the USA), and the Google Maps API, HousingMaps provides a map-based interface allowing users to search and display flats/houses for rent/sale within a specified price range or by city/state. By clicking placemarks displayed on the map further details about the properties of interest can be viewed.

Within the field of biology a small number of tools using the same principles have begun appearing. iSpecies [W25], produced by Rod Page at Glasgow University, is a test of Edward O. Wilson's idea of every biological species having a webpage of its own. iSpecies provides a simple web search query, which when a species name is entered, returns a single webpage containing multiple datasets, from disparate sources, through the use of web APIs. For example, searching for *Streptococcus*

pneumoniae returns a list of Yahoo Tags (using the Yahoo search API), the initial paragraph of text from the *S.pneumoniae* text entry at Wikipedia (using the Wikipedia API) [W26], a list of genome resources from NCBI [W27], images from Google and a list of articles from Google Scholar [W28].

A further example within the Health Sciences is HealthMap (Freifeld et al., 2008 [W29]) which conglomerates a number of text-based news feeds regarding infectious disease notification cases (e.g. ProMED, WHO, EuroSurveillance and Google News) and, using text-mining algorithms, produces a map displaying the location of cases based on the type of disease caused.

The basic, open development techniques described (HTML, CSS, server-side scripting etc.), in conjunction with the emerging open standards driving Web 2.0 (AJAX and APIs), provide web developers with tools for producing complex web applications. The choice of tools and application architecture obviously needs to be carefully thought through and, as with any software implementation, the right choice of tools should be guided by the data that is being used and the careful consideration of functionality and audience usability.

1.11 Usability

Usability is often defined as a measure of the ease with which an individual can employ a particular tool in order to achieve a particular goal. Within computer science, or the study of human-computer interaction, usability refers to the elegance and clarity with which the interaction with a computer program or web site is designed. While often neglected within bioinformatics web developments, the application of basic usability principles are important when developing applications that are to be used by any number of users, of whom, very little is known. Prior to the definition of application logic, the answers to simple questions such as ‘Who are the users, what do they know/want/need to do?’ and ‘What is the background of the users and in what context are they working?’ can be usefully applied within an application development framework. Although seemingly obvious, thorough testing of applications on various computer/browser/operating systems and the careful design of pages to cater for the ‘lowest common denominator’ are often overlooked.

Web usability has become a field of its own and publications by the more vocal proponents such as Jacob Nielson's 'Designing Web Usability' (Nielson, 1999) and Jef Raskin's 'The Humane Interface' (Raskin, 2000) are regarded by many as 'gold standards' for those involved in any kind of interactive design. When designing and building bioinformatic web applications, such issues are important and the careful design of all tiers of an application should take this into account.

This thesis details the design, development and usage of bioinformatic web applications. Each chapter focuses on a specific development created to address different, but not exclusive, areas of infectious disease microbiology. All utilise the web development methodologies detailed above, however, the scope and focus of each differ in a number of ways. Chapter 2 details a development utilised in the analysis of a discrete data set (CPS Web Application). Part of a collaborative project, a small number of individuals utilised the tool to understand aspects of a discrete biological system. Acting more like an intranet, development was undertaken gradually as the data were analysed and I undertook substantial parts of this analysis as well as all the web development. Chapter 3 details stages of an application for microbial molecular typing: Multilocus Sequence Typing (MLST; mlst.net). mlst.net is widely used by research and public health laboratories and consequently development is focussed on providing the necessary tools to allow a disparate number of researchers to answer their questions. Chapter 4 details the development of eMLSA.net, a generic facility for microbial electronic taxonomy, initially demonstrated for a particular bacterial species group, but with the aim of encouraging further research communities to host datasets addressing their speciation questions. Chapter 5 details a number of developments, which introduce, through the use of web mapping, applications for the geographic visualisation and analysis of molecular epidemiological databases with the aim to encourage an uptake of such an approach for researchers who utilise laboratory techniques such as MLST.

The underlying methods utilised were somewhat dependant on the time point the project was undertaken over the previous four years. Chapter 1 utilises a three-tiered architecture, Chapter 2 is in two parts, the first details a three-tiered and the second half (and chapter 4), concurrent with AJAX and API developments, utilise *n*-tiered approaches. Chapter 5 also utilises *n*-tiered approaches but increases the involvement of APIs further.

Chapter 2: *Streptococcus pneumoniae* Capsular Polysaccharide (CPS) Web Application

Available: http://spneumoniae.mlst.net/cps/database/v3

Summary

Large-scale comparative sequencing projects are becoming more routinely feasible and demand logical data storage, visualisation and analysis frameworks. Ideally, such frameworks should allow the subdivision of data into maneagable units, which can then be further investigated in a rational fashion. In this chapter I detail the development of such a framework and its application to the DNA sequences of the capsular polysaccharide locus (*cps*) from 90 serotypes of the pathogen *Streptococcus pneumoniae*.

The *S.pneumoniae cps* sequences are known to have a complicated evolutionary history involving homologous and illegitimate recombination and are therefore difficult to interpret. Through clustering those *cps* sequences that appear to be similar as they share the greatest number of similar genes, fine-scale differences between clusters (and associated biological data) can be investigated using traditional sequence alignment methodologies.

The development of a web application and its use for the analysis of the *S.pneumoniae cps* sequences are detailed and illustrative results of the analysis are presented in the context of user workflows. Discussion of the potential merits and limitations of the technique are described and further work and application of the framework to other systems are proposed.

Note: Further analysis, utilising the system described in this chapter have been included in the following publications, copies of which can be found in Appendix B:

Bentley, S.D., Aanensen, D.M., Mavroidi, A., Saunders, D et al. (2006) Genetic analysis of the capsular biosynthetic locus from all 90 pneumococcal serotypes. *PLoS Genet* 2(**3**)e31

Aanensen, D.M*., Mavroidi, A*., Bentley, S.D., Reeves, P.R. and Spratt, B.G. (2007) Predicted functions and linkage specificities of the products of the Streptococcus pneumoniae capsular biosynthetic loci. J Bacteriol 189(**21**):7856-76

Mavroidi, A*., Aanensen, D.M*., Godoy, D., Skovsted, I.C., Kaltoft, M.S., Reeves, P.R., Bentley, S.D. and Spratt, B.G. (2007) Genetic relatedness of the Streptococcus pneumoniae capsular biosynthetic loci. *J Bacteriol* 189(**21**):7841-55

*Joint first authors

2.1 Introduction

The scale and scope of eukaryotic and prokaryotic genomic sequencing projects in particular have lead to the rapidly expanding field of comparative genomics. Currently there are over 1000 publicly available (published) bacterial genomes, representing microorganisms from a broad range of lifestyles and niches [W30]. Such data have allowed investigation of the variation in the general features and characteristics of the genomes and the details of the sequences have allowed the revelation of mechanisms by which some of these genomes have evolved. Furthermore, comparative analyses have become an important method by which functional elements can be identified, based on homologous sequence identification and the inference of homologous function.

The rapid decrease in the cost of sequencing and the concomitant improvements in technology have also made large-scale comparative sequencing projects covering large regions of chromosomal DNA from a number of strains of a particular organism more routinely feasible, allowing evolutionary questions focussed on a particular biological system to be addressed.

When faced with the prospect of analysing such large datasets, methods must be employed that subdivide the data into more manageable units which can then be further investigated, by traditional sequence alignment methods, in a rational fashion. To address overall relationships between data from different organisms, or from many examples of the same organism, an obvious approach would be to attempt to infer a phylogeny based on a measure of sequence similarity. However, whereas phylogenetic methods and the value of single gene trees has been reasonably well established, the usefulness of large, up to genome scale, phylogenetic trees is not as commonly accepted due to the observation that different parts of the genome do not necessarily have the same evolutionary history. Indeed, the influence of lateral gene transfer, recombination and duplication and loss of genes within genomic regions has brought into question the possibility of ever producing a phylogeny at the genome level (Doolittle, 1999). To address these issues, various methods have been produced that, rather than attempting to infer phylogeny, attempt to cluster species based on characteristics of complete genomes as a genome tree rather than a phylogenetic tree.

Broadly, these techniques can be grouped into one of two categories:

1) Trees based on average sequence similarity, whereby all coding and non-coding DNA is utilised for an exhaustive search of the genomes to define an overall average sequence similarity.
2) Trees based on gene repertoire, comparing the presence and absence of genes as a measure of 'gene content'. Genes are defined based on initial annotation and the measure of gene content similarity arises from defining 'gene families' based on a level of sequence similarity.

Trees based purely on sequence similarity do not take into consideration the functional assignment undertaken during careful annotation. Also, incorporation of non-coding sections of the sequences can introduce a level of noise that could lead to ambiguous groupings. Therefore, a more immediately useful and widespread approach is to construct trees based on a level of gene content (Snel et al., 2005). These trees generally rely in part on annotation and, in contrast to a strict phylogenetic approach based on sequence alignment, seek to cluster genomes together based on the number of genes in common. Such approaches have been shown to be less affected by horizontal gene transfer (Snel et al., 1999) and have been further developed to include measures of sequence similarity. For example, Konstantinidis and Tiedje (2005) have analysed bacterial genomes by first extracting all shared genes and then expressing relatedness as the average nucleotide identity between the shared genes. Differences in methodologies can have an impact on the resultant tree, for example, differences in genome size. If numbers of shared genes are used, two large genomes that share more genes in common can cluster more closely than a smaller genome that has a higher proportion of genes (but a smaller number) shared with one of the larger genomes. To overcome this, a weighting must be given based on the total size of the genomic sequences under comparison, otherwise smaller and larger genomes will tend to cluster together.

The issues inherent in producing trees for full genomes are also pertinent when investigating any multi-gene system from a number of species or from many examples of the same species that have a complex evolutionary history. However, a clustering approach does serve as a useful framework for the closer investigation of sets of such sequences, to attempt to further analyse the relationships between the sequences.

Data handling and analysis systems utilising such approaches are scarce and as sequences continue to be generated at an ever-increasing rate, the development of system level frameworks in this area becomes a necessity. Such systems should allow, based on the rational grouping of entire sequence sets (such as multiple genomes), further investigation of fine-scale similarities ideally utilising well-proven, traditional sequence comparison methodologies where possible.

The three-tiered web application architecture detailed in section 1.4 of this thesis provides the necessary flexibility to allow such systems to be produced. The data layer is represented by a relational database containing logically organised sets of data, the top level (presentation logic) is composed of the client web browser that sends user requests, via the middle tier (application logic) to the database and formats the responses for display as a webpage.

The careful design of the database, and the choice of tools utilised within the application logic should take into consideration the nature of the biological system, and should allow the user to answer their particular question in a rational fashion. The focus of this chapter is the production of a system that combines a clustering based approach to initially group more similar sequences, together with a database system that holds all of the information required for further analysis, in a logical way, and allows tools to be integrated which address the questions a researcher may have. The emphasis is on the production of a system that is scaleable and has application to a variety of data sources.

Through a collaboration with The Sanger Institute the initial dataset utilised in this project is the genomic sequences encompassing the capsular biosynthetic locus from 90 serotypes of *Streptococcus pneumoniae**. This is complemented by biochemical data relating to the structure of the polysaccharide molecules produced and immunological data based on serological cross-reactivity between the 90 different capsular polysaccharide (CPS) structures.

The capsular biosynthetic locus is known to have a complex evolutionary history (Coffey et al., 1998, 1999) and has provided an opportunity for the design of software capable of handling such a diverse dataset in a fashion that is scaleable to other

systems. Prior to presenting the analysis system the biology of *Streptococcus pneumoniae* capsule biosynthesis will be described.

*It should be noted that since this study was undertaken, a new serotype has been discovered, taking the total to 91 (Park et al., 2007). However, the sequence data were not available for use in the development and analysis undertaken and described here.

2.1.1 Capsule biosynthesis in *Streptococcus pneumoniae*

The opportunistic pathogen *S. pneumoniae* (the pneumococcus) is a normal component of the microbiota of the human upper respiratory tract and has been a model organism in molecular biology having been closely scrutinised for well over a century since its first isolation and description (Austrian, 2004).

Under normal circumstances, the pneumococcus survives as a nasopharyngeal commensal and colonization may occur soon after birth. However, disease occasionally follows colonization, ranging in severity from meningitis, septicaemia and pneumonia to sinusitis and acute otitis media, and pneumococci are a significant cause of morbidity and mortality (Cartright, 2002). It is estimated that up to 1 million children die due to pneumococcal infections each year, most of whom are young children in developing countries (WHO, 1999).

Many important bacterial pathogens, including the pneumococcus, produce a polysaccharide capsule that is believed to contribute to virulence by aiding bacterial survival in blood (Taylor and Roberts, 2005; Finn, 2004). Capsular polysaccharides (CPS) are surface-exposed and immunogenic and provide a target for the host immune response. Consequently, there has been selection for mechanisms by which encapsulated bacteria evade the host immune system. In most encapsulated species, this has been achieved by the generation of antigenic diversity, such that strains of the pathogen may express one of a number of different CPS structures (Spratt et al., 2004).

At the time of this study, 90 immunochemically distinct CPS types had been identified in *S. pneumoniae* and the structural differences are recognised by anti-sera that resolve the pneumococci into individual serotypes, or serogroups which include immunologically-related serotypes based on sera cross-reactivity. (Heidelberger M 1983). The current number of serotypes is thought to represent a high proportion of

the total diversity in the species (Henrichsen, 1995) and patterns of reactivity between the CPS and typing sera provide the basis for laboratory based serotyping.

The virulence of a particular pneumococcal serotype appears, at least in part, to be determined by the chemical make-up of the polysaccharide and the amount synthesized (although other genes undoubtedly contribute to differential virulence) (Brueggemann et al., 2003; Taylor and Roberts, 2005). Along with capsular polysaccharide, pneumococci also produce an additional polymer, the C or cell wall polysaccharide, composed in part of techoic acid, which is common to all pneumococcal serotypes. It has been known for several decades that CPS can elicit protective antibodies and vaccine development has been based on the CPS from the serotypes most commonly associated with invasive disease in children. A 7-valent vaccine composed of protein-conjugated CPS has been introduced in the USA and elsewhere. (Finn, 2004; Poolman, 2004).

The CPS of *S. pneumoniae* differ in a number of ways. Each CPS is composed of a series of repeat-units, consisting of a number of single-sugar (monosaccharide) entities linked together by glycosidic bonds. The repeat-units are linked together and are located on the cell surface of the bacteria. CPS diversity is manifest in the number and nature of monosaccharide components, the type and anomeric configuration of the linkages between monosaccharides, non-sugar composition and decorative features such as acetylation (the addition of an acetyl group to a monosaccharide), and the nature of the linkages between repeat-units (See Appendix A: polysaccharide biochemistry). At the time of study 54 of the 90 known capsule structures had been biochemically elucidated (Kammerling, 2000). Some are relatively simple in structure, such as serotypes 3 and 37, whereas others are more similar to the techoic acids and contain ribitol or phosphorylcholine (e.g. serogroup 6 and 15). In common with other bacterial polysaccharides that are located on the bacterial surface, the genes required for the synthesis of *S. pneumoniae* capsules are located in a gene cluster referred to as the capsular biosynthetic locus (*cps*).

The importance of the CPS as the protective antigen in the current, highly successful, conjugate vaccines and the need to monitor and understand vaccine escape has maintained interest in CPS structures, biosynthesis and genetics. Consequently, efforts have been directed at the sequencing of the entire *cps* locus from several serotypes, the first of which was published in 1995 from a serotype 3 isolate

(Arrecubieta et al., 1995), followed by those of serotypes 1 (Munoz et al., 1997), 19F (Morona et al., 1997) and 14 (Kolkman et al., 1997) and by 2004 the DNA sequences of 17 serotypes had been deposited in public databases.

The sequences of the published pneumococcal *cps* loci have led to, not only a description of the genes involved in polysaccharide synthesis, but also description of the mechanisms and regulation of polysaccharide synthesis, mechanisms of genetic exchange and properties related to virulence.

Excepting serotypes 3 and 37, which shall be discussed later (Section 2.1.3), the *cps* locus is located at the same position on the pneumococcal chromosome between *dexB* and *aliA*, two genes that play no part in capsule biosynthesis, and in all cases (again excepting serotypes 3 and 37) CPS are synthesized by the Wzx/Wzy dependant pathway.

Briefly, synthesis of capsule by the Wzx/Wzy pathway involves the transfer of an initial sugar onto a lipid carrier by the action of an initial sugar phosphate transferase, followed by progressive addition of sugars on the cytoplasmic face of the membrane to produce a 'repeat-unit', through the action of further glycosyltransferases (GTs) encoded by genes present in the *cps* locus. The repeat-unit is then moved across the cytoplasmic membrane by a repeat-unit transporter (Wzx), or flippase, followed by polymerisation of repeat-units by a polysaccharide polymerase (Wzy), and finally, linkage and transfer to the peptidoglycan (Sørenson et al., 1990). Despite the diversity of the polysaccharide structures, many of the sequenced *cps* loci show general commonalities (Figure 2.1).

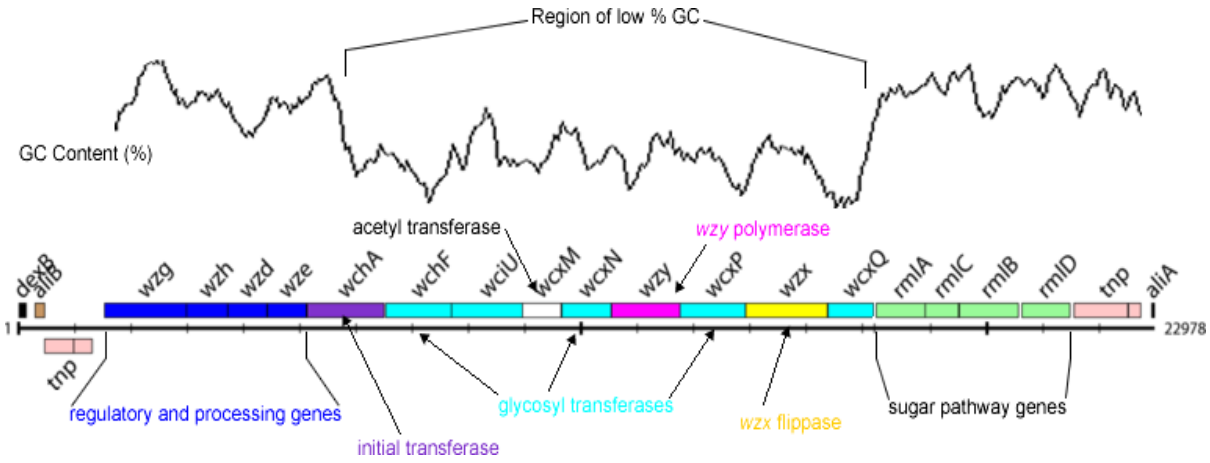


Figure 2.1 Makeup of a generic *cps* locus: for explanation of colour coding and graphical representation see methods.

Excepting insertion-sequence (IS) transposases (*tnp*), the first 4 genes, designated as *cap1*, *cap2*, *cap3* and *cap4* (Bernheimer and Wermundsen, 1972), or *cpsA*, *cpsB*, *cpsC* and *cpsD* in various studies and, definitively, using the Bacterial Polysaccharide Gene Naming (BPGN) scheme (Reeves et al., 1996) as *wzg*, *wzh*, *wzd and wze* are conserved, with high sequence similarities, and are present in all cases at the 5' end of the *cps* locus (Figure 2.1). In most cases the 5th gene encodes the initial sugar phosphate transferase WchA (also known as *cpsE*), which is responsible for the linkage of an activated glucose phosphate to the lipid carrier (Pelosi et al., 2005). Downstream of the 5th gene there is a serotype-specific region that exhibits a distinct low percentage G/C content (Figure 2.1), and contains GT genes, the CPS polymerase gene (*wzy*) and the CPS repeat-unit flippase gene (*wzx*) and often genes encoding acetyl transferases. In many cases the 3' end contains genes encoding enzymes for the biosynthesis of activated monosaccharide precursors (sugar pathway genes). Some of these are common to several CPS types, such as the rhamnose pathway genes whose products produce UDP-rhamnose and which is not available for CPS biosynthesis from central metabolism (termed non-house-keeping sugars). This description is somewhat generic, however, as there are often differences in the location of pathway genes, glycosyl transferases and the *wzx* and *wzy* genes can be found in the opposite order to the above.

2.1.2 Wzx/Wzy-dependant CPS biosynthesis

The Wzx/Wzy-dependent pathway involves the synthesis of the repeat-unit of the CPS, its translocation across the cytoplasmic membrane, and its polymerisation and linkage to the cell wall. Figure 2.2 depicts the production of the polysaccharide from the pneumococcal serotype 23F (Aanensen et al., 2007).

Synthesis begins by the transfer of an initial sugar, usually glucose-1-phosphate (Glc-1-P), from UDP-Glc to a lipid acceptor (undecaprenyl phosphate UndP) to create Und-PP-Glc, catalysed by a Glc-1-P glycosyltransferase (WchA) (Figure 2.2 (2) (Kolkman et al., 1997). As this reaction initiates synthesis of the CPS repeat-unit the transferase involved is referred to as the initial transferase (IT). In previously published cases, where capsules contain Galactose (Gal) and not Glc as the initial sugar, synthesis is initiated by the transfer of Gal-1-P, from UDP-Gal, to UndP, catalysed by an IT that recognises Gal-1-P.

A series of GTs transfer further sugars, from their respective nucleotide-sugar precursors, to the non-reducing end of the growing repeat-unit (Figure 2.2 (3)). Many transferase specificities have been characterised or suggested, for example, in serotypes 14 (Kolkmann et al., 1997), 2 (Iannelli et al., 1999) and 8 (Munoz et al., 1999; Saksouk et al., 2005; Pelosi et al., 2005) through a combination of sequence homology, biochemistry and site directed mutagenesis. Those capsules containing non-house-keeping sugars, for example the sugar rhamnose (Rha) in Figure 2.2, have, in their *cps* locus, a corresponding biosynthesis gene or pathway, and a GT or GTs with specificity for the non-house-keeping sugar. For Rha, the rhamnose biosynthetic pathway genes are present (*rmlA, rmlC, rmlB* and *rmlD*), the products of which are responsible for the conversion of UDP-Glc to TDP-Rha, and the GT genes, *wchF* and *wchW* the products of which transfers Rha to the repeat-unit (Figure 2.2 (1)).

The translocation of the mature repeat-unit across the cytoplasmic membrane, still attached to Und-PP, is proposed to be mediated by the polysaccharide repeat-unit transporter Wzx (flippase) (Figure 2.2 (4)). Following translocation, polymerisation of the repeat-units is achieved by the repeat-unit polymerase, Wzy, which transfers the reducing end of the growing polysaccharide chain onto the nonreducing end of the new repeat-unit to give an Und-PP-attached polymer (Figure 2.2 (5)). The presence of the *wzx* and *wzy* genes in the *cps* loci is therefore a useful predictor of the presence of the Wzx/Wzy-dependant pathway.

The products of the first four genes *wzg, wzh, wzd* and *wze* are implicated in the regulation of CPS production. Wzg functions as a transcriptional activator that regulates the level of CPS production in *Streptococcus agalactiae* and is thought to play a similar role in *S. pneumoniae* (Cieslewicz et al., 2001). Wzd and Wze are predicted to function together in the polymerization and export of CPS, and are thought to be functionally similar to ExoP in exopolysaccharide production in *Rhizobium meliloti* (Glucksmann et al., 1993) and Wzc from *Escherichia coli* K12 and K30 (Paiment et al., 2002). Studies have shown that Wze is an autophosphorylating protein-tyrosine kinase and that Wzd is required for Wze tyrosine phosphorylation (Morona et al., 2000). The final step in production of the capsule is the linkage of the polysaccharide to the cell wall (Figure 2.2(6)). The precise mechanism for this remains undetermined although it has recently been shown that the phosphorylated form of Wze promotes attachment to the cell wall (Morona et al., 2006).

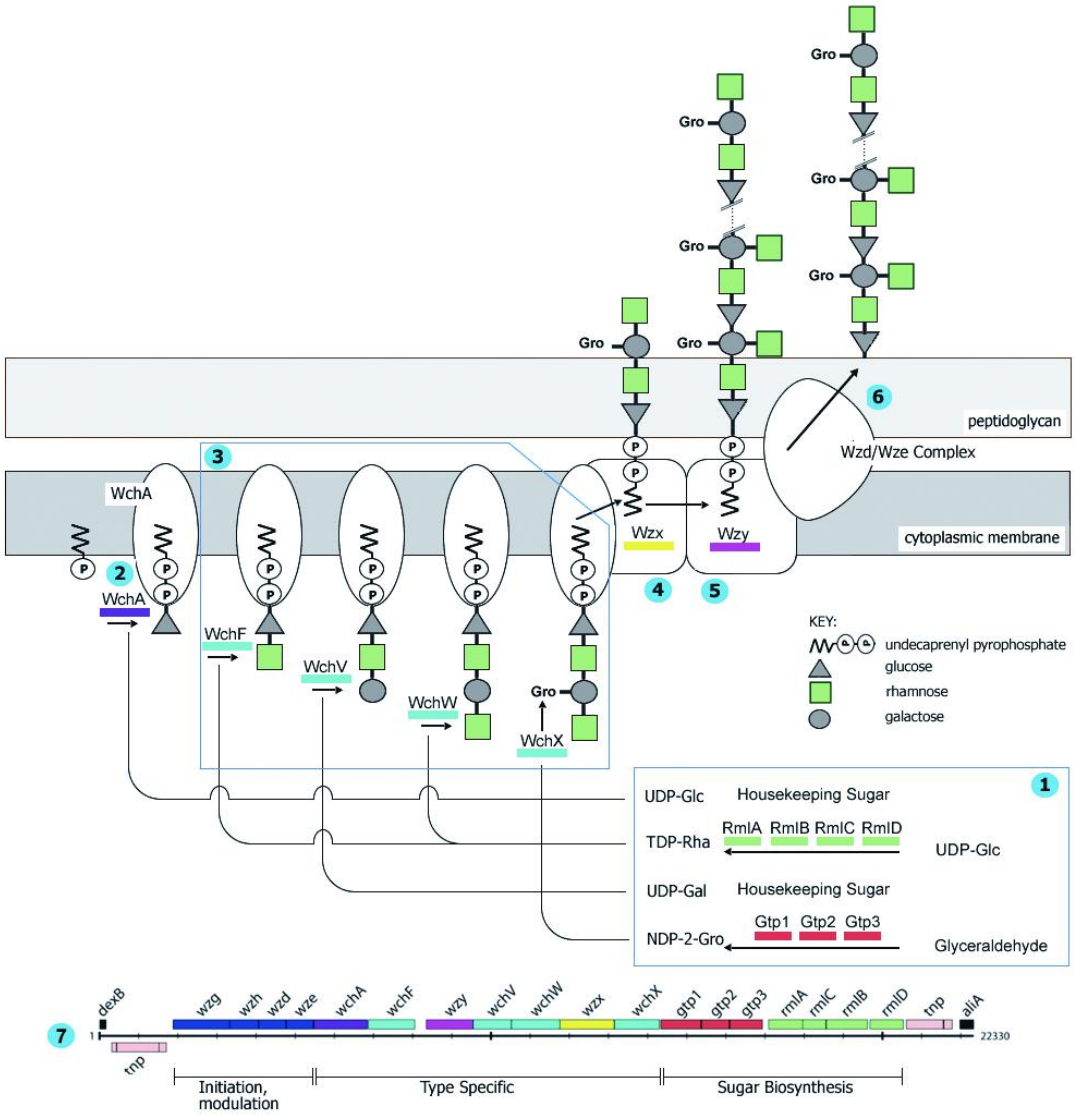


Figure 2.2 Schematic representation of the biosynthesis of CPS by the Wzx/Wzy-dependant pathway. The biosynthesis of serotype 23F is represented.

1) Specific nucleotide sugar biosynthesis.
2) The initial transferase (WchA here) links the initial sugar as a sugar phosphate (glucose-phosphate; Glc-P) to a membrane-associated lipid carrier (widely assumed to be undecaprenyl phosphate; Kolkman et al., 1997).
3) Glycosyltransferases sequentially link further sugars to generate the repeat-unit.
4) Wzx flippase transports the repeat-unit across the cytoplasmic membrane.
5) Wzy polymerase links individual repeat-units together to form lipid-linked CPS.
6) Lipid-linked CPS is linked to the cell wall by a poorly understood process involving the Wzd/Wze complex, with release of the undecaprenyl carrier.
7) The genes in the representation of the *cps* locus and their gene products are colour coded.

Reproduced from Aanensen et al., (2007).

2.1.3 Serotypes 3 and 37: Synthase-dependant pathway

The two simplest CPS structures are found in serotypes 3 and 37, both of which have been shown to be produced by an alternative to the Wzx/Wzy pathway: the synthase-dependant pathway. The CPS of both serotypes consist of a disaccharide repeat-unit, serotype 3 consisting of glucose (Glc) and glucuronic acid (GlcA) (Cartee et al., 2005), and type 37 consisting of so-called 'sophorosyl' units made up of a glucose main chain and an additional glucose side branch in each repeat-unit (Llull et al., 2001). The synthase pathway is dependant only on a single integral membrane protein, responsible for the sequential addition of sugars and the extrusion of the polymer across the cell membrane (Yother, 2004). The serotype 3 locus contains the synthase gene *wchE* and the gene *ugd*, the product of which is responsible for the synthesis of UDP-GlcA. Both genes are located between *dexB* and *aliA* along with non-functional copies of the first 4 genes of the Wzx/Wzy-dependant pathway (*wzg*, *wzh*, *wzd* and *wze).* Two other genes are present, *galU* and *pgm*, encoding a Glucose-1-phosphate uridylyltransferase and a phosphoglucomutase respectively, however, these have been shown not to be essential for type 3 CPS biosynthesis (Bonofiglio et al., 2005). Interestingly, unlike capsules formed by the Wzx/Wzy-dependant pathway, the type 3 polysaccharide is not linked to the cell wall and up to 50% of the polymer can be detected in culture supernatants (Hardy et al., 2000).

Serotype 37 is unusual, in all strains so far examined, there is a complete copy of a non-functional serotype 33F *cps* locus between *dexB* and *aliA*. However, the production of the serotype 37 CPS is directed by a single synthase gene *tts*, located elsewhere on the chromosome (Llull et al., 2001). Not only does this demonstrate that not all *S. pneumoniae* serotypes are directly determined from the products of the genes located between *dexB* and *AliA*, but that some caution must be used when assessing the correlation of biochemical structures, serology and gene content to bear in mind that inconsistencies can occur.

2.1.4 Sequencing of 90 *S.pneumoniae cps* loci

All laboratory work and initial sequence annotation was undertaken by collaborators at the Sanger Institute led by the project manager, Stephen Bentley, and will now be briefly described.

Strains used for the production of anti-sera for each of 90 pneumococcal serotypes utilised were provided by the Statens Serum Institut, Denmark and exploiting the conserved location of *cps* loci, between *dexB* and *aliA*, the *cps* loci were amplified and sequencing was undertaken. Prediction of the location of coding sequences (CDS) was undertaken using the program GLIMMER trained on the TIGR4 *S. pneumoniae* genome sequence (Tettelin et al., 2001), and annotation of gene products was based on similarities with known sequences deposited in primary databases such as GenBank. To aid the annotation of such a large set of chromosomally related loci, and to group similar proteins together into manageable subsets, the algorithm TribeMCL (Enright et al., 2003) was used (see methods). Subsets of similar protein sequences are assigned to a homology group (HG) and given a unique number which can be used to annotate across HGs based on previously identified proteins, and similarities to functional domains within proteins characterised using information held at the protein family database, Pfam (Sonnhammer et al., 1997).

From an evolutionary perspective, it can be envisioned that selective pressure, exerted from the human immune system, has favoured the emergence of novel serotypes. Mechanistically, over time, ancient gene shuffling events, including gene loss and duplication and the acquisition of new genes from within and outside of the pneumococcal gene pool, and minor sequence divergence of existing serotypes, would lead to the emergence of further novel serotypes, the consequence being the 90+ serotypes that we recognise today. It would also be expected that the genetic events leading to these immune evasion strategies continue to play a part in the evolution of the pneumococcal *cps* locus. Attempting to untangle the complete evolutionary history of such a complex locus is probably unreasonable, however it should be possible to untangle the more recent events that have led to the production of 'newer' serotypes given the sequences of 90 *cps* loci. Furthermore, relating the changes in the sequences to the gene products, the structural similarities and the immunological consequences of such structural differences should lead to a deeper understanding of the mechanisms and nature of capsule diversity. It was with this in

mind, and the knowledge that similar issues are present in the analysis of other biological systems that the development of the analysis framework presented began.

A consequence of using the TribeMCL method for annotating the *cps* loci is that each serotype can be defined as a string of integers based upon its HG content. This 'homology group profile' (HGp) can then be used to cluster similar serotypes together (see methods) based on the number of HGs in common. This approach has been utilised and refined to take into consideration the DNA similarities between genes encoding the shared HGs (see methods) and, rather than inferring direct phylogeny, forms the basis for investigating further the similarities and differences between groups of serotypes that share similar genes (HGs). The storage of all data in a web-accessible database along with the careful design of, and choice of tools, incorporated within the web application should allow the further analysis to be achieved.

The methods section of this chapter is split into three sections. Firstly, the nature and storage of the data, and ways of handling, gene naming and visualisation of data will be described. This is followed by an explanation of the sequence analysis methods utilised to provide further data for storage and display and to provide data for the overall clustering of *cps* loci. Finally, the database and web software will be detailed with reference to 'workflows' utilised in the analysis. The results section details the overall results of the sequencing and the dendrogram produced based on clustering of *cps* loci, followed by two pertinent examples detailing the relationships between two groups of *cps* loci, their structures and their immunological reactivity profiles.

2.2 Materials and Methods

The focus of development was on building a system that would be generic, scaleable and applicable to other systems, but to do so in a fashion that would address immediately the questions raised within the pneumococcal *cps* system.

The underlying methods are based on established sequence alignment methodologies, however the combination of these methods into workflows should be governed by the questions a researcher may have and the tools available within the system should, to some extent, pre-empt the user's question. The challenge in such systems is to make the software and workflows generic, efficient, easy to use and the results easily interpretable (visualising the data).

2.2.1 Data handling

The raw data used for the project were as follows:

1. DNA sequences: 90 EMBL format text files containing the raw *cps* sequences, annotation (including Pfam domain matches) and homology group membership provided by collaborators at the Wellcome Trust Sanger Institute.
2. CPS structures: Representations for the 54 known repeat-unit were sourced from Kammerling (2000).
3. Serology: Tables of reactions of each CPS with the typing sera were sourced from Henrichson (1994).

2.2.2 TribeMCL

As mentioned, to aid the annotation of such a large set of sequences (1.8Mbp in total) the algorithm TribeMCL was used by Stephen Bentley (Sanger Institute). TribeMCL is an algorithm that clusters sets of proteins into 'protein families' or 'homology groups (HGs). Clustering is achieved by analysing similarity patterns within a given dataset and using these patterns to assign proteins into related groups. TribeMCL uses Markov Clustering and the program takes as input the results of sequence alignments between all members of the protein set, typically using BLAST, and groupings are based on a defined cut-off level of similarity.

The entire protein set is initially formatted into a "BLASTable" database, and each individual protein sequence is then used as a query to search against the database to return significant matches. A level of similarity can then be set within which TribeMCL groups the proteins into HGs.

The level of similarity utilised was based upon the BLAST Expect (e) value of e^{-40}. The e-value is a parameter describing the number of hits within a database search that would be expected to be seen by chance. The lower the e-value, the more significant is the match and the chosen cut-off value (e^{-40}) should produce groups of proteins that are sufficiently similar to be considered to have common ancestry (homology). Using HGs provided a basis for assigning similar function to sets of proteins and therefore aided in the annotation of the sequences. HG membership for each assigned coding sequence (CDS) was included in the EMBL files and provided the basis for subsequent clustering.

2.2.3 Gene naming

Many of the TribeMCL HGs encoded by the *cps* genes were unique and had not been assigned a gene name, excepting those members identified from analyses of the previously published sequences of pneumococcal *cps* loci. Once the EMBL files were received, the Bacterial Polysaccharide Gene Nomenclature (BPGN) was used and several 'alphabets' of new gene names were assigned specifically for the *cps* sequences (Reeves et al., 1996). HGs were used as a basis for assignment of gene name, and all members of the same HG were given the same gene name except the Wzx and Wzy proteins. The Wzx (flippase) and Wzy (polymerase) proteins were divergent and fell into several HGs, but as *wzx* and *wzy* have been widely used for these genes, we have retained the gene names for all but add the HG number (e.g *wzx*-xx where 'xx' is the HG). If a member of a HG had already been published and assigned a name, this name was given to all other members of the group. EMBL annotation was updated to reflect the newly assigned gene names.

2.2.4 Encoding of raw data

2.2.4.1 DNA sequences

2.2.4.1.1 Graphical encoding of *cps* loci

Traditionally, for publication, a linear representation of each *cps* locus has been drawn by hand. The requirements of displaying whole genomes in a visual manner has led to the development of genome-level linear diagrams and scripts, written using the programming language Perl. The 90 *cps* loci were rendered as linear diagrams using methods previously utilised for whole genome diagrams using a Perl script kindly provided by Kim Rutherford (University of Cambridge) which, since undertaking this study, has been incorporated within the software package DNAPlotter (Carver et al., 2009). The entire *cps* locus is represented linearly around a scaffold bar representing the entire length of the sequence. Each CDS is then mapped onto the bar in the correct position and orientation, based on the EMBL annotation. To aid visual comparisons, each class of CDS (sugar pathway genes, glycosyltransferases, flippase, polymerase and sugar biosynthesis pathway genes) was coloured differently.

2.2.4.1.2 *In silico* storage of sequences

Each EMBL file contains the raw DNA sequence along with information regarding each CDS, based around a coordinate start and end point in the sequence. Information contained includes, the predicted product of the gene, the gene name (where applicable), and the homology group membership of the CDS. The following example shows a simplified version of the annotation of an *rmlB* gene, the product of which is part of the rhamnose biosynthetic pathway encoded within the *cps* locus of several serotypes:

```
-----------------------------------
FT   CDS  18324..19373
FT        /note="member of homology group 0008"
FT        /product="dTDP-D-glucose 4,6-dehydratase RmlB"
FT        /gene="rmlB"
-----------------------------------
```

The first line indicates the start and end coordinate of the CDS within the *cps* sequence, the second line shows the homology group membership, the third line contains the predicted product and the fourth the gene name.

Should a CDS be frame-shifted (a deletion or insertion that causes the sequence to be translated in a different reading frame), this can be represented in the CDS line of the EMBL file as a 'join', indicating the sequence start and end points in each frame, for example:

```
-----------------------------------
FT   CDS             join (13416..13583,13587..14417)
-----------------------------------
```

The raw DNA sequence is present as one long string at the end of the EMBL file.

The EMBL format was initially produced as a descriptive way of storing sequence data in flat files to aid submissions to central databases. However, even though the file format is optimised for description, the data contained within is not immediately amenable for transfer to a database other than EMBL. Consequently, Perl scripts were written to extract all of the information into a TAB-delimited format that would be amenable for entry into a database, the schema for which is described in section 2.2.6.1

2.2.4.2 Biochemical repeat-unit structures

2.2.4.2.1 Graphical encoding of CPS structures

The complexity of carbohydrate molecules provides a challenge for diagrammatic and, more importantly, digital representation and a number of different formats have been produced with varying degrees of recommendation. The nomenclature recommended by The International Union of Pure and Applied Chemistry – International Union of Biochemistry and Molecular Biology (IUPAC-IUBMB) describes oligosaccharides using a 3-letter code to characterise monosaccharide units, for example Glc = glucose, Gal = galactose etc., with glycosidic bond linkage information in parenthesis between the monosaccharides (see Appendix A). However, within the biological literature there is an alternative representation, used rather less frequently, where each monosaccharide is represented by a different symbol, joined

together by various types of lines representing the different linkages. For the purposes of this project a symbol based approach, using previously published sugar symbols (where available), and new symbols or plain text for the infrequently seen sugars (mostly those encoded for by pathways encoded within the *cps* loci) was used. Furthermore, those sugars which have pathways encoded within the *cps* loci have been colour-coded according to the colour used in the representation of the relevant pathway genes in the linear *cps* gene diagrams to further aid visual interpretation. It is important to note that, as opposed to the *cps* locus diagrams, the structural diagrams, in common with general practice in the literature, read from right to left with the Und-P linked first sugar on the far right of the diagram. Sugar residues within the repeat-unit can be referred to numerically from right to left.

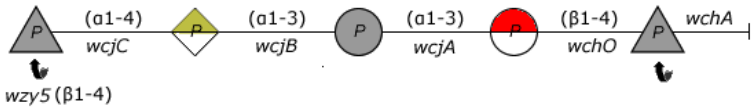
2.2.4.2.2 *In silico* encoding of CPS structures

The string-based nature of DNA sequence data and its annotation makes it easily manageable within relational database systems, however, encoding the branched structures of carbohydrate molecules make this digital representation a more challenging prospect. As the field of comparative glycomics is in its infancy, digital representation schemas are relatively few and not wholly satisfactory, however, the two most mature representations are CabosML (Kikuchi et al., 2005) and LINUCS (Bohne-Lang et al., 2001). Building on the fact that an oligosaccharide unit is essentially a tree-like structure, CabosML uses the tree like XML data description format to encode oligosaccharides, and LINUCS represents a structure using a linear string of nested parentheses. For the purposes of this project a modified version of the LINUCS notation has been used to encode the available structures into one string that can be stored easily in a database.

2.2.4.3 Combining sequence and structure representations

Rendering the *cps* loci and polysaccharide repeat-unit structures graphically in this fashion allows us to move towards combining information from the two data sources to allow the specificity of the enzymes encoded by transferase genes (and the Wzy polymerase linkage) to be incorporated into one diagram. For example, the previously published serotype 9A repeat-unit structure with its glycosyltransferase specificity can be represented as in Figure 2.3a, and the linear *cps* locus diagram can be seen in Figure 2.3b.

a)



b)

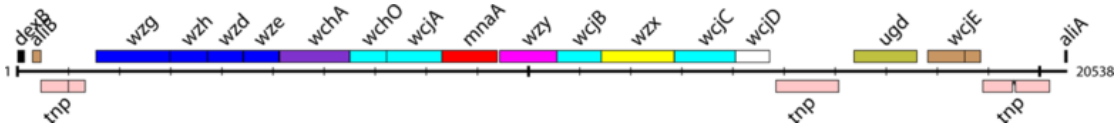


Figure 2.3 Graphical representations of a) the CPS repeat-unit structure and b) the *cps* locus of serotype 9A.

In the structural representation of the repeat-unit of serotype 9A, the gene whose product is believed to catalyse each linkage is shown, e.g. *wchA* encodes the IT which attaches the first sugar to the UND-P lipid carrier. The linkage between repeat-units, catalysed by the repeat-unit polymerase Wzy, is indicated by thick black arrows. House-keeping sugars (i.e. those available from central metabolism) are coloured grey whereas non-house-keeping sugars are coloured in relation to the sugar pathway genes present in the *cps* locus. In this example the UDP-N-acetylglucosamine-2-epimerase encoding gene *mnaA* is coloured red and corresponds to the presence of the red/white symbol for the monosaccharide, N-acetylmannosamine (ManNac), in the representation of the structure.

2.2.4.4 Serology

Each pneumococcal serotype is associated with a characteristic pattern of reactivity with the panel of pneumococcal typing sera. A table of these sera reactivity patterns for 90 serotypes is available (Henrichsen, 1995) and details were stored in the database, on a per-serotype basis**.**

2.2.5 Sequence alignment methods

The identification of regions of similarity between sequences is a fundamental process in biological sequence analysis. Sequence alignments can be broadly split into two categories, global and local. Global alignments are based on aligning with the maximum number of matching bases along the entire sequence length, and are most suited for comparisons between fairly similar sequences of a similar length, whereas local alignments attempt to identify the best sub-sequence matches and are more useful when applied to sequences which have regions of similarity interspersed with dissimilar regions, rearrangements or differing lengths.

2.2.5.1 Basic Local Alignment Search Tool (BLAST)

The largest possible unit of comparison is that of a full *cps* locus aligned against any other. As mentioned previously, global alignment of the sequences of the different *cps* loci is not appropriate due to the complex evolutionary history of the *cps* system. However, visualisation of a one against one local alignment can be used to investigate closely related serotypes, or similar regions within different serotypes, to tease out fine scale similarities.

Perl scripts were written to allow the production of all versus all BLAST comparisons at both the DNA sequence (blastn) and protein sequence (tblastx) level for the 90 serotypes in both directions (1 vs 2 and also 2 vs 1), in all possible combinations, representing a total of 16200 comparisons. The textual output produced by BLAST can then be used for visualisation of sequence similarities within software such as the Artemis Comparison Tool (Carver et al., 2005; see 2.2.6.4).

2.2.5.2 Clustal

Clustal is a tool allowing the simultaneous global alignment of many sequences (Thomson et al., 1994). Briefly, the multiple alignment method firstly aligns all pairs of sequences, using a fast approximate method, to calculate a distance matrix giving the divergence of each pair of sequences. This matrix is then used to produce a guide tree and subsequently, the sequences are progressively aligned according to the branching order in the guide tree. This allows for a minimal set of more robust alignments to be produced, cutting down compute time.

ClustalW was used to produce alignments at both the DNA and amino acid sequence level for each homology group, and results and trees made available through the website.

2.2.5.3 Needle

The TribeMCL method groups similar protein sequences together into a HG (section 2.2.2). However, in order to investigate the sequence similarities among the genes encoding members of the same HG (and hence to infer evolutionary relationships), the program Needle was used.

Needle, which forms part of the software suite EMBOSS (Rice et al., 2000), uses the Needleman-Wunsch global alignment algorithm to find the optimum alignment of two sequences using a scoring matrix that contains values for every possible nucleotide match. The maximum scoring alignment is found, where the score of an alignment is equal to the sum of the matches from the scoring matrix and the overall similarity score between any pair of sequences can be expressed as a percentage. For every HG, all vs all nucleotide and amino acid sequence comparisons were generated by Needle and stored as a % similarity score in the database, allowing overall % similarity levels for a HG to be calculated and also to allow direct comparisons of members of a HG at the DNA level when clustering the loci (detailed in section 2.2.7.2).

2.2.6 CPS Web Application

The core analysis team involved in the day to day handling of the data were myself, Stephen Bentley (Sanger Institute), Angeliki Mavroidi (Imperial), Peter Reeves (University of Sydney) and Brian Spratt (Imperial). All needed access to the data and relevant tools in a form that would make analysis a group effort. Furthermore, all should all be able to ask different questions of the data, depending on expertise, using the same tools. The bioinformatics expertise within the group differed and the platforms and software commonly used by individuals also differed. For example, Stephen Bentley works on a Unix platform and is comfortable with command line methods as well as desktop software. Angeliki Mavroidi and Peter Reeves are more comfortable with software packages with graphical interfaces, allowing data to be loaded and analysed. One solution would be to develop a desktop software package. However, this would involve multiple copies of the data and could not be easily amended to include new tools as the project developed. The one common platform all have access to is the web and as such was the only real choice when deciding how to progress. The application started in a simple form (just a standard database search), however, as the data were investigated further, different tools were needed which could easily be added to a web application, once, with immediate access all.

The following sections firstly describe the database schema, then the web application, and finally, its functionality will be explored.

2.2.6.1 Database schema

As mentioned in section 2.2.4.1.2, the EMBL format used to store sequences and their annotation is not readily amenable for querying and further analysis without importing into third-party applications. Data were therefore entered into a database in a number of tables designed to minimise redundancy, and to serve as a framework for building a user interface to allow the *cps* loci to be explored. The database system used is SQL Server 2000 and the import of data was achieved using a combination of Perl, Visual Basic and SQL. Figure 2.4 shows the database schema and relationships between tables graphically.

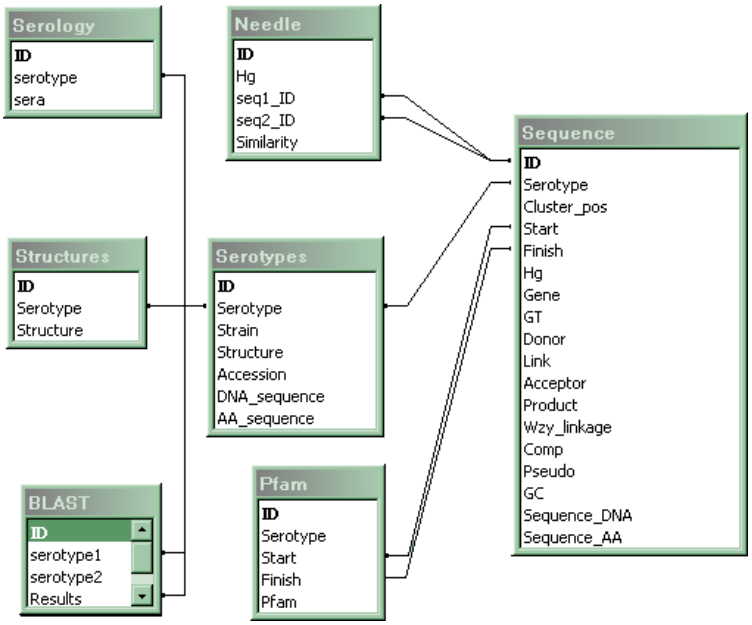


Figure 2.4 CPS Application database schema.

The central table 'Serotypes' contains information specific to each serotype including the strain the *cps* locus was sequenced from, its EMBL accession number, a boolean indicator as to whether the associated CPS structure is available and the entire raw DNA and translated amino acid sequences. Should the CPS structure be available, a link is present to the 'Structures' table, which contains the LINUCS notation for the corresponding structure. Information relating to the CDS content of each serotype is stored in the 'Sequence' table which, for every CDS from every serotype, contains data including the CDS start and end co-ordinate in the *cps* locus, its homology group membership and gene name. Every CDS is assigned a unique identifier (ID) and this is utilised in the table containing the results of the all vs all Needle comparisons undertaken and utilised in the clustering of the *cps* loci ('Needle'). Serological data are stored in the table 'Serology' and are linked to the serotypes table based on the serotype columns. Two further tables are present: 'BLAST' contains the results from the all vs all full length *cps* locus BLAST alignments undertaken for presentation to a user using ACT (See 2.2.6.4) and 'Pfam' contains information on the location within CDS of functional domains based on searching the Pfam database. The start and end co-ordinates of the Pfam matches are based on the co-ordinates of the CDS sequences, this allows overlapping Pfam matches to be stored successfully.

The design and storage of data in such a fashion allows server-side scripting to be used to interrogate the data in a logical fashion utilising the Structured Query Language (SQL). Complex queries can be built incorporating data from multiple tables. Further tables of information can be added easily and ad-hoc queries can be built and executed based on the relationships between the data.

2.2.6.2 Presentation and application logic

The website is hosted on a server within the Department of Infectious Disease Epidemiology, Imperial College London. The web interface includes a navigation bar, along the top, that allows access to the different sections within the site as follows:

1) Database
2) WebACT
3) Serotypes
4) Structures
5) Statistics

2.2.6.3 Database

2.2.6.3.1 Database querying

Below the tab-based menu navigation are two options (Figure 2.5), the first allows the dendrogram based on sequence similarities between clusters to be interrogated and will be detailed in the section on serotype clustering (section 2.2.8). The second contains a link to a BLAST server, which allows any DNA or amino acid sequence to be entered and queried against the entire set of *cps* sequences. The 'Search sequence database' section of the page contains a form allowing entry of a single criterion or a combination of criteria, for example, serotype = '1' and gene = 'pgm', or HG = '4'.

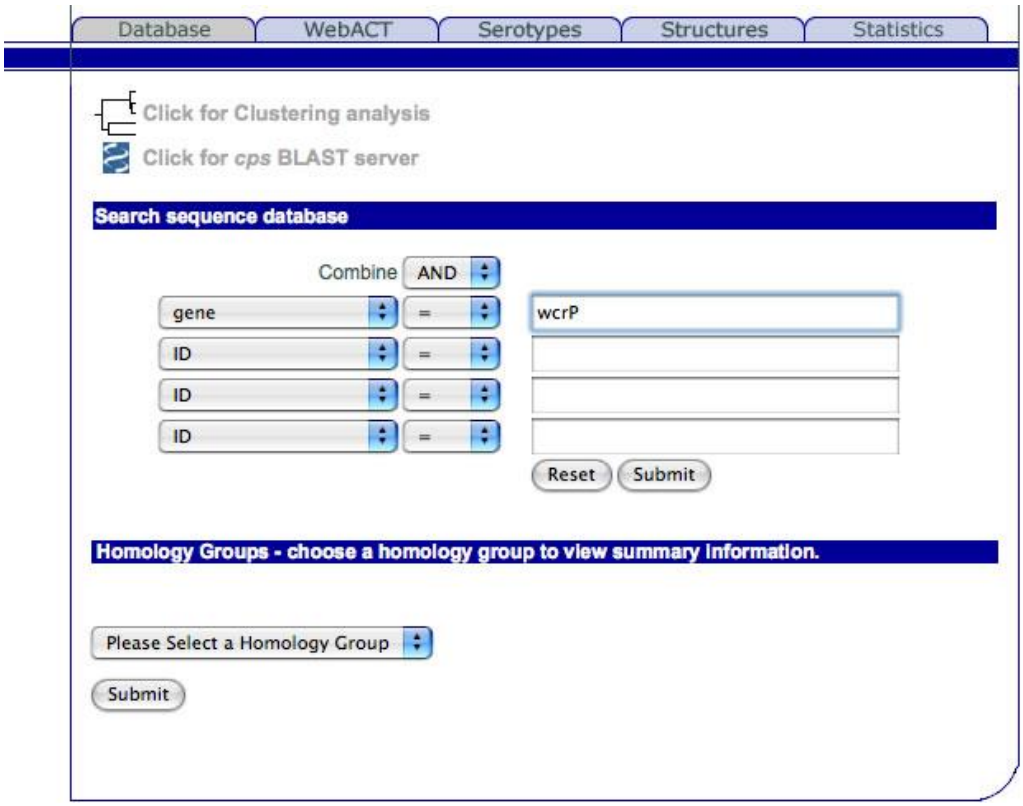


Figure 2.5 Database Query form.

Upon entering a query (in Figure 2.5, 'gene=wcrP') and submitting the form, the results returned include the annotation and database fields for each CDS (Figure 2.6). The corresponding DNA sequences can then be submitted to ClustalW for sequence alignment and subsequent tree production. ClustalW is run at the server and resulting trees are visualised using a Java applet version of the tree-viewing software ATV (Zmasek and Eddy, 2001)

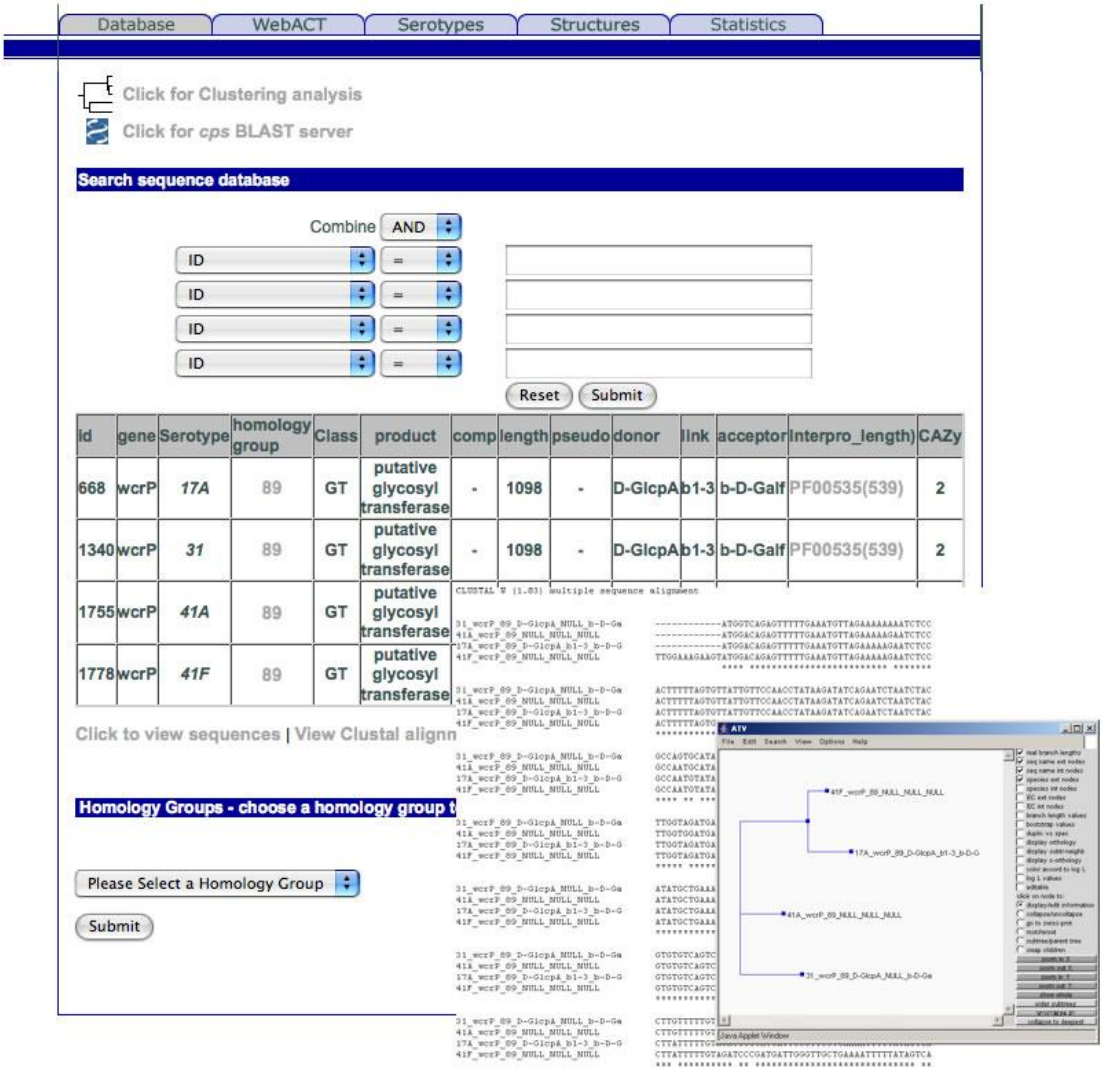


Figure 2.6 Database search results including subsequent clustalW alignment and dendrogram visualised using ATV.

2.2.6.3.2 Homology group querying

The latter section of the database search page (Figure 2.5) contains a form allowing any HG to be selected. Following form submission, a summary for each HG is displayed as follows (and in Figure 2.7):

- HG number and gene name
- A list of serotypes that the HG is found in.
- Average overall % sequence similarity at both the DNA and amino acid level, based on pairwise Needle comparisons.
- Average proportion GC content for the HG.

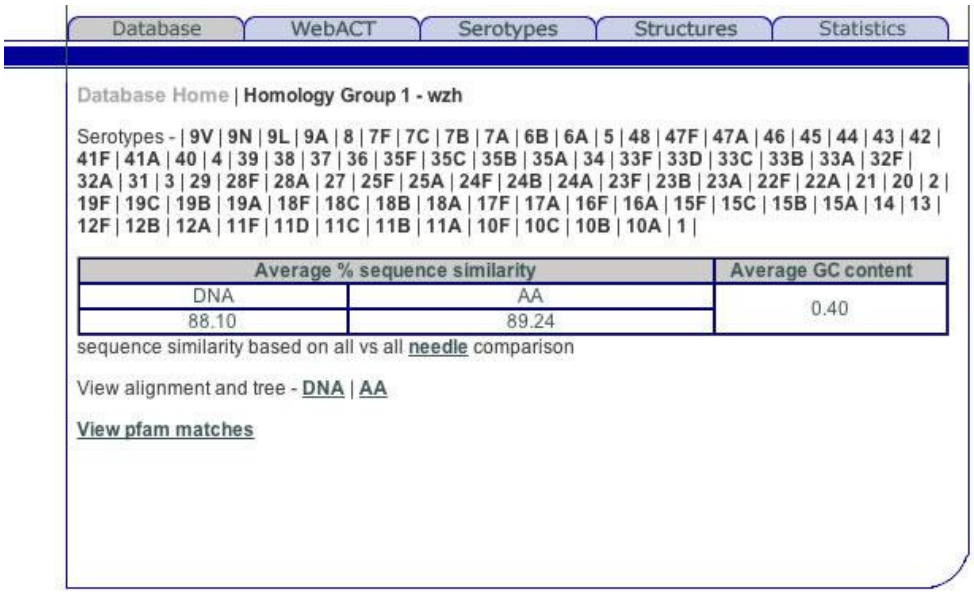


Figure 2.7 Homology group query results.

ClustalW alignments and resultant trees based on the HG can also be viewed directly from this point along with protein domain matches, based on searching Pfam, found in the HG.

2.2.6.4 WebACT

The Artemis Comparison Tool (ACT) (Carver et al., 2005) is an interactive graphical DNA sequence comparison viewer, which permits the visualisation of pairwise comparisons created using BLAST. Multiple comparisons between a number of sequences can be stacked. For example, a three-way comparison can be visualised where pairwise alignments have been generated between sequences 1 and 2, and sequences 2 and 3. It should however be noted that the order of the sequences in such multi-way comparisons can have a significant impact upon the interpretation of the results, since regions of similarity between sequences 1 and 3 are not identified in the example. Therefore, a thorough investigation of a group of sequences necessitates the production of all vs all comparisons, which traditionally has been undertaken on a one by one basis by a user.

Using the all vs all BLAST comparisons detailed in section 2.2.5.1, a simple web interface was developed to allow combinations of *cps* clusters to be viewed. Firstly, a user chooses how many sequence comparisons they wish to view, then the serotypes, and in which order they would like to view the comparisons (and which type of BLAST comparison they wish to use: blastn or tblastx). Upon form submission, the

linear graphical representations of the *cps* clusters are presented and a link to view the comparisons loaded into ACT is presented. The results of a tblastx comparison between 3 serotypes (18A vs 18B vs 18C) can be seen in Figure 2.8.

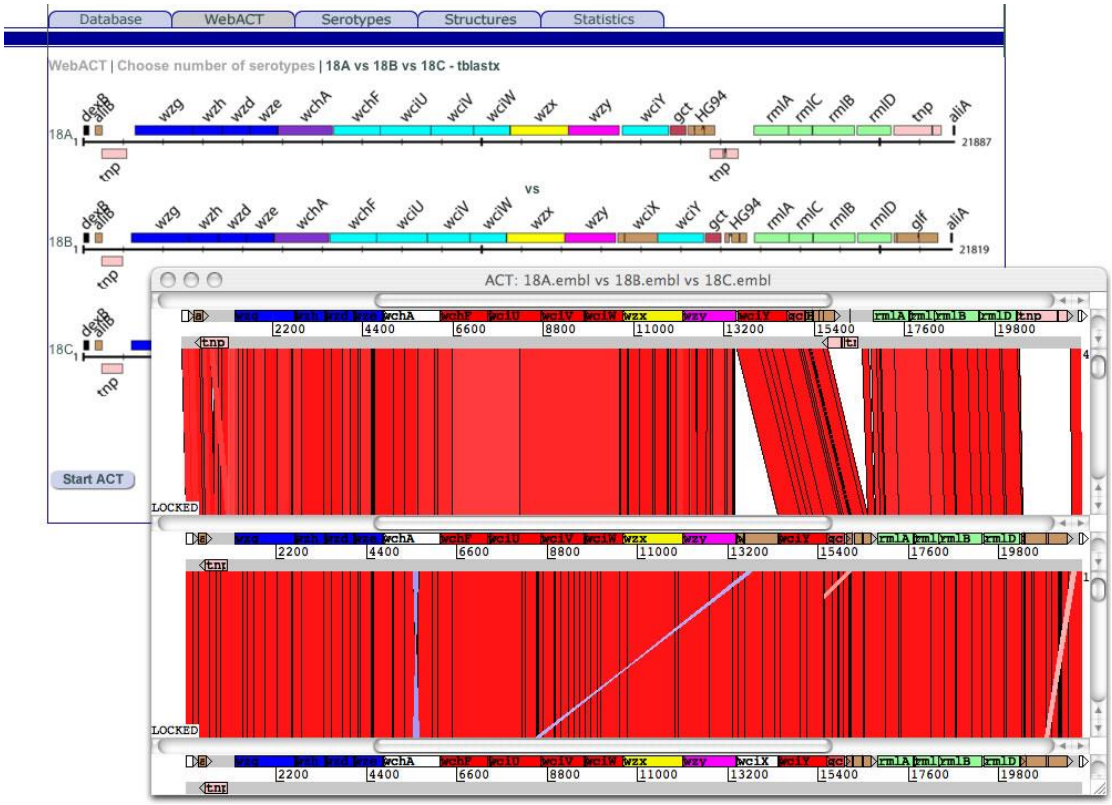


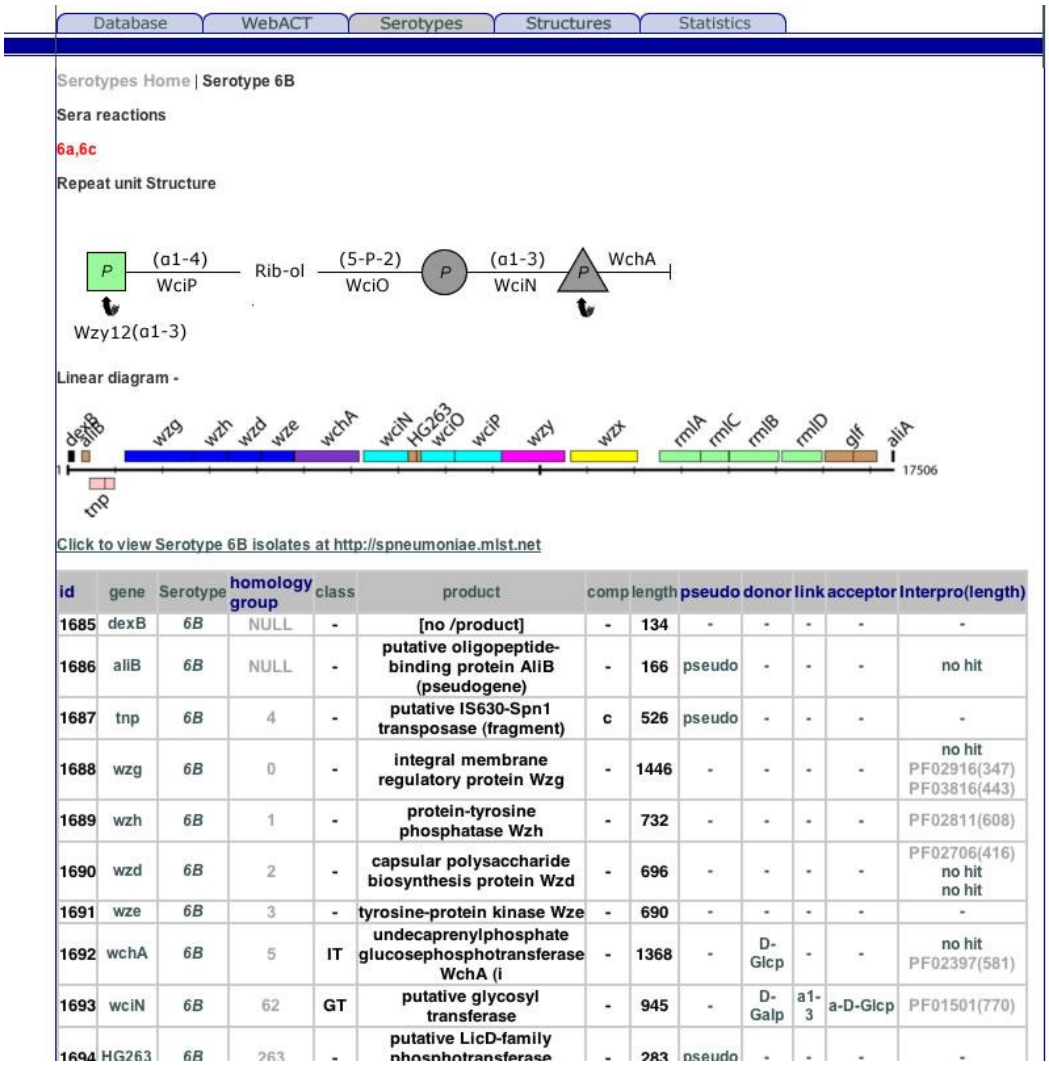
Fig 2.8 A three-way ACT comparison: Serotypes 18A vs 18B vs 18C. Each red band between two serotypes represents a BLAST hit and ACT can be used to zoom all the way to sequence level to investigate regions further.

While developing the interface, it became clear that such a method would be applicable for any user who wished to view and generate multiple comparisons for any set of genomic sequences. A collaboration was initiated with James Abbott at The Bioinformatics Support Service, Imperial College London, and the online tool WebACT was produced (Abbott et al., 2005, 2007). WebACT allows the production of and viewing, using ACT, of pre-computed BLAST comparisons between any bacterial genome sequences available in public databases or for on-the-fly comparison generation from user-entered queries.

2.2.6.5 Serotypes

A search form is provided to view a summary page for each serotype. Following serotype selection, the results can be seen as in Figure 2.9. The following information is detailed:

- The graphical representation of the structure of the CPS repeat-unit (if known).
- Linear diagram of the *cps* locus.
- Total number of CDS within the locus.
- List of CDS within the *cps* locus, gene names and products, Pfam matches with links to the Pfam website, length of CDS and, for GTs and the *wzy* gene, the predicted glycosidic linkage made by the gene product etc.
- Typing sera that react with the serotype.
- Link to epidemiological information held on strains of a particular serotype at http://spneumoniae.mlst.net. mlst.net is the subject of chapter 3 of this thesis.



id	gene	Serotype	homology group	class	product	comp	length	pseudo	donor	link	acceptor	Interpro(length)
1685	dexB	6B	NULL	-	[no /product]	-	134	-	-	-	-	-
1686	aliB	6B	NULL	-	putative oligopeptide-binding protein AliB (pseudogene)	-	166	pseudo	-	-	-	no hit
1687	tnp	6B	4	-	putative IS630-Spn1 transposase (fragment)	c	526	pseudo	-	-	-	-
1688	wzg	6B	0	-	integral membrane regulatory protein Wzg	-	1446	-	-	-	-	no hit PF02916(347) PF03816(443)
1689	wzh	6B	1	-	protein-tyrosine phosphatase Wzh	-	732	-	-	-	-	PF02811(608)
1690	wzd	6B	2	-	capsular polysaccharide biosynthesis protein Wzd	-	696	-	-	-	-	PF02706(416) no hit no hit
1691	wze	6B	3	-	tyrosine-protein kinase Wze	-	690	-	-	-	-	-
1692	wchA	6B	5	IT	undecaprenylphosphate glucosephosphotransferase WchA (i	-	1368	-	D-Glcp	-	-	no hit PF02397(581)
1693	wciN	6B	62	GT	putative glycosyl transferase	-	945	-	D-Galp	a1-3	a-D-Glcp	PF01501(770)
1694	HG263	6B	263	-	putative LicD-family phosphotransferase	-	283	pseudo	-	-	-	-

Figure 2.9 Serotype summary detailing serotype 6B.

2.2.6.6 Structures

The ‘Structures’ section of the application details predictions of GT, IT and Wzy polymerase linkage specificity based on correlations made between the presence of genes in the *cps* loci with specific linkages within the known CPS repeat-unit structures. Details of methods and results are beyond the scope of this chapter, which focuses on the development of the web software and its use. However, the rationale behind the methods and results for these tentative predictions are fully detailed in Aanensen et al, (2007). Briefly, the predictions allowed us to assign a specific linkage within a CPS structure to a specific GT gene within the *cps* locus and furthermore allowed confidence to be placed on the linkage polymerised by Wzy and the initial sugar transferred by an IT. Predictions were stored in the database and can be viewed during database searching and are also included on the linear representations of the *cps* loci (Figure 2.3). Within the structures section, either a GT gene, a particular

sub-section of a structure or functional protein domains (Pfam) can be used to query the database and the results return all known graphical structure representations containing the query.

2.2.6.7 Statistics

Functional groupings of CDS' from CPS produced via the Wzx/Wzy pathway are summarised as counts based on the entire sample set. CDS not involved in CPS biosynthesis (*dexB, aliA, aliB*) were counted separately.
The functional groupings are as follows:

- Regulatory genes (*wzg, wzh, wzd, wze*)
- *wzx* flippase gene
- *wzy* polymerase gene
- Transferases (broken down into initial transferases, glycosyltransferases - including those such as rhamnosyltransferases, acetyl transferases, sugar phosphate transferases and pyruvyl transferases).
- Sugar pathway genes (each individual pathway gene).
- Insertion Sequence (IS) transposases.
- Others (hypothetical proteins where a function has not been assigned, glycosyl transferase enhancers, group II intron proteins).

It is also possible to view all CDS' based on these functional groupings and the serotypes they occur in and furthermore, the Pfam domains can be viewed through links to the Pfam website.

2.2.7 Serotype clustering

Serotypes 3 and 37 were not included in the analysis due to the differing mode of CPS biosynthesis, therefore only the 88 CPS, presumed to be produced via the Wzx/Wzy pathway, were included.

CDS' not involved in CPS synthesis were also not included (*tnp, dexB, aliB*) and furthermore, the first four genes, which are present in all 90 serotypes (and in each

case their products are very similar and fall into a single HG), were excluded so that the tree was not weighted unnecessarily.

2.2.7.1 Dendrogram based on homology group sharing

Initially, a dendrogram was produced based solely on differences in HG content.
The HGs encoded by each *cps* locus were used to define the HG profile (gene content) and the HG profile of each serotype was compared to every other, in turn, in both directions (i.e. 1 vs 2 and also 2 vs 1). The similarity between two serotypes was defined as the ratio of the number of HGs in common divided by the total number of HGs in the second serotype, expressed as a percentage, using the following formula:

$$\frac{sHG_1 + sHG_2 + ... sHG_n}{N_{HGs}} \text{ X } 100$$

where sHG_n represents a shared HG and N_{HGs} represents the total number of HGs in the second serotype. Comparing in both directions takes into consideration the unequal numbers of genes within each *cps* locus and gives us two values for the same pair of serotypes. As an example, for the comparison of two serotypes (S1 and S2) that share five HGs where S1 has a total of ten HGs and S2 has a total of fifteen HGs, there are two comparisons:

S1 vs S2 = (5/10)*100 = 50%
S2 vs S1 = (5/15)*100 = 33.3%

The resulting 88 x 88 two-sided matrix was subjected to Agglomerative Hierarchical clustering (AHC) and transformed into a one-sided distance matrix utilising the Pearsons dissimilarity distance. AHC is a bottom-up clustering method that is used extensively for the analysis of microarray gene expression data. Initially, each object (in this case a serotype) is treated as a separate group. Following this, each serotype is successively combined with others based on similarity until there is only one group remaining. The dendrogram was then produced using the unweighted pair-group method with arithmetic averages (UPGMA) method.

Perl scripts were written to produce the matrix from data held in the database and AHC and subsequent dendrogram production was undertaken using the software XLSTAT v7.5.3 (Addinsoft, 1995-2005).

2.2.7.2 Dendrogram incorporating the degree of sequence similarity

The grouping of HGs using TribeMCL is based on similarities in their amino acid sequences and proteins assigned to the same HG could be 100% identical or could be moderately divergent. However, due to the redundancy inherent in the genetic code, the DNA sequences of the genes encoding the members of an HG could be substantially more divergent than their amino acid sequences. Thus, two *cps* loci that are syntenic (share the same gene content and order) could share genes that are almost identical in sequence or genes that differ by 20% or more in sequence. Clearly, the two *cps* loci in the former case are more closely related than those in the latter and, to include this evolutionary information within the tree, a refinement of the above clustering method was used that incorporates a measure of the % DNA sequence similarity between the genes encoding the shared HGs, using values obtained from the Needle comparisons (see section 2.2.5.3). For shared HGs the % DNA sequence similarity was obtained and an average % similarity was scaled based on the total number of HGs in the second serotype, using the formula:

$$\frac{sHG_1\% + sHG_2\% + ... sHG_n\%}{N_{HGs}}$$

where $sHG_n\%$ represents the % DNA similarity between the two members of sHG_n and N_{HGs} represents the total number of HGs in the second serotype.

To demonstrate the increased robustness inherent in this method, the resultant clustering of the four serotypes of serogroup 15 and serotype 14 shall be briefly described.



Figure 2.10 Clustering of serotype 14 and serogroup 15 based on a) gene content and b) sequence weighted gene content.

Serotype 14 has more HGs in common with serotypes 15A, 15B and 15C than the latter serotypes do with serotype 15F; thus serotype 14 appears more similar to serotypes 15A, 15B and 15C than the latter are to serotype 15F (Figure 4a). However, the sequences of the genes encoding the shared HGs are rather similar among serogroup 15, whereas those of serotype 14 are more divergent. Consequently, when the similarity of the genes of the shared HGs is taken into account, the four serotypes of serogroup 15 cluster together, with serotype 14 appearing somewhat less related (Figure 4b).

For the purposes of analysing the *cps* loci the sequence-weighted tree was used, and further details and applications are presented in Mavroidi et al, (2007).

2.2.8 Interrogating the Dendrogram within the CPS Web Application

Within the 'Database' section of the web application detailed above (section 2.2.6.3) an option is available to 'Click for clustering analysis'. This presents an interactive version of the sequence-weighted tree allowing 'drilling down' to view ACT comparisons of sub-clusters. Clustering details will be given in section 2.3.3. However, initially the entire dendrogram is visible and any sub-cluster can be clicked to view further details as in Figure 2.11.

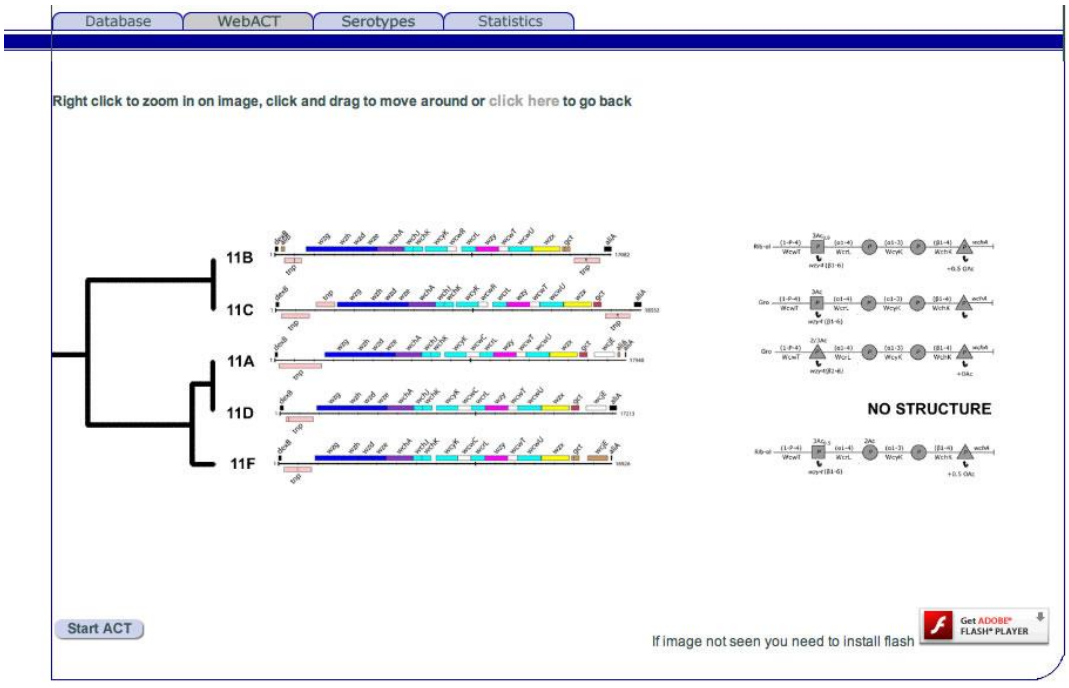


Figure 2.11 Dendrogram of sub-cluster containing serogroup 11, viewed within the CPS application.

The page presents the section of the tree selected, the serotype represented by each terminal node and the linear representation of the corresponding *cps* cluster. Additionally, the structural representations of the CPS repeat-units (if known) are shown. The sub-clusters are rendered using Adobe Flash and 'Right-clicking' the figure allows zooming in to view further detail. A link to 'Start ACT' is present at the bottom left of the page that, when clicked, launches ACT pre-loaded with the BLAST comparisons in the order represented on the dendrogram (Figure 2.12).

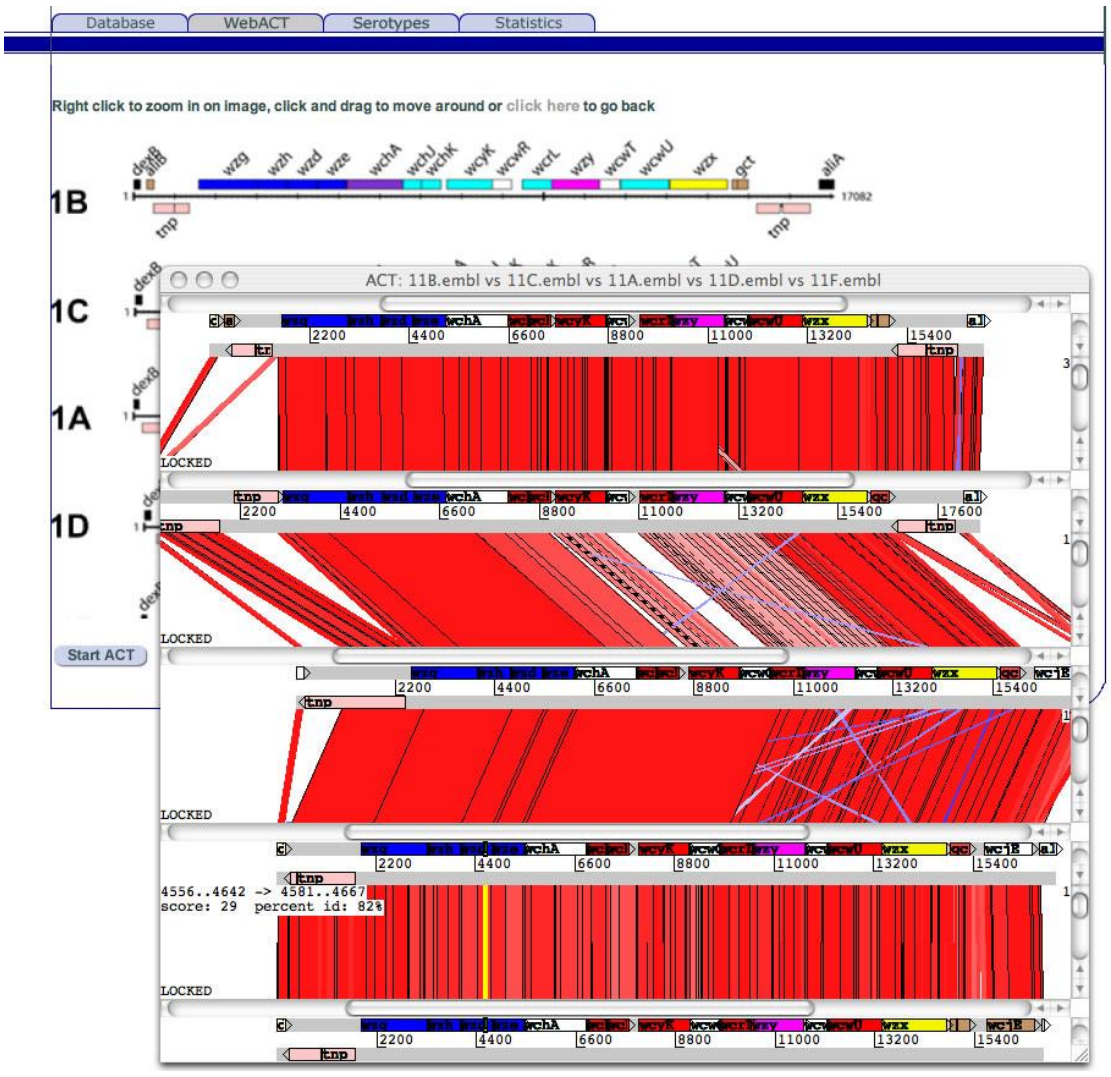


Figure 2.12 ACT comparison based on dendrogram interrogation: Serotypes 11B vs 11C vs 11A vs 11D vs 11F.

Using this method any sub-cluster of the tree can be viewed within ACT based on the results of the *cps* clustering method detailed in section 2.2.7.2.

2.2.9 Demonstration of workflow for analysing clusters of serotypes

To demonstrate the applicability of the web application for analysing clusters, from a user perspective, a workflow will be described based on investigating the relationships between serotypes 18B and 18C. The *cps* loci of serotypes 18B and 18C are syntenic and both of the CPS repeat-unit structures are known.

1) When accessing the sequence-weighted tree as described in section 2.2.8, both serotypes 18B and 18C fall within cluster 8.
2) Clicking on cluster 8 displays a 'zoomed in' view of the tree, which is delineated further into subclusters. Serotype 18B and 18C are located within sub-cluster 6.
3) Clicking on sub-cluster 6 displays the branch of the tree leading to serotypes 18B and 18C along with the linear *cps* locus diagrams and the structural CPS representation which can be zoomed in on by right clicking and selecting 'zoom'.
4) Serotypes 18B and 18C are syntenic, and the only difference that can be seen is the presence of a non-functional copy of the *wciX* gene in 18B, which in serotype 18C is functional. The only difference between the two CPS repeat-unit structures is the acetylation of the terminal glucose in serotype 18C, which is un-acetylated in 18B. WebACT can be launched directly at this point pre-loaded with the comparisons of serogroup 18. However, we shall take an alternative route.
5) To investigate further we shall search the database for the *wciX* gene. Clicking the 'Database' menu option takes us to the database search form. Entering the search term 'Gene = wciX' and clicking submit returns all instances of this gene in the database. The acetyltransferase *wciX* (HG107) is found in only 3 serotypes: 18B, 18C and 18F.
6) To view the amino acid and nucleotide differences we shall view the two serotypes, 18B and 18C, using the WebACT facility. Clicking the 'WebACT' tab takes one to the page asking how many serotypes would like to be viewed.
7) Select '2', which then offers two drop down boxes containing lists of serotypes.
8) Select 18B in the first box and 18C in the second box.
9) We shall look at the tblastx comparison, which was generated by BLASTing the amino acid sequences against each other. By default, tblastx is selected. Click 'submit'.
10) The results page shows the two linear cluster diagrams highlighting the *wciX* difference.
11) Clicking 'Start ACT' launches a WebStart version of ACT preloaded with the BLAST comparison between the two serotypes.

12) The ACT window shows the sequence of 18B at the top and 18C at the bottom. The red bands in-between indicate the position and strength of the BLAST hits between the two sequences. Clicking on a particular hit shows the BLAST score and the % identity. Sequences can be scrolled from left to right and zoomed in by using the scroll bars on the top, bottom and to the right.

13) Clicking once on the *wciX* gene of serotype 18B selects the gene and also highlights the BLAST matches that cover the sequence. Right clicking and selecting 'Zoom to selection' will zoom so that the whole of the gene is in view. To zoom in further the top right scrollbar can be used. Zoom until the amino acid and DNA sequences can be viewed.

14) To centre the corresponding region of serotype 18C, double-click the yellow BLAST hit directly below the 18B *wciZ* gene.

15) Scroll along *wciZ* until the stop codon is found. It can be seen that the three bases 'TGA' encode the stop codon. By looking at the corresponding amino acids and DNA in serotype 18C it can be seen that instead of a stop codon a G (glycine) is encoded by the bases 'GGA'

16) Therefore, a single base change of T to G has encoded a stop codon making the *wciX* gene in 18B non-functional.

17) It is also possible to view the DNA and amino acid sequence alignments using ClustalW from the homology group summary and, if required from the database results page.

18) The result can be used to suggest that a single base difference has introduced a stop codon in the acetyl transferase encoding gene *wciX* in serotype 18B with subsequent loss of acetylation in the CPS compared to that in 18C.

Utilising the web application described, and combinations of similar workflows as that detailed above, allowed analysis of the CDS makeup and general features of the 90 *S. pneumoniae cps* loci to be investigated and relationships between biochemical structures and serological reactions to be inferred based on clustering. Prior to presenting specific examples of the analysis of resulting clusters, the general features of the dataset, as described in Bentley et al, (2006) shall be detailed.

2.3 Results

2.3.1 General features of the *dexB-aliA* locus from 90 *S.pneumoniae* serotypes

Sequence length for the 90 *cps* loci ranged from 13,337 base pairs (serotype 3) to 30,298 bp (serotype 38) with an average length of 20,714 bp. In total, 1999 CDS' were identified and a functional designation could be assigned to all but 26 of them. As previously reported, the synthase gene (*wchE*) of serotype 3 is located within the *cps* locus, however the type 37 *cps* locus, which was very similar to that of serotype 33F, is defective and the 37 serotype is determined by the synthase gene (*tts*) which is located elsewhere on the chromosome (Llull et al., 2001).

The general characteristics of the loci, in common with previously sequenced clusters can be described as follows:

1 The regulatory genes *wzg*, *wzh*, *wzd* and *wze* are conserved, in all cases displaying high sequence identity, and are almost always present in the same order at the 5' end of the locus.

2 65 *cps* loci contained *wchA* as the fifth gene, encoding initial transferases with specificity for glucose phosphate.

3 The polymerase (*wzy*) and flippase (*wzx*) genes are always located downstream, together with a varying set of genes encoding GTs, sugar pathway genes, acetyltransferases and other transferases.

4 In every case the central region of the *cps* locus, which includes the genes that tend to be specific to each serotype or serogroup, have anomalously low % G+C content.

5 In almost every instance there is evidence of mobile genetic elements flanking the *dexB* and *aliA* genes, consisting of intact or disrupted IS transposases.

6 Nine sugars not normally available through central metabolism (non-house-keeping) are found within the *S. pneumoniae* capsules and this corresponds perfectly with the presence of the relevant pathway genes in the *cps* loci.

91% of the proteins fell into 175 HGs containing between 2 and 90 members, with the remaining 74 proteins forming single member HGs. The products of the first four genes *wzg*, *wzh*, *wzd* and *wze* each fell into a single HG covering every serotype.

Ignoring IS transposases, the next largest HG comprises the 65 *wchA* initial transferase encoding genes. At the other extreme, the serotype-specific gene products show great diversity with 87 HGs for non-initial sugar transferases and 40 and 13 groups of Wzy polymerases and Wzx flippases respectively.

2.3.2 Initial transferases and Wzy linkages

There is biochemical evidence from previously studied CPS biosynthesis that the initial transferase WchA adds glucose-1-phosphate to undecaprenol phosphate to create Und-PP-Glc (Pelosi et al., 2005) and it is assumed that it performs the same function in all 65 serotypes where found. For those structures that are known, this correlates perfectly with the presence of glucose in the repeat-unit. Where *wchA* is not found, the products of the fifth gene fall into three HGs (designated as WciI, WcjG and WcjH), all of which contain the same Pfam domain as WchA and display similar hydrophobicity profiles to the carboxy-terminal domain of WchA. It is suggested that they function as initial transferases due to evidence showing that the carboxy-terminal domain of the WchA homologue, WbaP in *Salmonella enterica* is sufficient for transferase activity (Wang et al., 1996).

The initial sugar also acts as the donor sugar in the polymerisation of the repeat-units and the presumed specificity of the Wzy polymerase determines the other component of the linkage. For the known CPS structures, identification of the initial sugar allowed determination of the polymerase linkage as both donor and acceptor sugar, and the linkage was defined once the initial sugar had been identified. Where there was any ambiguity as to the nature of the linkage due to two residues of the initial sugar being present in the repeat-unit, the polymerase linkage was provisionally identified by considering the linkage catalysed by other members of the same HG.

2.3.3 *cps* locus clustering

The HGs encoded by each *cps* locus were used to produce a HG profile (gene content) for each of the 88 serotypes produced via the Wzx/Wzy pathway. The similarities in the profiles were displayed by constructing a tree based on a weighted measure of the extent of sharing of HGs and the degree of sequence similarity of the genes encoding the shared HGs (as described in section 2.2.7.2).

In order to rationally discuss sets of closely related serotypes the tree was truncated at levels that define major clusters of serotypes (cut-off of 0.3) and also within these clusters, sub-clusters were defined (cut-off of 0.05) to identify very closely related groups of serotypes.

At a truncation level of 0.3, the *cps* loci fall into eight major clusters (See Figure 2.13). In all cases, serotypes within the same serogroup fall into the same cluster and 63 of the 88 *cps* sequences fell into the two largest clusters (clusters 3 and 8). In most cases the clusters consisted of the loci of several different serogroups, however the *cps* locus of serotype 8 and serogroups 9 and 11 are sufficiently different from all other *cps* loci to cluster alone.

21 sub-clusters were defined using the 0.05 cut-off. Serotypes from within the same serogroup generally fall into the same sub-cluster, however, serotype members from within serogroups 7, 16, 17, 33, 35 and 47 fell into different sub-clusters. It is also apparent that the *cps* loci of different serotypes may fall into the same sub-cluster and can be very similar or indeed syntenic, and furthermore, the similarities between different serotypes can sometimes be greater than that between some members of the same serogroup. Of the 21 sub-clusters, 9 were composed of different serogroups /serotypes, and of these, 3 contained completely syntenic *cps* loci. Serotypes 35A, 35C and 42 are syntenic and show a high sequence similarity, serotype 40 along with serotypes 7B and 7C are syntenic and within this subcluster, serotype 40 is more similar to serotype 7B than 7B is to 7C. Similarly, the three members of serogroup 12 are syntenic with serotypes 44 and 46. Below, I briefly describe examples of the similarities between *cps* loci, that were identified by clustering, relating these to the CPS structures (where known) and patterns of cross-reactivity with the typing sera. These are: the sub-cluster consisting of the syntenic serotypes of serogroup 12 and serotypes 44 and 46 (Figure 2.13A), and cluster 6 consisting of serogroup 11 (Figure 2.13B).

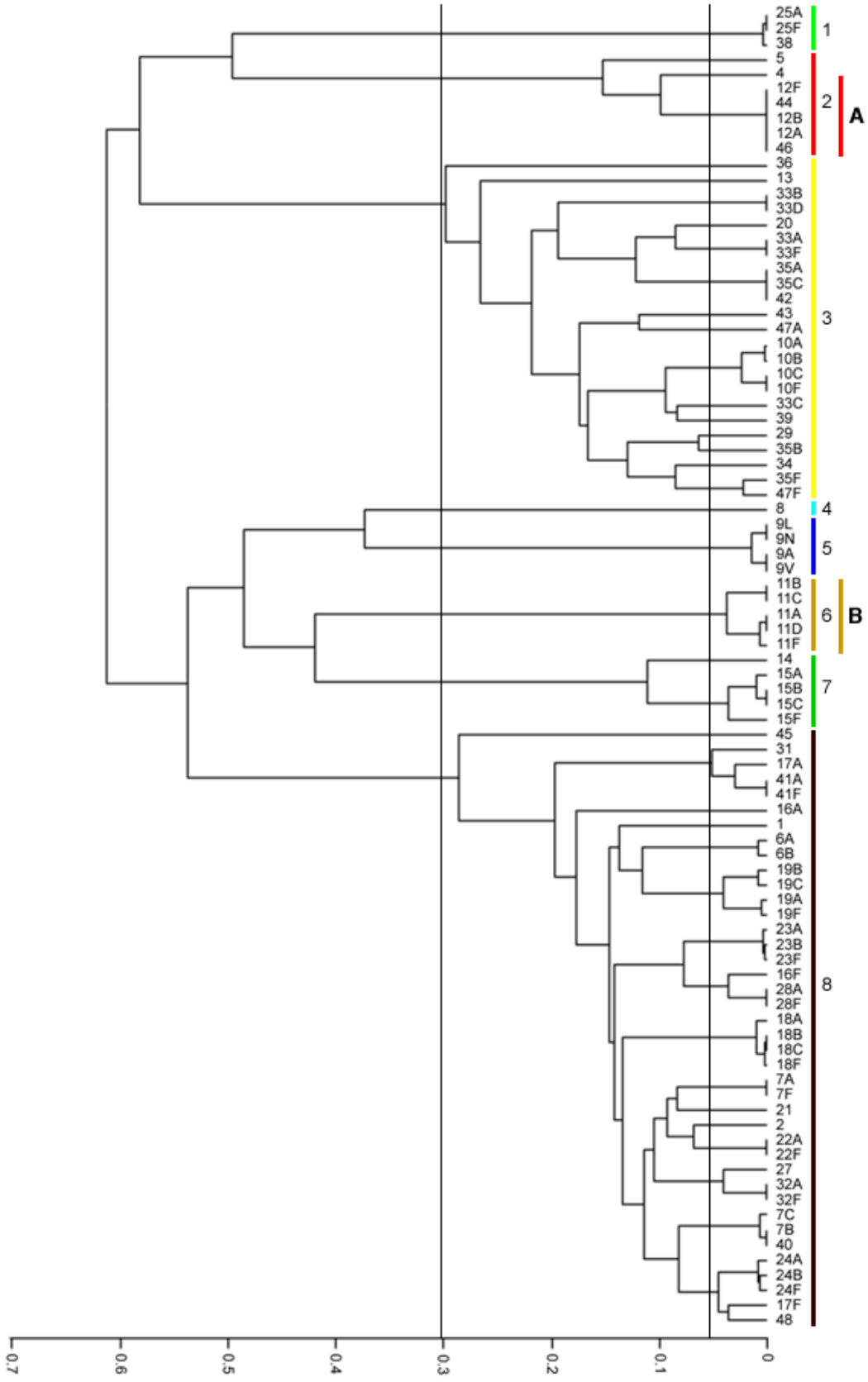


Figure 2.13. Dendrogram based on proportion of shared HGs, adjusted for sequence similarities between the shared HGs. The major clusters identified at a cut-off of 0.3 are shown (1-8). The line at 0.05 identifies very similar *cps* clusters (defined as sub-clusters). Highlighted regions of the tree (A and B) are discussed further in the text.

2.3.4 Serotypes 44 and 46 are related to serogroup 12

Excepting IS transposase genes, the *cps* clusters of serogroup 12 and serotypes 44 and 46 are syntenic and have presumably recently diverged from a common ancestor. This is not apparent from serology alone (i.e. serological investigations used to define serotypes did not recognise the similarity of these serotypes) (See Figure 2.14). When examining serological cross-reactions, no single factor sera cross-reacts with all five serotypes but cross-reactions do indicate some antigenic commonalities. Serotype 44 cross-reacts with factor sera 12b and 12d, while serotype 46 cross reacts with sera 12c.

Although structures have only been determined for 12A and 12F, the constituents for serotype 46 CPS are known and all are present in 12F and 12A (Kammerling, 2000). The CPS structures of 12F and 12A differ only in the initial sugar and a side branch. As there is synteny, these differences in repeat-unit structure are presumably due to the amino acid variation in the IT WciI and the GT (WcxB). The structures of types 12B, 44 and 46 are unknown, but the synteny and immunological cross-reactivity mentioned above suggest structural similarities with the CPS of serotypes 12F and 12A.

The *N*-acetylfucosamine (L-Fuc*p*NAc) biosynthesis pathway genes (*fnlA, fnlB* and *fnlC*) and the *N*-acetylmannosaminuronic acid (Man*p*NAcA) biosynthesis genes (*mnaA* and *mnaB)* are present in all loci of serogroup 12 and serotypes 44 and 46 and both L-Fuc*p*NAc and Man*p*NAcA can be seen in the CPS of serotypes 12A and 12F. This would predict structural similarities in the unknown structures, however, the type 44 *cps* locus has *fnlC* frameshifted and therefore, L-Fuc*p*NAc is not expected in its CPS structure.

The IT gene *wciI* is present in all loci, however the IT WciI appears to have different specificities as the initial sugar in known structures varies: *N*-acetylglucosamine (D-Glc*p*NAc) in serotype 12A and *N*-acetylgalactosamine (D-Gal*p*NAc) in serotype 12F). The amino acid sequences of WciI within this sub-cluster are very similar (70 to 100% pairwise identity; average, 81.4%) and this highlights the difficulties in predicting the enzymatic specificities of transferases from their amino acid sequence similarities.

The GT genes *wcxD*, *wcxE* and *wcxF* are present in all loci and their products are predicted to form the linkages common to these serotypes. Inspection of the two known structures suggests that these GTs catalyse the formation of the common tetrasaccharide element:

D-Glc*p*-(α1-2)-D-Glc*p*(α1-3)-D-Man*p*NAcA-(β1-4)-L-Fuc*p*NAc.

The GT gene *wciJ* is present in all of these serotypes and the L-Fuc*p*NAc is linked to the repeat-units but to different initial sugars. Therefore, we have tentatively assigned *wciJ* as the putative α-1,3-L-Fuc*p*NAc transferase gene. As in the case of the IT gene, the specificities of the linkage made by similar members (average pairwise identity, 83.9%) of the same HG (WciJ) appears to vary. The GT gene *wcxB* is present in serogroup 12 and types 44 and 46 and is suggested to encode the GT catalysing the α-1,3 linkage of D-Gal*p*NAc to α-L-Fuc*p*NAc in the CPS of type 12A and of galactose (D-Gal*p*) to α-L-Fuc*p*NAc in type 12F.

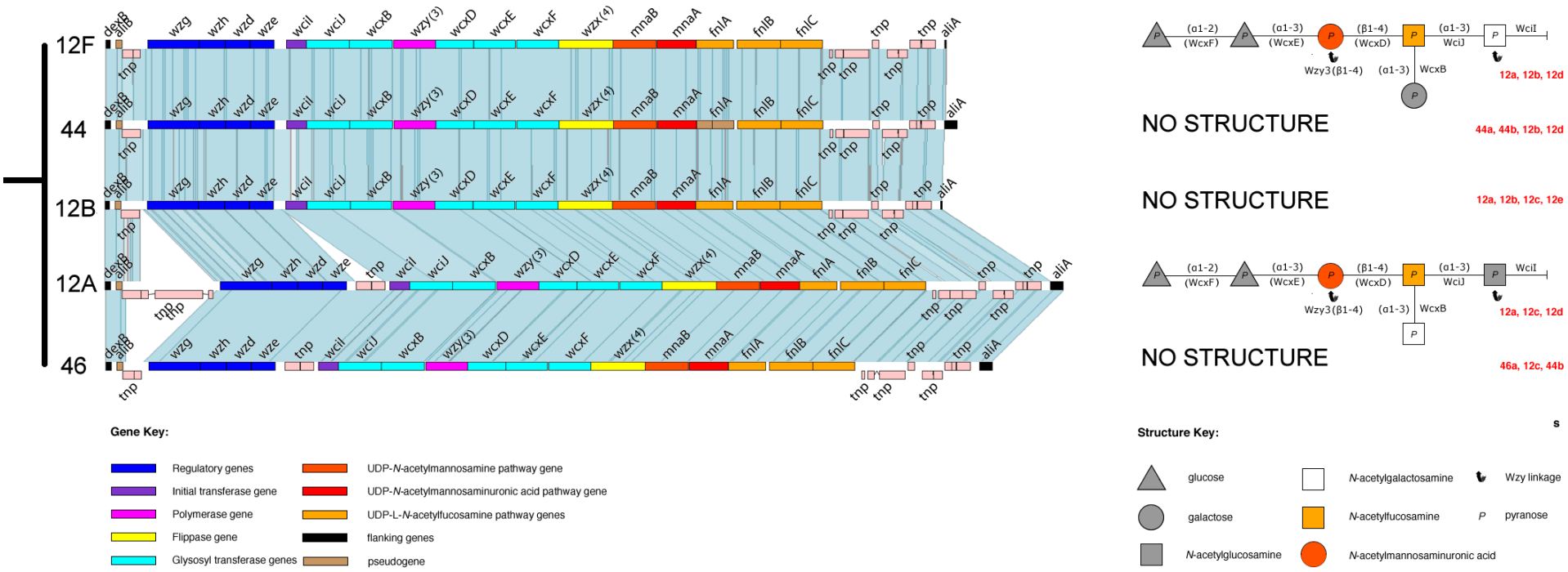


Figure 2.14 Comparison of those *cps* loci within cluster 2 assigned to the same subcluster: Serogroup 12 and serotypes 44 and 46. Light blue blocks between linear representations of the gene clusters indicate regions of sequence similarity. The Wzy homology group membership is indicated on the *wzy* gene in parenthesis. Repeat-unit structures are represented with the initial sugar that is attached to the lipid carrier on the right and reactive sera are listed in red below each structure (if known).

2.3.5 Serogroup 11

The five *cps* loci within cluster 6 (serogroup 11) consist of two syntenic sub-groups (11F-11A-11D and 11B-11C) differing only in their acetyltransferase genes, and have, similarly to the syntenic *cps* loci in cluster 2a (above), presumably diverged from a recent common ancestral *cps* locus (see Figure 2.15).

All five loci possess the acetyltransferase gene *wcwT*, but types 11F, 11A and 11D have two extra acetyltransferase genes (*wcjE* and *wcwC*) whereas types 11B and 11C have one extra acetyltransferase gene (*wcwR*). WcwU has been assigned as a glycerol phosphate transferase, and the presence of Gro-1P correlates with an intact *gct* gene in types 11A and 11C; *gct* is frameshifted in types 11F and 11B, and Rib-ol is present in the CPS instead of Gro (presumably house-keeping).

The *wchJ* and *wchK* genes are present in serogroup 11 *cps* loci. These genes are also present in serotype 14 *cps* (data not shown), where biochemical studies have shown that WchK is a β-1,4-galactosyltransferase that catalyses the linkage D-Gal*p*-(β1-4)- β-D-Glc*p*, whereas WchJ is suggested to act as an enhancer of WchK activity (Kolkman et al., 1997). It is therefore likely that the homologous gene products in serogroup 11 are responsible for the D-Gal*p*-(β1-4)- β-D-Glc*p* linkage found in their CPS. WcrL possesses the same Pfam domain (PF04488) as WciT of type 8, which has been suggested to catalyse the D-Glc*p*-(α1-4)-a-D-Gal*p* linkage (Broadbent et al., 2003).

The last sugar of the repeat-unit is D-Glc*p* in type 11A but D-Glc*p*NAc in types 11B, 11C and 11F, linked in all cases via an α-1,4 bond to α-D-Gal*p*. Thus it is suggested that the last sugar of the repeat-unit in serogroup 11 is transferred by WcrL and differences in amino acid sequence probably affect the specificity for the donor sugar. WcyK would therefore, by a process of elimination, catalyse the linkage D-Gal*p*-(α1-3)- β-D-Gal*p* in serogroup 11 CPS.

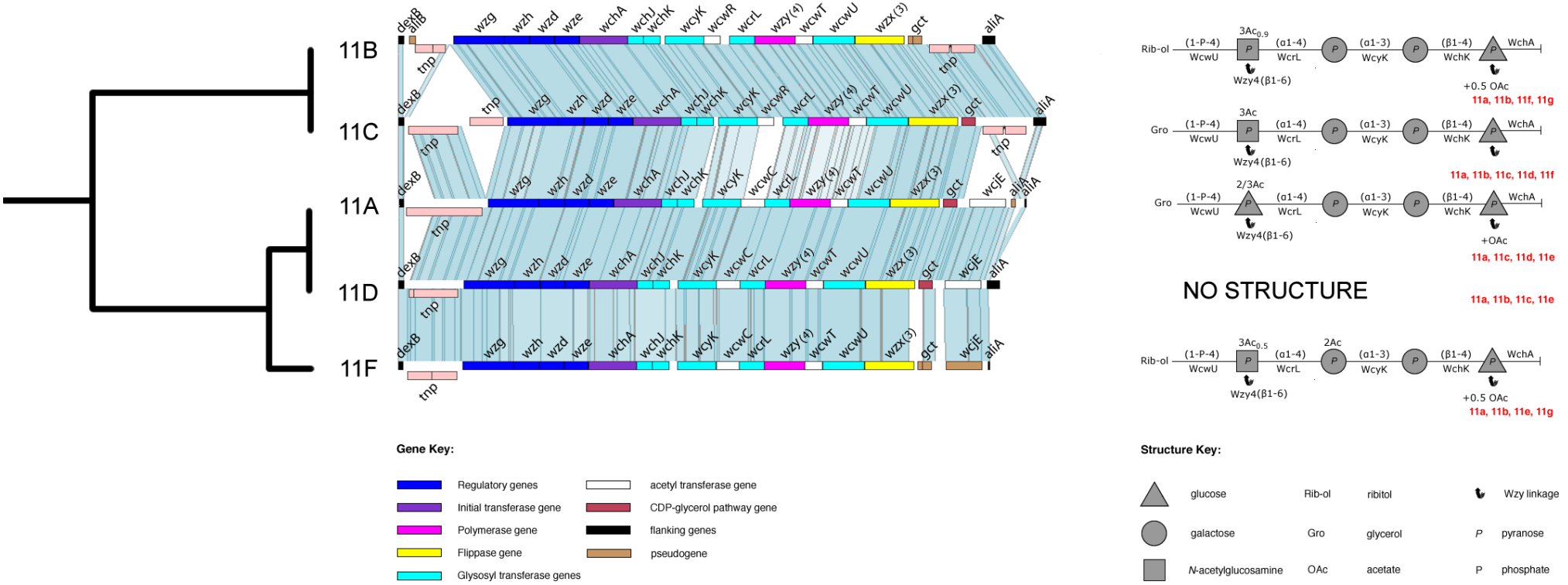


Figure 2.15 All serotypes within serogroup 11 fall into cluster 6. Light blue blocks between linear representations of the gene clusters indicate regions of sequence similarity. Wzy homology group membership is indicated on the *wzy* gene in parenthesis. Repeat-unit structures are represented with the initial sugar that is attached to the lipid carrier on the right and reactive sera are listed in red below each structure (if known).

2.4 Discussion

Initially utilising the *cps* sequences from the 90 *S. pneumoniae* serotypes, the aim of this chapter was to provide a framework for analysis applicable to large-scale comparative sequencing projects, which produce data that are present within a complicated evolutionary background. To delimit such data into more manageable units, a clustering approach was utilised based on the shared gene content between *cps* loci, refined to include a measure of sequence similarity between shared genes. In tandem, a web application containing all of the information was developed allowing the further investigation and visualisation of the fine scale similarities between loci.

Utilising this approach has allowed previously unseen relationships between genetic elements of *cps* loci, their corresponding biochemical structures and patterns of sera reactivity to be described. Prior to discussing the general applicableness of the analysis framework, the system-specific results will be discussed.

Within the pneumococcal community there is substantial interest in how (and how frequently) new pneumococcal serotypes arise and the analysis of all 90 *cps* loci does suggest that it may be relatively easy for the generation of new serotypes by a wide variety of mechanisms. The examples discussed have demonstrated that gene insertion or loss, single base differences introducing loss of function in encoded proteins or subtle differences in amino acid sequences can give rise to differences in serotype (i.e. differences in structure that can be recognised by sera). The production of different repeat-units composed of differentially linked combinations of sugars indicates that the potential for diversity within pneunococcal serotypes is immense. The provision of the 90 *cps* sequences together with their associated polysaccharide structures and serological profiles present an extensive dataset, the detailed analysis of which will allow a clearer understanding of capsule biochemistry, genetics and evolution.

The sequence-weighted clustering approach has proved a useful and robust framework for investigating the fine scale similarities between related serotypes. The initial use of a purely gene-content based approach was refined to include sequence similarities between shared HGs and, as discussed in the Methods section 2.2.7.2, provides a more realistic grouping than would be seen by gene content alone. Truncating the tree has identified groups of loci that can be seen as sufficiently similar to each other to

suggest that they have evolved from a recent common ancestor. The clustering also served as a basis for correlating gene content and function to the known CPS structures and their pattern of serological reactivity (for further discussion see Mavroidi et al., 2007).

In the majority of cases, functional inference can be undertaken with confidence (based on shared membership of a HG and knowledge of function of other members of that HG) such as the specificity of ITs, sugar pathway genes and the majority of transferases. However, assignment of specificity becomes somewhat more complicated when assigning the linkage of a glycosyltransferase as it has been shown that there are cases where even a single amino acid change can alter the specificity of a linkage produced (e.g. Mavroidi et al., 2005). Within the majority of HGs, the similarity between members is high, however in some cases members can be more divergent and this is a direct result of the e-value cut-off initially utilized within all vs all BLAST comparisons to define a HG (discussed further below). Due to the almost total lack of biochemical studies undertaken on gene products involved in pneumococcal CPS biosynthesis there is little option but to predict function based on sequence. It must be stressed however that the assignments of genes and gene products to specific linkages are tentative (for further discussion see Aanensen et al., 2007).

The 90 pneumococcal strains utilised for the *cps* sequencing were provided by the Statens Serum Institute and are those utilised in anti-sera production allowing confidence between the serological data and the sequence data. The 54 structures (Kammerling, 2000) were elucidated from different strains, from different studies, over a long time period, and as such it should be noted that anomalies may occur. This has become evident in two of the examples discussed where more recent, updated structural data has shown small differences in capsule structure (Lemercinier and Jones, 2006). However, the more recent data does correlate well with the observed sequence and serological data and as such provides a further confirmation of the results.

To address the relationships between any large datasets, it is imperative to break down the data into more maneagable units that can be investigated in fine detail. The sequence-weighted clustering approach has proved a useful and robust framework for investigating the fine scale similarities between related pneumococcal serotypes.

Utilising traditional sequence alignment methods, such as ClustalW and BLAST results loaded into ACT, allows further investigation, and as such, the sequence-weighted gene content tree, as opposed to a phylogenetic approach acted purely as a way of delimiting the dataset into more manageable closely related sets of data.

One of the issues that is not addressed utilising this approach is the presence of mosaic serotypes, e.g. the production of hybrids through large recombination events between ancestral serotypes, as these would not be recognised by the clustering approach. This issue warrants further investigation and also may become more pertinent when applying such a framework in other systems.

The development of the application, based on the pneumococcal *cps* sequences and related data is in some respects system-specific; however, if we regard a serotype as being a 'genomic unit' which could take many forms (from a discrete locus up to an entire genome) then the design of the system can be viewed, more generically, in relation to any such sequencing project. The database schema (section 2.2.6.1), can be seen as containing a set of 'core' and 'system-specific' tables. The core tables contain the sequence data and the results of sequence alignments and the system-specific tables contain information particular to the biological system, in the case of the *cps* data, those tables relating to known structures and serology. The application logic and visualisation of data is somewhat less system-specific. The HG methodology can be applied to any sets of multiple genomic units and HG membership, along with CDS annotation, can be stored in the database in a similar fashion to that presented. Likewise, the clustering approach can be utilized to produce a system level framework in the same manner. The web interface, including linear visualisation of genomic units, database interrogation and subsequent tree drawing, HG interrogation, WebACT and the overall data statistics, can all be applied to any set of annotated genomic units. This separation into core and system-specific units will become more important should the framework be applied to other systems. Utilising a 3-tiered architecture, the addition of further system-specific data tables along with software to analyse them becomes trivial and should allow established toolsets for dealing with system specific data to be incorporated within the system to closer link clustered sequence data with other lines of evidence.

TribeMCL provides the basis for clustering analysis, however there are certain aspects of this methodology that warrant consideration. The grouping of similar proteins is based on a defined BLAST 'e' value cut-off following all vs all protein alignments. The e-value of -40 was chosen based on the experience of our collaborators at the Wellcome Trust Sanger Institute. This value seems acceptable as a more stringent measure of homology (defining HGs using e^{-50}) resulted in identical HGs (data not shown). However, the overall amino acid similarity within a HG (based on summing all pairwise amino acid Needle comparisons and scaling based on the number of comparisons) differs considerably. HGs range from containing nearly 100% identical sequences to those that have overall similarities of roughly 50% and this figure is a direct consequence of the e-value used. This has an immediate impact on functional assignment within the *cps* sequences, as confidence in assignment where members of a HG are almost 100% identical is clearly much greater than when members have much lower levels of similarity, and also needs careful consideration when addressing other systems where more or less stringent criteria could be warranted.

The *cps* dataset, once received, was static, however to apply more generically it would be necessary, and indeed desirable, to add further sequence data and re-run TribeMCL. For example, other *cps* loci from closely related bacteria could be utilised to investigate cross-species HG membership and as such another factor that could influence HG groupings is the size of the sample. It could be envisioned that HGs that have more divergent members may 'break apart' should additional sequences be added, and preliminary investigation has shown this to be the case (data not shown). For the *cps* dataset, HG membership has been used as the criterion for gene naming and the potential division of HGs and gene naming will need to be addressed in further work. In some ways, the problems that have faced taxonomists at the species level that, particularly in bacteria, are being rigorously questioned (See Chapter 4), have parallels at the single gene level when attempting to place a name on a group of genes using any level of sequence similarity: a gene name can be as fuzzy as a species in these cases, however, refinements will come when samples are much bigger and definition of HG membership becomes more robust.

Another consequence of delimiting HGs at a defined cut-off is that there is no implicit measure of the similarity between different HGs. For a group of functionally similar,

but divergent enzymes, such as GTs, introducing a measure of 'between-HG' similarity may provide further evidence of gene duplication.

The *cps* system-specific data have been utilized to relate the genetic components of pneumococcal CPS biosynthesis to biochemical and immunological data, however, the apparent lack of system-specific tools focussed on the structural components does indicate an area for further investigation. The storage of carbohydrate data utilising the LINUCS format was not exploited outside of actual storage within the database. A more carbohydrate-centric analysis could therefore be undertaken. The clustering approach utilised for the *cps* loci could also be applied to the structural data, as each structure can be represented as a series of integers based on the sugar composition and/or the linkages between sugars. This would provide a parallel method of clustering structural data to investigate congruence between trees at different levels of the system, and this approach could be further exploited. Only 54 of the 90 CPS structures are known, however the complete genetic dataset and subsequent assignment of potential GT specificity presents the opportunity for a somewhat tentative prediction of the makeup (e.g. the constituents of the repeat-unit) of the unknown structures. Correlating the presence or absence of glycosyltransferase HG's with structural elements between serotypes could aid in such a prediction, however confidence limits would need to be addressed, as our current assignments are tentative.

All aspects of the analysis system have been built using server-side scripting, excepting the production of the dendrogram, which was undertaken using the software XLSTAT. Automation of this step could be undertaken using the statistical programming language 'R' and utilised on the web server so that tree generation can be undertaken automatically.

The system presented has proved essential to the analysis of the 90 pneumococcal *cps* loci, however, its more generic use will only be assessed after addressing other systems where sequences have a complicated evolutionary history and it is expected that further systems will provide refinements and modifications that will allow more generic, useful tools to be developed.

2.5 Further work

Many other bacteria contain clusters of genes involved in the production of polysaccharide structures and many of these have a similar genetic organisation to those found in *S. pneumoniae*. For example, closely related streptococcal species such as *S. mitis* and *S. oralis* have gene clusters, located in the same chromosomal location between *dexB* and *aliA*, that synthesise an external surface polysaccharide. Similarly, several other streptococcal species (e.g. S. *agalactiae and S. suis*) produce capsules by the Wzx/Wzy pathway and TribeMCL identifies some gene products of these pathways that fall in the HGs identified in the analysis of the pneumococcal *cps* loci (data not shown). Indeed the *S. mitis* and *S. oralis* clusters for external polysaccharide synthesis cluster with some of the pneumococcal *cps* loci and some even fall in the same sub-cluster as the pneumococcal serotype 21 *cps* locus (detailed in Mavroidi et al., 2007).

The Bacterial Polysaccharide Gene Database (BPGD) hosted at the University of Sydney, Australia (Reeves, 1996 [W31]) contains information regarding polysaccharide gene clusters and biochemical structures from a wide range of bacterial species. Structures are held as images taken from publications and gene cluster information is purely textual and no sequence information is stored within the database. However, GenBank identifiers are listed and link to the entries within the primary database. In parallel, the Bacterial Carbohydrate Structure Database (BCSDB) hosted at the N.D. Zelinsky Institute of Organic Chemistry, Moscow, Russia [W32] contains structure representations from many bacterial carbohydrates based on literature searches and manual insertion into the database using the LINUCS format mentioned previously. Further work is ongoing with Prof. Peter Reeves at the University of Sydney to combine these datasets and produce a global set of bacterial polysaccharide HGs, utilising the annotated DNA sequences linked to in BPGD.

There are other examples of sets of large sequences where there has been gene acquisition and loss, re-arrangements and probably ancient recombinational exchanges generating mosaic genomes. One good example is bacteriophage genomes, which often seem to share common features, however it has not been clear how to explore the relatedness between their genomes. I plan to look further at sets of phage that attack similar bacteria to explore whether sequence-based clustering can provide a more

rational way of showing their relatedness and if so will develop a web application for phage genomes. A preliminary analysis of prophage sequences within the genomes of *Streptococcus pyogenes*, *S. equi* and *S. zooepidemicus* was utilised by Dr. Matthew Holden at the Sanger institute and recently published in PLoS Pathogens on which both Brian Spratt and I were kindly included as authors (Holden et al., 2009).

Further application within phage biology will warrant careful consideration. It is well known from previously sequenced phage genomes that they have a modular structure consisting of discrete sets of genes, and as mentioned previously, the sequence-weighted clustering method does not take this into consideration. This is an issue that will need to be addressed and may warrant different ways of clustering sequences based on shared sets of genes as opposed to presence or absence of single HGs. Another issue that has presented itself is based on the fact that it is often difficult to annotate smaller CDS within phage genomes, as many of the genes present show no homology to any previously sequenced DNA. This has an immediate impact on any clustering based approach that utilises annotated sequences as its basis and as such provides an exciting challenge for applying the system.

The pneumococcal *cps* dataset consists of a single example of each of 90 strains assigned by sera profiling to a particular serotype. To address the evolution of individual serotypes it would be necessary to explore the patterns of sequence diversity among isolates of a single serotype and relate these to the levels of genetic relatedness and patterns of evolutionary descent of the isolates.

Such a study has been carried out by Mavroidi et al, (2006) who investigated the evolutionary history of the *cps* loci of 102 serotype 6A and 6B isolates. Sequence variarion within three *cps* genes (*wciP*, *wzx* and *wzy*) was related to the genetic relatedness of the isolates, as determined by multilocus sequence typing (MLST). By exploring inferred patterns of recent evolutionary descent (using the algorithm eBURST) they postulated the introduction of a single ancestral *cps* sequence, followed by slight divergence to create serotypes 6A and 6B. Furthermore, they demonstrated that a minority of 6B isolates had *cps* sequences (within the three genes) that were ~5% divergent from those of other 6B isolates and which may have arisen by a second, more recent introduction from a related but distinct source. Since this paper was published, Serotype 6C has been described by Park et al, (2008) and

consequently, these data need to be re-interpreted following identification of which strains belong to 6C.

Web applications used to explore the genetic relatedness of populations of isolates indexed by MLST and the eBURST algorithm are the subjects of the following chapter of this thesis.

Chapter 3: The Multilocus Sequence Typing Network: mlst.net

Available: http://www.mlst.net

Summary

The unambiguous characterisation of strains of a pathogen is crucial to a number of areas of infectious disease microbiology. Multilocus Sequence Typing (MLST), which defines strains from the sequences at several house-keeping loci, has become the method of choice for molecular typing of many bacterial and fungal pathogens (and non-pathogens). Sequence data can be stored and analysed easily within web applications allowing laboratories undertaking MLST to quickly progress from sequencing gene fragments to characterizing their strains and relating them to the population as a whole.

In this chapter I detail development of the gateway to a number of MLST schemes that are widely used by research and public health laboratories: mlst.net. Each species-specific site contains a set of tools for the initial characterisation of strains, and methods for relating query strains to other strains of the species. The first half of this chapter details the current version of mlst.net, and the second half, concurrent with the maturity of AJAX technologies detailed in the thesis introduction, details a BETA version, designed for greater efficiency and usability. Furthermore, a number of additional tools are introduced, including geographical interrogation of MLST databases using Google Maps and comparative eBURST, which provide users (and ourselves) with a richer toolset for the exploration of the populations held within each database.

Note: The application currently in use at mlst.net has been described in the following publication, a copy of which can be found in Appendix B:

Aanensen, D.M. and Spratt, B.G. (2005) The multilocus sequence typing network: mlst.net. *Nucleic Acids Res* 33(Web Server Issue):W728-33

3.1 Introduction

To address questions relating to the epidemiology, population and evolutionary biology of microbial populations it is vital to index genetic variation between strains. Molecular methods are widely utilised to characterise (or 'type') strains and this allows the relationships between isolates within different contexts (e.g. between different patients, hospitals or countries) to be addressed. Methods utilised are largely dependant on the questions being asked (for example local or global epidemiology) and the biology of the organism in question. However, in all cases the method chosen should be highly discriminatory such that those isolates assigned to a particular 'type' are likely to share a recent common ancestor.

Broadly, molecular methods used can be grouped into those that use DNA banding patterns and those that use precise digital data (e.g. alleles assigned from DNA sequences). A good example of the former approach is pulsed-field gel electrophoresis (PFGE), which assays the presence of unique recognition sequences (sites within the organisms genome where restriction enzymes cut the DNA). Due to the varying distances between these sequences between strains, restriction digests produce fragments of differing length. Gel electrophoresis allows the comparison of banding patterns based on these DNA fragments and provides the ability to 'barcode' the genome of the strains. The comparison of barcodes (i.e. the fraction of fragments in common) provides a distance measure that can be used for clustering to indicate relatedness between strains.

PFGE is widely used and can be applied to a wide range of bacteria but does suffer from important limitations. The comparison of banding patterns is less precise than comparison of DNA sequences, and this loss of fidelity is compounded when comparing banding patterns between laboratories. Pattern variation can accumulate rapidly in some species so those isolates which have descended from a common ancestor many decades ago may not be identifiable as minor variants of the same clone. The rapid accumulation of variation can be advantageous when looking at outbreaks, or local epidemiology, as it provides good levels of discrimination between strains. However, rapid accumulation is less useful for global epidemiology where assessing relatedness between strains from different decades or countries is required. Furthermore, the nature of the genetic variation, or the mechanisms by which it

changes, is often poorly understood which can limit its use for addressing questions of population and evolutionary biology.

Public health laboratories tend to focus on the strains that are causing disease within their own country. However, as disease is a global issue, molecular typing methods should allow precise digital comparisons of strains recovered locally with those from a wider area. Global epidemiological questions require the use of variation that accumulates slowly and, to satisfactorily address both strain characterisation and population biology, such variation should be selectively neutral (i.e. not under selective pressure) and the mechanism underlying the accumulation of variation should be well understood (Spratt et al., 2001). This has been addressed by using sequence variation in house-keeping genes (e.g. those that encode essential metabolic functions). Variation in house-keeping genes is generally considered to be selectively neutral and the origin of variation is relatively well understood. However, a single house-keeping gene, due to the slow accumulation of variation, does not provide enough discrimination to offer a satisfactory typing scheme and furthermore, any single locus typing scheme is sensitive to the effects of recombination.

One of the first techniques to index variation at multiple loci was multilocus enzyme electrophoresis (MLEE) which indirectly assays nucleotide sequence variation by indexing variation in the electrophoretic mobility of their gene products (Selander et al., 1986). MLEE assigns alleles at between 15-35 loci and each isolate is defined by a string of integers (the allelic profile) and each unique allelic profile is used to define the electrophoretic type (ET). Allelic profiles can be used to infer relationships between ETs using dendrograms based on a matrix of pairwise mismatches between profiles. MLEE has been used to successfully identify ETs (clones) and closely related ETs (clonal complexes) that are particularly likely to cause disease (Selander et al., 1987). However, as with PFGE, MLEE suffers from limitations. Comparison between laboratories is difficult to achieve and as much of the genetic variation is only indexed indirectly, many loci must be assayed to obtain sufficient discrimination between strains. Furthermore, given the widespread availability and ever decreasing cost of DNA sequencing indexing variation using electrophoretic mobility of gene products is no longer necessary and as such MLEE is rapidly becoming a legacy technique.

Multilocus sequence typing (MLST) adapts the proven concepts of MLEE but rather than using the inference of genetic variation achieved using MLEE, variation at multiple house-keeping loci is directly indexed by nucleotide sequencing (Maiden et al., 1998).

MLST involves obtaining the sequences of internal fragments of (normally) seven house-keeping genes for each strain of a particular species. The sequences of each fragment are compared with all previously identified sequences (alleles) at that locus and are assigned allele numbers. The combination of the seven allele numbers (representing each locus) defines the allelic profile of the strain and each different allelic profile is assigned as a unique sequence type (ST).

Nucleotide sequencing is relatively cheap and easy to perform and the data produced by MLST are ideal for the characterisation of strains of microbial species via the internet. Data can be stored and accessed from web accessible databases and as such, strains studied by different groups and corresponding isolate epidemiological data can be compared easily. MLST is a widely used molecular typing method both within research and public health laboratories and MLST schemes have been developed for almost 50 bacterial (mostly pathogenic) and several fungal species (Bougnoux et al., 2004; Dodgson et al., 2003). As such, MLST databases and associated websites are an important resource and have been available for a number of years (Chan et al., 2001).

MLST databases are currently hosted on four main web servers located at Imperial College London [W33], Oxford University [W34], the University of Cork [W35] and the Pasteur Institute [W36]. Each server provides a website which acts as a gateway to different organism-specific MLST databases. Each website offers similar core functionality for the querying and assigning of alleles and sequence types and also for interrogating epidemiological information deposited for a growing number of strains. However, the accessory functionality offered at each site can differ.

The subject of this chapter is the organism-specific databases/web applications located at Imperial College London (http://www.mlst.net), which I develop and manage. Two versions of the mlst.net software are detailed which I shall refer to as ‘mlst.net v1’ and ‘mlst.net v2’ for clarity. mlst.net v1 (Aanensen and Spratt, 2005) represents a complete rewrite of the MLST software produced by Chan et al, (2001),

however full acknowledgment must be given as this formed the framework for much of the application logic. Re-development was undertaken to increase flexibility and the underlying database structure changed to reduce redundancy. This version of the software is currently in use and described in the first part of this chapter. mlst.net v2 has recently been developed and is currently in BETA testing with a number of users and will be described in the second half of this chapter.

Concurrent with the introduction of web development technologies such as AJAX (detailed in the thesis introduction) mlst.net v2 takes advantage of some of the more novel web development methodologies to offer increased accessory tools whilst preserving the core functionality described for mlst.net v1.

3.2 mlst.net v1

The mlst.net homepage acts as a gateway to a number of organism-specific sub-sites each accessible at a defined URL in the format http://*organism*.mlst.net. (see Table 1). Laboratories undertaking MLST can access species-specific information, including sequencing protocols and primer sequences, allowing a laboratory to begin producing data rapidly. Characterisation of a strain requires the generation of the sequences of the seven gene fragments and, once these are available, they are used to query the appropriate mlst.net sub-site, to assign the alleles at each locus and thereby to obtain the allelic profile of the query strain.

Organism	**URL**	**Reference**
Borrelia burgdorferi	http://borrelia.mlst.net	Margos et al., 2008
Burkholderia pseudomallei	http://bpseudomallei.mlst.net	Godoy et al., 2003
Candida albicans (legacy)	http://test1.mlst.net	Bougnoux et al., 2002
Candida albicans (consensus)	http://calbicans.mlst.net	Bougnoux et al., 2003
Candida glabrata	http://cglabrata.mlst.net	Dodgon et al., 2003
Cryptococcus neoformans	http://cneoformans.mlst.net	Meyer et al., 2009
Enterococcus faecalis	http://efaecium.mlst.net	Ruiz-Garbajosa et al., 2006
Enterococcus faecium	http://efaecalis.mlst.net	Homan et al., 2002
Haemophilus influenzae	http://haemophilus.mlst.net	Meats et al., 2003
Leptospira spp.	http://leptospira.mlst.net	Thaipadungpanit et al., 2007
Staphylococcus aureus	http://saureus.mlst.net	Enright et al., 2000
Staphylococcus epidermidis	http://sepidermidis.mlst.net	Thomas et al., 2007
Streptococcus pneumoniae	http://spneumoniae.mlst.net	Enright and Spratt, 1998
Streptococcus pyogenes	http://spyogenes.mlst.net	Enright, 2001
Streptococcus suis	http://ssuis.mlst.net	King et al., 2002

Table 3.1 Organism-specific databases hosted at mlst.net.

3.2.1 mlst.net database schema

Each species-specific website links to a database which stores the sequences of all known alleles at each of the seven loci, all known allelic profiles and epidemiological information for strains submitted. The schema for each underlying database is similar to that seen in Figure 3.1 (which is the schema for the *S. pneumoniae* MLST database).

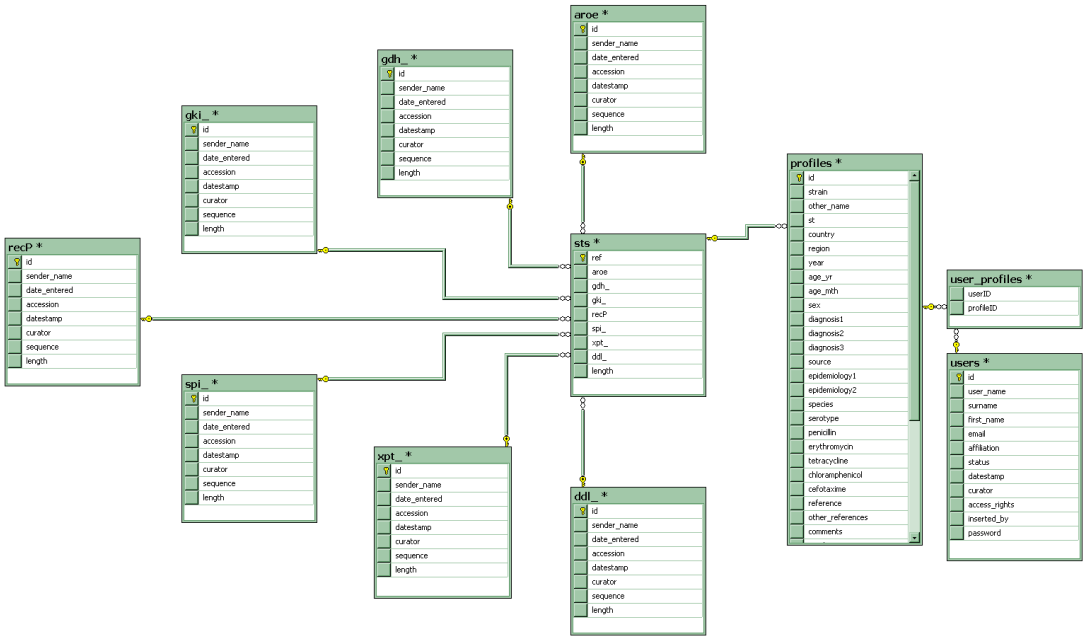


Figure 3.1 mlst.net v1 database schema.

The locus tables house sequence data representing alleles. The allele number is represented in the table with the primary key 'ID', and stored along with details regarding the submitter, date of assignment and the raw DNA sequence as a string.
The 'STs' table contains a list of all known allelic profiles. The primary key 'ST' is the Sequence Type and seven columns are present containing the allele numbers at each of the seven loci (except for schemes with fewer; e.g. *C. albicans* or greater; e.g. *B. burgdorferi* than seven loci). Each allele number references the corresponding locus table ID.

Epidemiological information pertaining to isolates of a particular ST are held within the 'profiles' table. The primary key 'ID' is assigned to any new isolate entered into the database and, along with epidemiological information, date stamps are recorded to indicate when the isolate was deposited. The 'ST' column of the profiles table refers

to the ‘ST’ column of the STs table, which in turn refers to the ‘ID’ column(s) of the locus tables.

3.2.2 Curator’s interface

A curator’s interface is available and password protected. A single (or multiple) curator(s) are assigned to each database and as such act as the first point of contact for a user wishing to enter a new allele / ST / isolate. The curator’s interface is web-based, and therefore curator’s can be located anywhere and can actively update the database with little intervention. To assign a new allele we request that the laboratory submitter sends the forward and reverse sequence trace file(s) for quality control. As any new allele will produce a new ST we also request at least one example of epidemiological information for an isolate with a new ST. Likewise, for any new ST generated from a new combination of existing alleles. Each data submitter has a corresponding record in the database ‘Users’ table (Figure 3.1), and the ownership of each isolate is recorded in the ‘user_profiles’ table.

3.2.3 User’s interface

For a user, each MLST website offers a number of analysis steps. First, alleles can be assigned from the users sequence data by one of three options (Figure 3.2 A and D):

- *Single/Batch Locus Query*: allowing a single sequence or a batch of sequences for a single locus to be compared with all known alleles.
- *Multiple Locus Query*: allowing the input of the sequences of all seven loci for a single strain.
- *Batch Strain Query*: allowing input of the sequences of all seven loci for a batch of strains.

In all cases, the user’s sequence is checked for correct length for that locus, and for the absence of unexpected characters, and is then queried against all other sequences in the corresponding locus table for that species. For any diploid organism (e.g. *C. albicans*), the standard IUPAC ambiguity codes are allowed and are used to assign heterozygous nucleotide sites (Bougnoux et al., 2004).

If the query sequence is found within the corresponding locus table, the allele number is returned, whereas if the query is a novel sequence, the percentage identity to the closest allele in the database is returned and the user is advised to check carefully those nucleotide sites that differ from the most similar allele or alleles in the database (Figure 3.2B). This can be carried out using the Jalview alignment editor (Clamp et al., 2004), which displays the query sequence aligned with all other allele sequences for the corresponding locus. Alternatively, the nucleotide differences can be displayed between the query sequence and the most similar alleles (Figure 3.2C). The latter method allows the user to view the flanking sequence around each nucleotide difference between the query and the most similar alleles, allowing the trace files of their proposed new allele to be searched easily for any potential ambiguities or sequencing errors. If a user is confident that they have a new allele, the forward and reverse trace files are submitted to the MLST curator before a new allele number for the novel allele is assigned and entered into the database, by the curator.

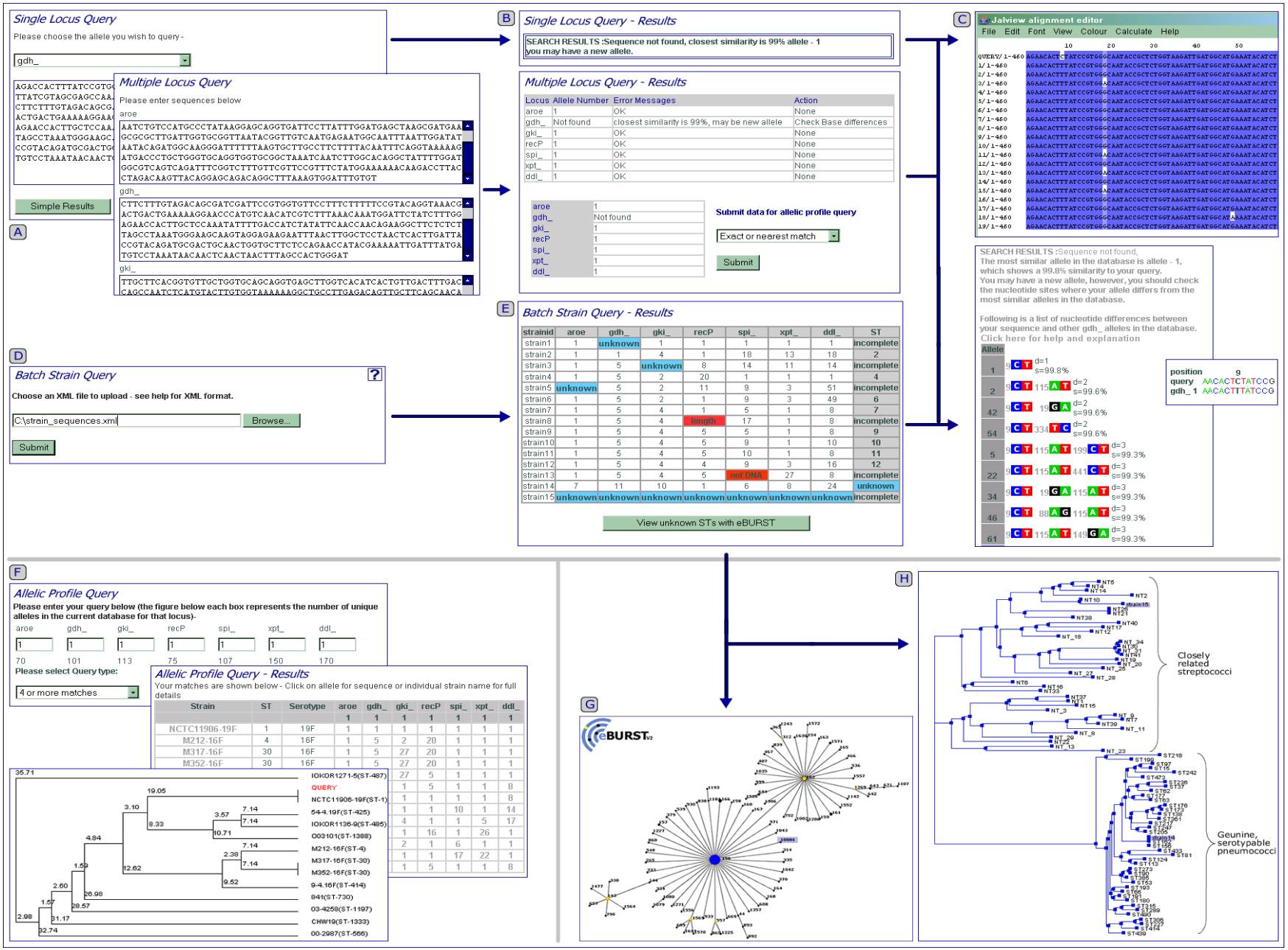


Figure 3.2 MLST Workflow: mlst.net v1.

Repeating this process for each locus provides the seven digit allelic profile for the query strain. The seven digit allelic profile can then be entered into the allelic profile query to discover if the strain is identical, or similar, to any of the strains already in the database (Figure 3.2F). The multiple locus query represents a batch processing method for a single strain, allowing all seven sequences for the query strain to be entered at once and for the allele numbers at each locus to be returned.

The repeated querying of single strains can become time consuming for laboratories undertaking MLST on many hundred strains of a particular species. The need to analyse sequences from multiple genes from a large number of strains at the same time precludes the use of standard sequence formats such as FASTA of MEGA. Therefore a simple XML format can be utilised that allows the batch processing of many strains at one time (Figure 3.3).

```
<dataroot>
	<strain_sequences>
		<strainid>Strain1</strainid>
			<locus1>…GCCGTGA…</locus1>
			<locus2>…GCCGTGA…</locus2>
			<locus3>…GCCGTGA…</locus3>
			<locus4>…GCCGTGA…</locus4>
			<locus5>…GCCGTGA…</locus5>
			<locus6>…GCCGTGA…</locus6>
			<locus7>…GCCGTGA…</locus7>
	</strain_sequences>
	<strain_sequences>
		<strainid>Strain2</strainid>
			<locus1>…GCCGTGA…</locus1>
			<locus2>…GCCGTGA…</locus2>
			<locus3>…GCCGTGA…</locus3>
			<locus4>…GCCGTGA…</locus4>
			<locus5>…GCCGTGA…</locus5>
			<locus6>…GCCGTGA…</locus6>
			<locus7>…GCCGTGA…</locus7>
	</strain_sequences>
</dataroot>
```

Figure 3.3 Simple XML Format used for Batch Strain Query.

As mentioned in the thesis introduction, XML allows one to define TAGS to describe data. This allows scripts to be written which will parse the XML much more easily than plain text.

Formatting the input data with this basic XML syntax for a set of seven sequences for each strain allows a user to produce a file, for an unlimited number of strains, which can be used for batch processing. To aid production of such a format, each of the MLST species-specific sites at mlst.net provides a modified MS Access database that

allows users to store their sequences and strain information in one place, and allows the data to be exported in bulk in the correct XML format, without the need to manually produce the document. Furthermore, for sequencing laboratories utilising the STARS [W37] platform for MLST, a script allows converting the FASTA files generated by STARS into the XML format via a web form.

When a user uploads the generated XML file (figure 3.2D), the sequences for each of the seven loci in all of the strains are checked for invalid characters and correct length. Each sequence is then queried against the appropriate locus table in the species database (Figure 3.2E). If found, the allele number is returned and, if unknown, the user can look further into the sequence differences between the query allele and the most similar alleles in the database (Figure 3.2C). If all the seven loci are found, the allelic profile of the strain is queried against the ST table and, if a match is found, the corresponding ST is returned. If the allelic profile is not found in the database, this information is returned and the user is encouraged to contact the curator. The batch procedure, therefore, automatically returns a table with the alleles, allelic profiles and STs of all the input strains, flagging up those alleles and STs that are previously unknown (Figure 3.2E).

3.2.4 Clustering MLST data

Depending on a user's question, it is highly likely they will wish to relate their own strains to those within the database to assess their similarity to other members of the population. Several methods are provided.

3.2.4.1 Comparing query strains to the batabase: Allelic profiles

The simplest approach within MLST datasets is to identify those strains in the database that have some minimum level of similarity in their allelic profile to each query strain (e.g. sharing alleles at ≥4 of the seven loci), and to show the relationship of the query strain to those returned from the database query with a dendrogram based on the matrix of pairwise differences between the allelic profiles of the strains. This can be achieved using a purpose-developed module of Jalview (Figure 3.2F).

3.2.4.2 Comparing query strains to the database: eBURST

Traditionally, dendrograms have been the method of choice for displaying the implied relationships between strains of a bacterial population or species. However, although dendrograms are good at visualising clusters of identical or very similar strains, the bifurcating process of lineage splitting implied by a dendrogram is a very poor representation of the way in which bacterial lineages emerge and diversify. To address this, an interactive implementation of the BURST (Based Upon Related Sequence Types) algorithm, eBURST, has been integrated within the mlst.net sites (Feil et al., 2004; Spratt et al., 2004). The first version of eBURST was developed by Bao C Li and the current version, eBURSTv3, by Derek Huntley. I have implemented the connectivity within mlst.net.

The algorithm underlying eBURST does not impose a tree-like pattern of descent, but rather uses an appropriate model of recent bacterial evolution. In addition, it is very difficult to display the relatedness of all strains in a large MLST database, including thousands of STs, on a dendrogram and better ways of visualising the relationships among all strains in large MLST databases are therefore required.

Briefly, the model incorporated into eBURST assumes that, due to selection or genetic drift, some genotypes will occasionally increase in frequency in the population and will then gradually diversify by the accumulation of mutation(s) and/or recombinational replacements, resulting in slight variants of the founding genotype. Initially, members of this emerging clone will be indistinguishable in allelic profile by MLST, however with time, the clone will diversify to produce a number of variants in which one of the seven MLST loci has been altered: single locus variants (SLVs). Further diversification will produce variants of the founder ST that differ at two out of the seven loci: double locus variants (DLVs). In this simple model, bacterial populations will consist of a series of clonal complexes (sets of variants of a founding genotype) that can be recognised from the allelic profiles of the strains within a MLST database (Feil et al., 2004; Spratt et al., 2004)

eBURST can be used to explore the relationships of newly characterised strains to those in the database (see Fig 3.2G). eBURST uses the STs and their associated allelic profiles as input and, using the default settings, divides the strains into groups in

which all STs in the same group share ≥6 out of 7 loci with at least one other member of that group, resulting in non-overlapping groups or clonal complexes. The predicted founding genotype is that which has the greatest number of SLVs within a clonal complex and eBURST represents the number of isolates of each ST by the size of the circles representing each ST. Of particular value is the ability to display all the STs in a large MLST database in a single diagram (a population snapshot; Figure 3.4), showing all the major and minor clonal complexes, and individual STs that are relatively distantly related to all other STs.

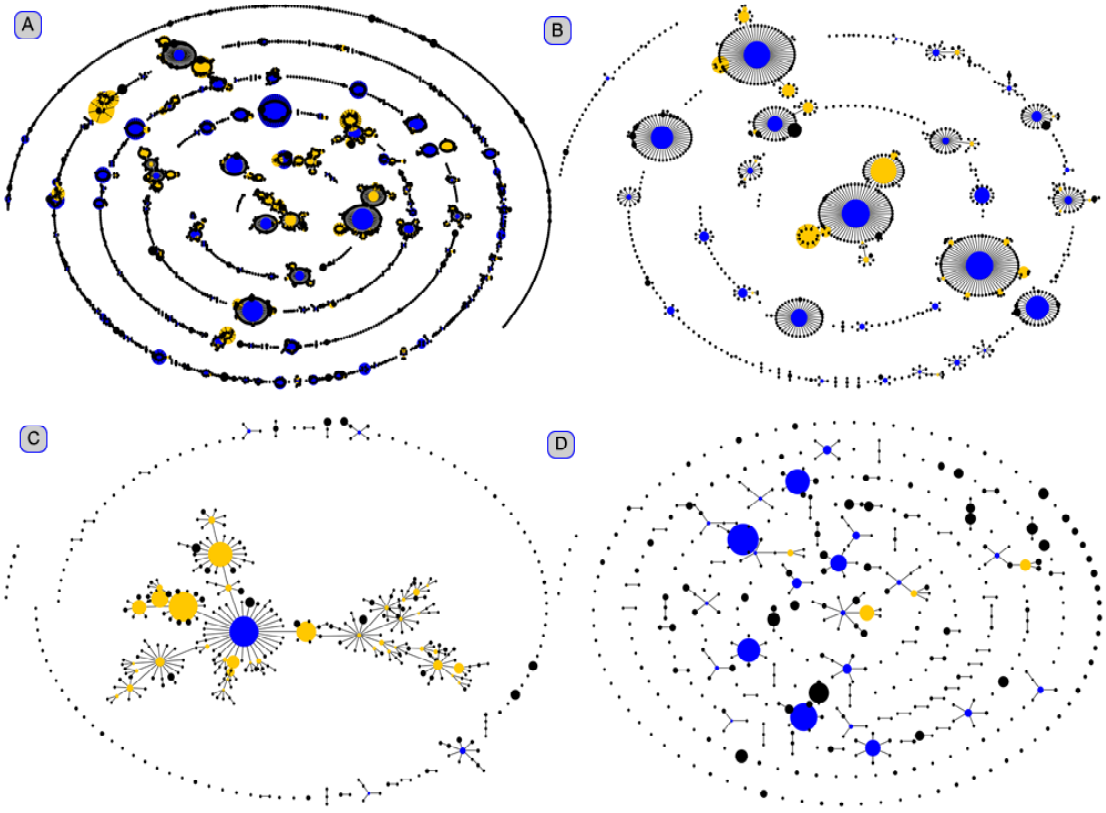


Figure 3.4 eBURST population snapshots of the *Streptococcus pneumoniae* (A), *Staphylococcus aureus* (B), *Enterococcus faecium* (C) and *Streptococcus pyogenes* (D) MLST databases.

It should be stressed that eBURST provides a hypothesis about the ancestry and patterns of recent descent of strains within a population. The different structures observed in population snapshots can be largely attributed to the different signatures of recombination / mutation rates within the evolution of a particular species. This should always be considered when drawing strong conclusions from such hypotheses. Further details are beyond the scope of this thesis, however, the reliability and limitations of eBURST have been tested by Turner et al., (2007). In this study, the

authors used simulated MLST data, with known ancestry, to test the effect of different population recombination and mutation rates on the reliability of eBURST. Their study has provided useful guidelines for users when interpreting the patterns of clustering provided by eBURST.

3.2.4.3 Comparing query strains to the database: Concatenated sequences

The ability to concatenate the sequences at the seven loci, maintaining the correct reading frame, and to construct a neighbor-joining tree based on these sequences is provided, but needs to be used with considerable caution. A module from MEGA (kindly provided by Sudhir Kumar and Joel Dudley; Kumar et al., 2008) produces an alignment, a distance matrix and the tree topology in NEWICK format, which is then displayed via the website using the ATV Java applet (Zmasek and Eddy, 2001). Allelic changes at the MLST loci will occur (to a varying degree depending on the species) by recombination, and in many cases the relative contribution of recombination and point mutation to the diversification of strains will be unknown (Feil et al., 2001). A long history of recombination will preclude the recovery of the true phylogenetic relationships between distantly related bacterial strains and even the relatedness between similar strains may be better represented on a tree based on differences in allelic profiles than one based on differences in the concatenated sequences. However there are specific issues that can be usefully addressed by using the concatenated sequences.

The *B. pseudomallei* database includes strains of closely related species and the *B. pseudomallei* website provides a facility to examine the position of a query strain on the tree constructed using concatenated sequences. The position of the query sequence on the tree can establish whether the query strain is *B.pseudomallei* or something similar to, but distinct from *B.pseudomallei* (Godoy et al., 2003). Similarly, there is considerable confusion about whether strains that appear to be *S. pneumoniae*, but which cannot be assigned a pneumococcal capsular serotype, are authentic pneumococci that either do not produce a capsule or are members of a similar but distinct streptococcal population. The pneumococcal MLST website has a facility to examine whether a query strain clusters within a reference set of *S. pneumoniae* strains, or with the related population, using a tree based on concatenated sequences, which can resolve this issue in most cases (Figure 3.2H). Trees based on concatenated

sequences may also be useful for assigning *H. influenzae* strains to major lineages (Meats et al., 2003), or for *S. aureus* (Feil et al., 2003) where recombination appears to be rare.

3.3 mlst.net v2

mlst.net v1 is widely used by research communities and public health laboratories to investigate the epidemiology, population and evolutionary biology of the organisms listed in Table 1. Data is submitted frequently and with ever increasing access to, and the falling costs associated with, DNA sequencing, deposition rates are expected to continue to increase until the procedure is superceded by new approaches to strain characterisation.

As service providers (and researchers), we have a responsibility to develop methods of analysis and data visualisation and to assess and implement well-proven software methodologies for the benefit of the user communities (and our research). Any new methods/facilities should provide benefit to users and their (and our) understanding of the data held within the databases. In many respects, methods should ideally be organised in a way that pre-empts the steps in a users workflow, prompting them towards the analysis required to answer their questions.

Developments under the umbrella of 'Web 2.0', described in the introduction to this thesis, are providing new and more sophisticated methods for building web applications and provide the tools to increase usability. Access to varying different data sources and software using APIs (such as cartographic data provided for web mapping) facilitate methods of visualising and interacting with data, which were not possible when mlst.net v1 was developed.

To take advantage of these techniques mlst.net v1 has been re-developed in several ways, grouped broadly as follows:

1) Interface re-development.
2) Integration of mapping tools to visualise and interact with the databases.

3) Increased integration and the introduction of comparative and geographical eBURST.
4) Additional features.

It should be stressed that, as with any major change to a facility routinely accessed by a community of users, one has to be certain of providing tangible benefits whilst not disrupting current practices. It is with this in mind that mlst.net v2 has been released as a BETA version, initially for the *S.pneumoniae* database, to allow users to assess and provide feedback on new features. To adhere with common practice, both versions will be accessible in parallel for an indefinite period of time. During this time the advantages and disadvantages of each version will be assessed (by ourselves and users) and appropriate amendments made. mlst.net v2 is available at http://spneumoniae.mlst.net/beta

3.3.1 Interface re-development

The database structure for mlst v2 remains unchanged as in Figure 3.1 excepting the addition to the profiles table of two columns, 'latitude' and 'longitude', allowing more precise geographical information to be recorded for an isolate.

A single new database that all organism specific mlst.net sub-sites can link to provides basic geographical information including the latitude / longitudes of the centre point (centroid) for all countries and cities, should this information be required. This data was retrieved using geocoding facilities available at our site http://www.spatialepidemiology.net, which is the subject of, and will be discussed (along with more in-depth details of web mapping) in chapter 5 of this thesis.

mlst.net v1 is an example of a page-click-page web site adhering to the common, three tiered-architecture. As discussed in the thesis introduction, the principles of AJAX web development can provide a more efficient method of content delivery and usability and these principles are applied within mlst.net v2.

To demonstrate, a simple, but common workflow for a user who has sequenced all seven MLST loci is as follows: Using the 'multiple locus query', enter seven sequences to discover allele numbers (if known), then enter the allele numbers into

the ‘allelic profile query’ to discover the ST (if known) and view similar STs found in the database.

Illustrative versions of this workflow using the application logic underlying both mlst.net v1 and mlst.net v2 can be seen in Figure 3.5. Using mlst.net v1 a user pastes query locus sequences into the locus query form (Figure 3.5A), then clicks submit. Query sequences are sent, via HTTP, to the web server to be processed. The allelic profile (if found) is returned, formatted within HTML as a new web page, which includes the HTML representing the initial form for further querying (Figure 3.5B). The user then navigates, using the menu, to the ‘Profile Query’, which loads a new page, requested from the server, containing the profile query form (Figure 3.5C). Entering the allelic profile and clicking submit sends this data to the server and a new HTML page is loaded containing the ST, and similar STs (if found). In each step of the workflow a new page is loaded, including the images and menu HTML.

Using AJAX (within mlst.net v2), a single page is loaded once (Figure 3.5E) and within the HEAD of the HTML document a single file is linked containing a set of JavaScript functions (Figure 3.5G). The initial page loaded contains the HTML corresponding to both forms embedded within ‘DIV’ elements. However, when the page loads the JavaScript function ‘`onLoad()`’ is triggered by the web browsers JavaScript engine. This function causes the DIV containing the Locus Query form to be displayed and the DIV containing the profile query form to be hidden. This presents to the user a view identical to that seen in mlst v1 (Figure 3.5A).

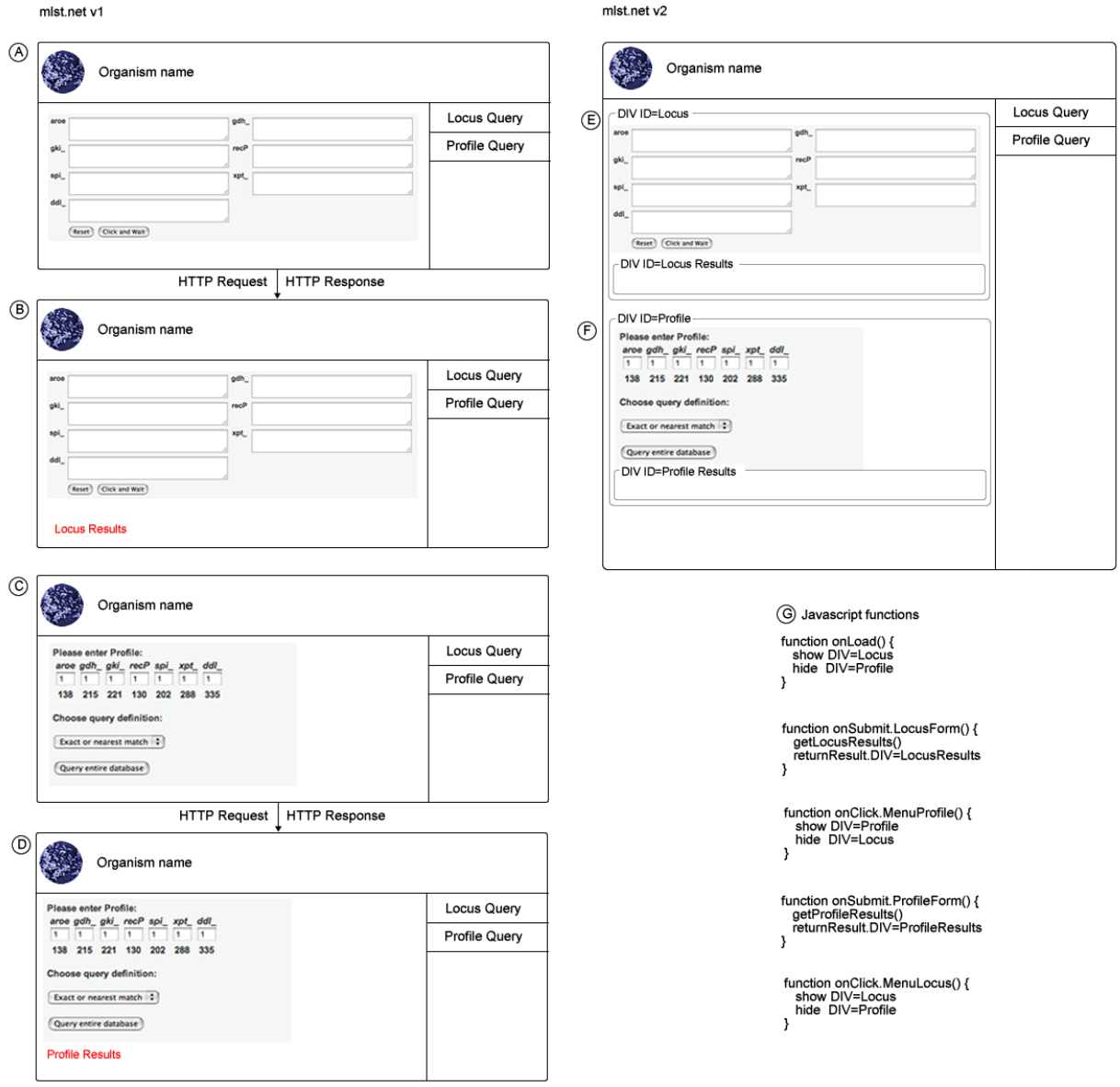


Figure 3.5 Application Logic : mlst.net v1 (page-click-page) versus mlst.net v2 (AJAX).

When pasting sequences in to the form and clicking submit, the JavaScript function '`onSubmit.LocusForm()`' is triggered which submits the sequences to the server using the *XMLHttpRequest* DOM object (as discussed in the thesis introduction). Instead of an entirely new page, as in mlst.net v1 (Figure 3.5B), a simple XML result is returned, asynchronously, to JavaScript in the format '<xml>1,2,4,2,3,4,5</xml>' (in this case all sequences are identified, error messages are returned if not). This returned result is then used to update the 'LocusResults' DIV. The user is presented with a similar view to that in mlst.net v1 (Figure 3.5B) and the only extra data loaded into the browser is the XML, rather than an entire page.

Using the navigation menu, the user can now select the 'Profile Query'. In mlst.net v1 this loads an entire page containing the profile query form (Figure 3.5C), however in mlst.net v2, when clicking 'Profile Query' the JavaScript function '`MenuProfile()`' is triggered (Figure 3.5G) which hides the 'Locus' DIV (still containing the results and pasted sequences) and displays the 'Profile' DIV. The user is now presented with a view identical to that in Figure 3.5C. Submitting the form triggers the '`onSubmit.ProfileForm()`' JavaScript function which submits the user's allelic profile via *XMLHttpRequest* and the results are returned in simple XML, but this time the 'Profile Results' DIV is updated (with a view identical to Figure 3.5D).

In this fashion the amount of HTML exchanged between client (web browser) and server is kept to a minimum and much of the application logic, which in mlst.net v1 is entirely implemented within server-side scripts, is handled by the web browser's JavaScript engine. Furthermore, after submitting the profile query, should the user discover that they had a single base mistake in one of their allele sequences, they can return to the Locus Query form by clicking the 'Locus Query' menu option (which triggers the '`onClick.MenuLocus()`' function, hiding the Profile DIV and displaying the Locus DIV) and update the sequences: the initially pasted query sequences and subsequent result will still be present in the form text boxes (as the Locus DIV was only hidden). Within mlst.net v1 this would require the user to re-load the initial form page (Figure 3.5A) and re-paste their query sequences.

This example is a simplified workflow, however the entire navigation logic of an MLST site can be presented in a similar fashion, allowing much of the 'hard work' of site navigation to be undertaken via JavaScript. Furthermore, the minimal data exchange between browser and server once the application has loaded, combined with the persistant presence of data entered by the user presents a much richer experience (although the actual impact from a users perspective remains to be assessed fully).

3.3.2 Integration of Mapping tools for viewing MLST databases

Isolate data submitted to mlst.net is inherently geographical in nature. A global molecular typing tool that allows the submission of user sequences and epidemiological information about isolates should be able to offer intuitive methods to place such data (and allow subsequent interrogation) in its geographical context. The potential for relating strains geographically through an intuitive user interface has not been fully realised within current implementations of MLST software (mlst.net and other sites) and historically, the majority of MLST databases allow the deposition of isolate origin limited to country and, in some cases, city or town. Sampling biases (e.g. many isolates being submitted from one hospital which covers a large geographic area) undoubtedly play a part in this trend. However, through providing and demonstrating the benefit of tools that allow better geographic visualisation / interrogation of MLST databases, our aim is to encourage the sampling and submission of isolate data at higher geographic resolution.

A sub-domain of mlst.net, MLST-Maps (http://maps.mlst.net) was developed and made publicly available in June 2007. MLST-Maps focuses on the geographical interrogation of globally deposited STs from multiple MLST database in a slightly more basic form than discussed below. MLST-Maps will be discussed further in chapter 5 and the description of mapping tools below pertains to those utilised in mlst.net v2.

Web mapping API's, discussed briefly in this thesis introduction, and discussed in greater detail in chapter 5, provide the ability for a developer to embed an interactive map (consisting of cartographic images) within a webpage and, using JavaScript, display data on the map. The principles of AJAX are extensible and through writing custom JavaScript code, additional application logic can be implemented within the client browser to add interaction with maps catered to a developers (and therefore users) needs.

To present the data held in MLST databases geographically the Google Maps API was chosen due to its greater flexibility than similar initiatives (see Chapter 5 for justification). The Google Maps API [W50] allows data specified by a developer to be

displayed as points on a map. For the purposes of mlst.net v2, a number of JavaScript functions were written to extend the API allowing integration with MLST databases.

A user may wish to ask the simple question 'Which STs are found in my country and where else are these STs found?' Within mlst.net v1 this would be undertaken using the 'advanced query'. Firstly, the country of interest would be searched for, and a list of STs returned. Further 'advanced queries' would need to be undertaken (one by one) and results manually collated to give an answer. This is a time-consuming and unsatisfactory procedure and furthermore relies solely on textual output for a user to interpret.

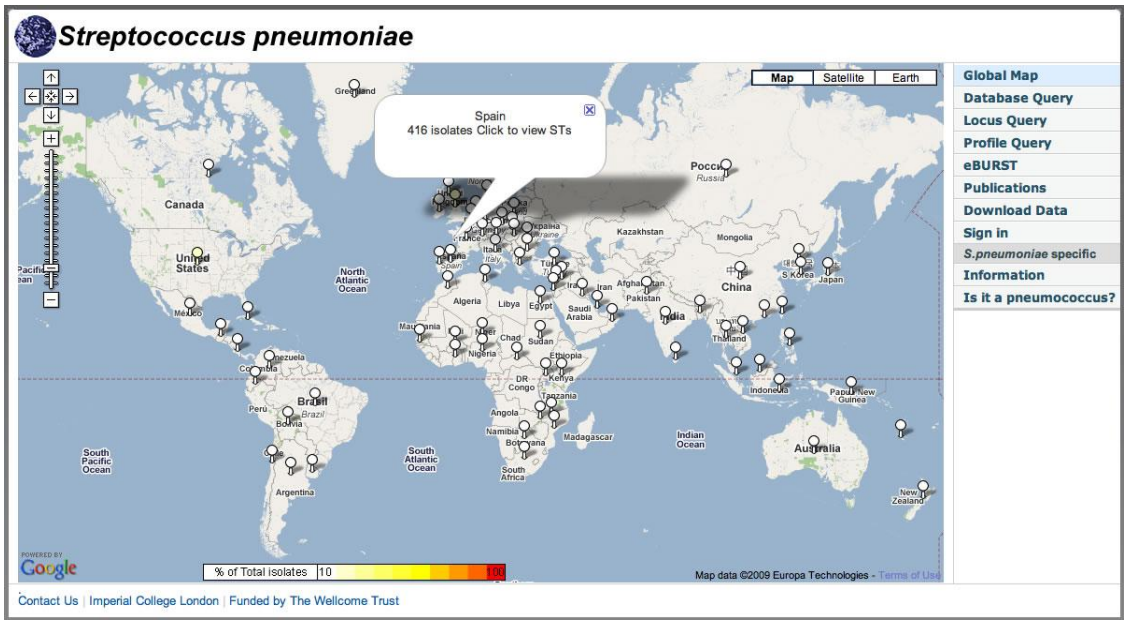


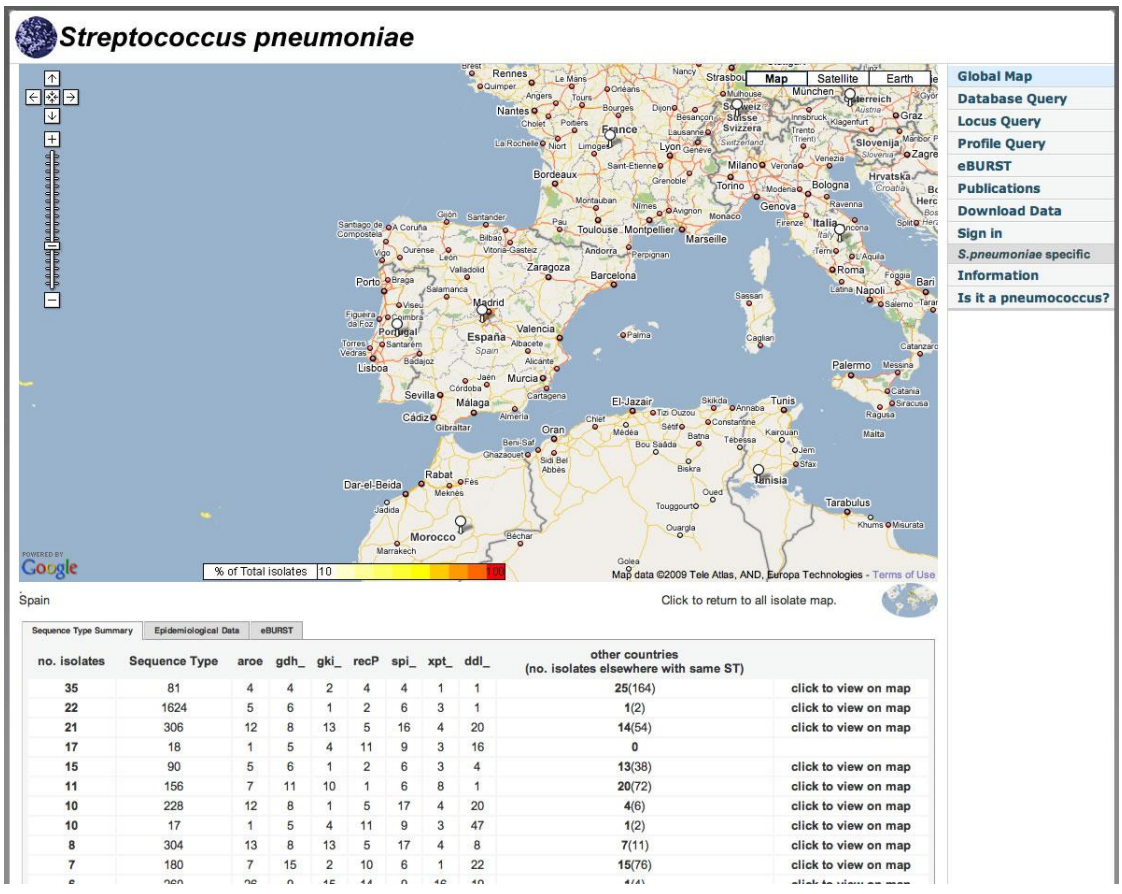
Figure 3.6 mlst.net v2: Initial display.

Within mlst.net v2 (http://spneumoniae.mlst.net/beta), the initial view, presented to a user, displays a global map containing 'placemarks' indicating any country from where an isolate(s) epidemiological data have been deposited in the database (the placemark is located at the centroid of that country). The colour of the placemark represents the % of total isolates within the database that have been deposited from that country and a key is positioned at the bottom of the map. The navigation menu (similar to that described in section 3.3.1) can be seen on the right hand side of the screen (Figure 3.6).

When clicking on a placemark for a particular country (e.g. 'Spain') an 'Info Window' appears indicating the number of isolates within the database from that

country and a message to ‘click to view STs’ (Figure 3.6). On clicking the link, the map zooms in to Spain and a ‘loading’ icon appears below the navigation menu indicating that the data is loading. When complete, the ‘Sequence Type Summary’ detailing all isolate STs found in Spain appears below the map window (Figure 3.7).

STs are listed based on prevalence (within Spain) and data returned includes the number of isolates with a particular ST, the ST and accompanying allelic profile, the number of other countries (and number of isolates in other countries with that ST) and, should that ST be found elsewhere, a link (‘click to view on map’) to view the global distribution of that ST on the map. We can see that at the time of writing ST18, which is the fourth most recorded ST in Spain, is not found elsewhere.



no. isolates	Sequence Type	aroe	gdh_	gki_	recP	spi_	xpt_	ddl_	other countries (no. isolates elsewhere with same ST)	
35	81	4	4	2	4	4	1	1	25(164)	click to view on map
22	1624	5	6	1	2	6	3	1	1(2)	click to view on map
21	306	12	8	13	5	16	4	20	14(54)	click to view on map
17	18	1	5	4	11	9	3	16	0	
15	90	5	6	1	2	6	3	4	13(38)	click to view on map
11	156	7	11	10	1	6	8	1	20(72)	click to view on map
10	228	12	8	1	5	17	4	20	4(6)	click to view on map
10	17	1	5	4	11	9	3	47	1(2)	click to view on map
8	304	13	8	13	5	17	4	8	7(11)	click to view on map
7	180	7	15	2	10	6	1	22	15(76)	click to view on map

Figure 3.7 Sequence Types from Spain displayed within mlst.net v2.

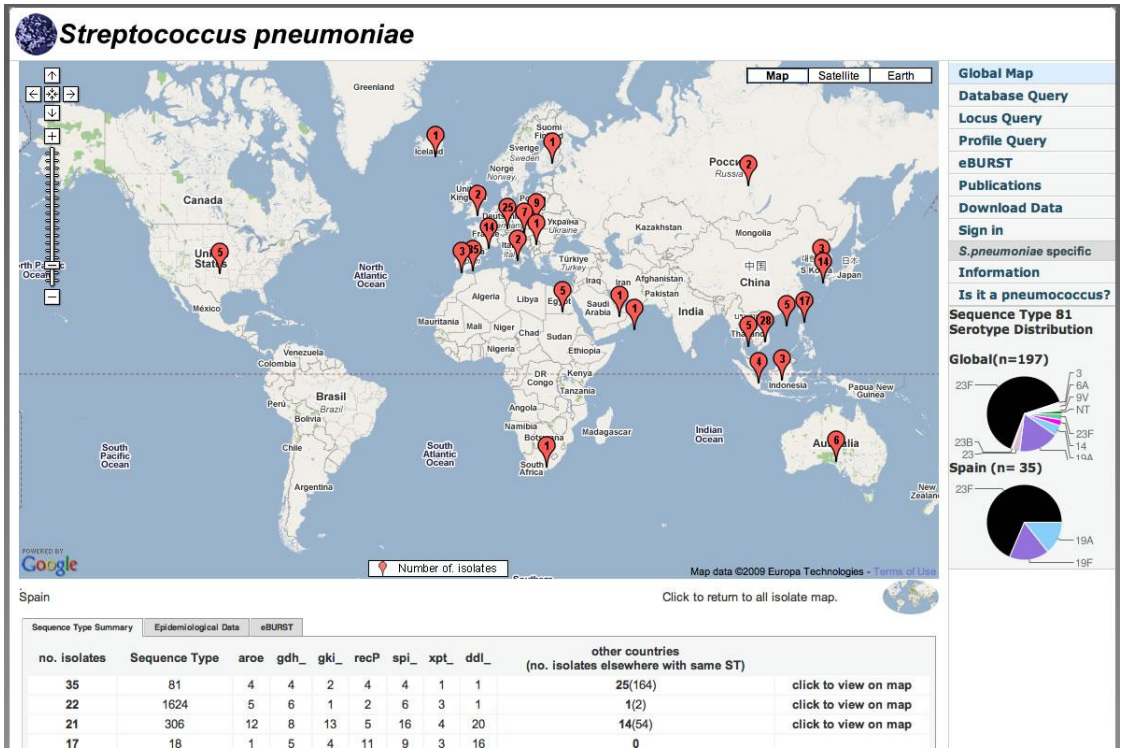


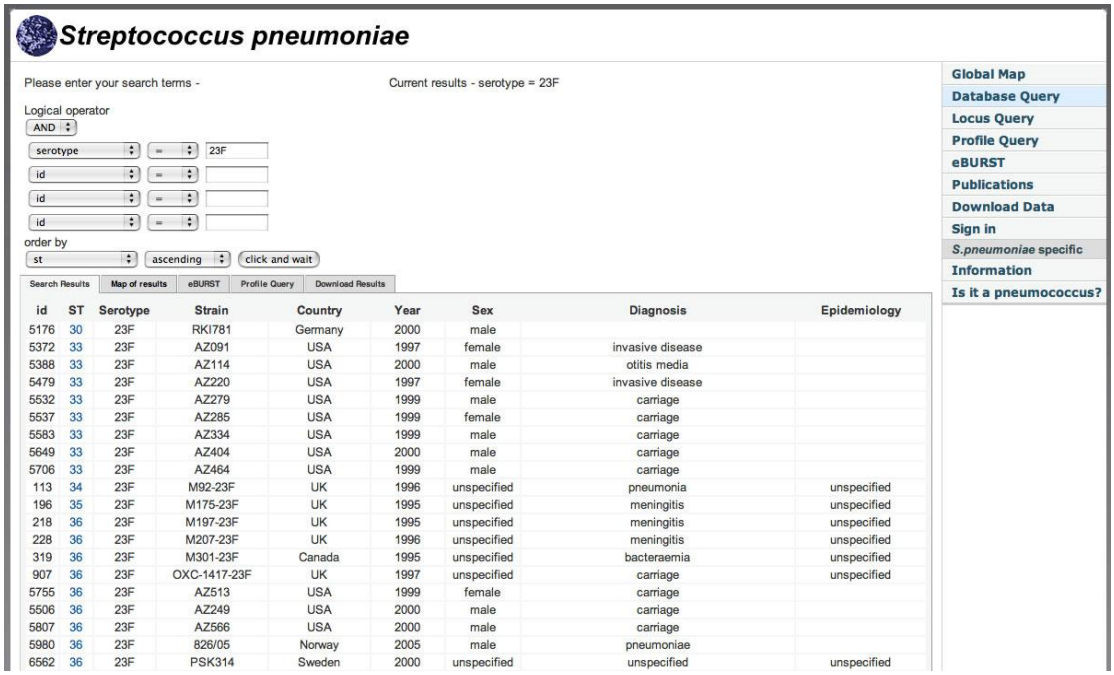
Figure 3.8 Global distribution of *S. pneumoniae* ST 81 isolates.

When clicking the 'click to view on map' link for ST81 (of which 35 isolates are found in Spain, and which currently is found in 25 other countries, with 164 isolates elsewhere), the map returns to the global view and the country locations of isolates of ST81 are displayed as a series of placemarks, with the number of isolates in those countries of ST81 displayed numerically within the placemark (Figure 3.8). Additionally, the serotype distribution of ST81 isolates, globally, and for Spain, are displayed as pie charts below the navigation menu. At this time it can be seen that the majority of ST81 isolates in Spain, and also globally, are of serotype 23F, which is in accordance with the original Pneumococcal Molecular Epidemiology Network (PMEN) global clone 'Spain 23F-1' which was typed using MLST as ST81 (McGee et al., 2001).

Within this example (*S. pneumoniae*), capsular serotype distribution is graphically displayed as pie charts for a particular ST. However, any stored epidemiological variable, or variables, could be displayed in such a fashion. Serotype was chosen as this is most often the question that users of the *S. pneumoniae* MLST database wish to ask.

The area below the map containing the list of STs found in Spain (Figure 3.8) is composed of three 'TABS' (in reality, three 'DIV' elements, the display of which is controlled, via JavaScript as detailed previously). By default the 'Sequence Type Summary' is displayed, however, clicking on the Second TAB, labeled 'Epidemiological Data', allows a user to view the deposited epidemiological data for isolates from Spain ordered by ST. The third TAB, labeled 'eBURST' provides a facility for viewing eBURST either loaded with only those isolate STs found in Spain or, through using comparative eBURST, comparing the isolate STs found in Spain to those found in any other country or set of countries. (This will be detailed further in section 3.3.3.1). An icon allowing one to return to the global distribution of isolates is present above these TABS. This map based interrogation and overview of the database provides an intuitive interface for users to explore the contents of MLST databases geographically. Further integration of maps is present within the 'Database Query'.

Clicking the 'Database Query' menu option hides the Map DIV and displays the Database Query DIV presenting a similar interface to the advanced query option from mlst v1. When entering a query, for example 'Serotype = 23F' and 'order by ST', and clicking the 'Click and Wait' button, a 'Data loading' icon appears whilst the server responds, passing the results for display to JavaScript (Figure 3.9).



id	ST	Serotype	Strain	Country	Year	Sex	Diagnosis	Epidemiology
5176	30	23F	RKI781	Germany	2000	male		
5372	33	23F	AZ091	USA	1997	female	invasive disease	
5388	33	23F	AZ114	USA	2000	male	otitis media	
5479	33	23F	AZ220	USA	1997	female	invasive disease	
5532	33	23F	AZ279	USA	1999	male	carriage	
5537	33	23F	AZ285	USA	1999	female	carriage	
5583	33	23F	AZ334	USA	1999	male	carriage	
5649	33	23F	AZ404	USA	2000	male	carriage	
5706	33	23F	AZ464	USA	1999	male	carriage	
113	34	23F	M92-23F	UK	1996	unspecified	pneumonia	unspecified
196	35	23F	M175-23F	UK	1995	unspecified	meningitis	unspecified
218	36	23F	M197-23F	UK	1995	unspecified	meningitis	unspecified
228	36	23F	M207-23F	UK	1996	unspecified	meningitis	unspecified
319	36	23F	M301-23F	Canada	1995	unspecified	bacteraemia	unspecified
907	36	23F	OXC-1417-23F	UK	1997	unspecified	carriage	unspecified
5755	36	23F	AZ513	USA	1999	female	carriage	
5506	36	23F	AZ249	USA	2000	male	carriage	
5807	36	23F	AZ566	USA	2000	male	carriage	
5980	36	23F	826/05	Norway	2005	male	pneumoniae	
6562	36	23F	PSK314	Sweden	2000	unspecified	unspecified	unspecified

Figure 3.9 Database query results: Epidemiological view.

The results are used to update the five 'TABS' that can be seen below the search form. The 'Search results' TAB contains the epidemiological details for isolates returned from the search query (i.e. all those that have been serotyped as 23F) and data are listed in order of ST.

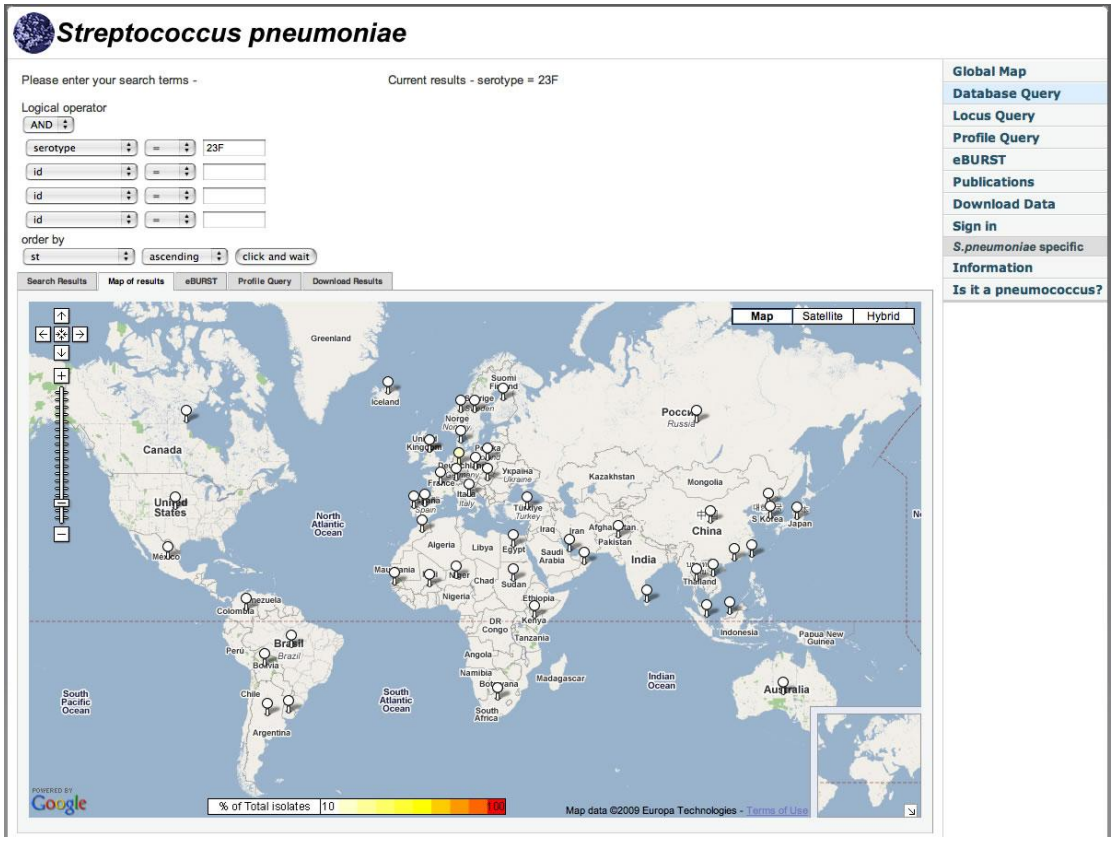


Figure 3.10 Database Query results: Map-based view.

Clicking the 'Map of Results' TAB displays all country locations (as placemarks) based on the results of the database query. In this example, all countries from where serotype 23F isolate details have been deposited in the MLST database (Figure 3.10). Similarly to the main map view described above, the colour of the placemark denotes the percentages of the total number of isolates returned that are found in each country. Clicking on a placemark indicates, within an Info Window, exactly how many isolates are found from that country.

The 'eBURST' TAB shall be detailed in section 3.3.3.2. The 'Profile Query' TAB allows closely related STs within the MLST database to be interrogated based on an initial allelic profile. From the 'Search Results' TAB, any ST (highlighted) can be clicked on and the allelic profile automatically entered into the Profile Query TAB

and the TAB displayed. The allelic profile can then be queried as normal but in the context of the database search.

The 'Download Results' TAB allows the download of all data based on the initial database query in a number of formats. The contatenated sequences for all STs returned from a database query can be downloaded in MEGA, FASTA, CSV or TAB delimited formats. Allele sequences for each locus can also be downloaded in similar formats and all isolate epidemiological data (again based on the query) can be downloaded.

3.3.3 Integration of comparative and geographical eBURST

eBURST is implemented as a Java 'Web Start' application [W44]. Web Start allows stand-alone Java applications to be deployed, by a single click, over a network (in a similar fashion to ACT, detailed in Chapter 2). Using this method allows eBURST to be launched from within a webpage. When launching, certain parameters can be specified, such as the name and location of a text file containing data to be loaded. eBURST takes lists of allelic profiles as input and this allows one to create text files on the web server (using server-side scripting), and to specify the location of the text file when eBURST is launched. Thus the data is immediately viewable.

eBURST provides an option to compare two MLST datasets and to highlight STs which are common to both. Figure 3.11 shows a 'comparative eBURST' population snapshot based on loading two sets of STs. In this example, the first 100 STs in the *S. aureus* MLST database were used. STs 1 to 50 are used as the initial dataset (termed the 'reference'), and STs 50 to 100 are used as a second dataset (termed the 'query'). ST 50 is therefore present in both the reference and the query datasets. Both were loaded into eBURST using the comparative option at a further sub-domain of mlst.net: http://eburst.mlst.net. This site was developed to function as a single access point to all MLST databases, pre-loaded into eBURST, and to allow the entering of user datasets directly (full instructions for the use of eBURST are also provided).

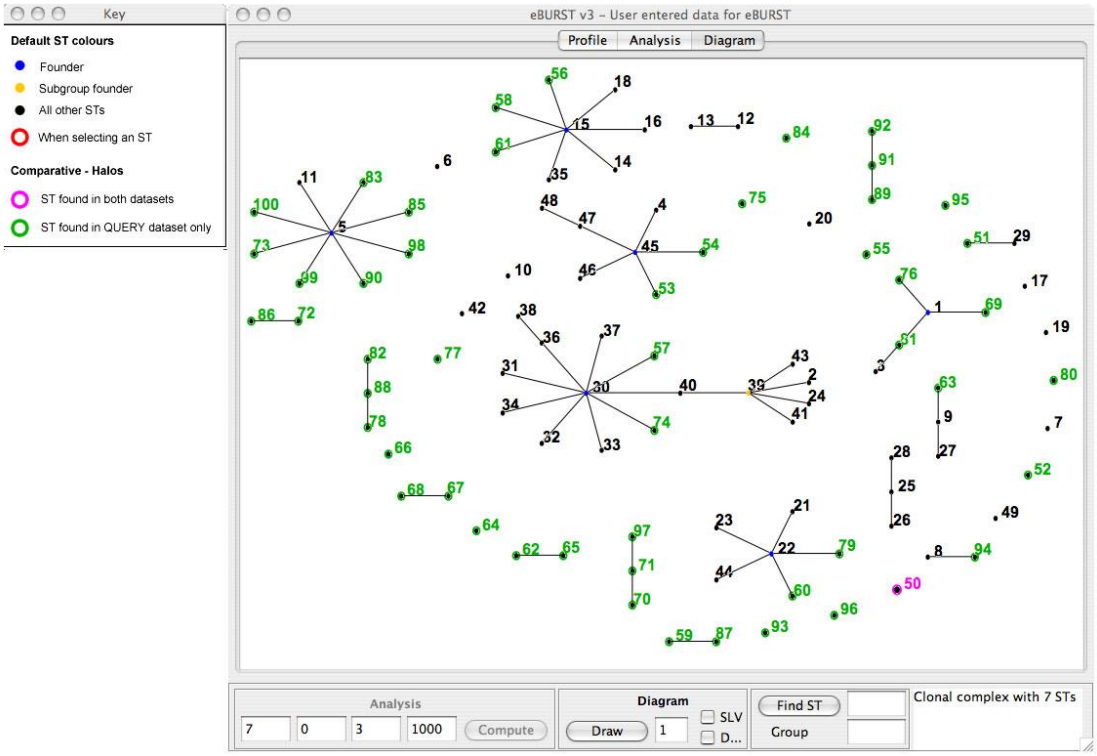


Figure 3.11 Comparative eBURST snapshot: Demonstration.

The population snapshot displays all of the clonal complexes defined by eBURST within the combined datasets. All STs found only in the reference dataset (STs 1 to 49) are highlighted in BLACK, and all STs found only in the query dataset are highlighted in GREEN (STs 51 to 100). Any ST found in both datasets is highlighted in PINK (ST 50). Using comparative eBURST in this fashion it is possible to compare any MLST dataset to any other.

Within mlst.net v1, eBURST can be launched, pre-loaded with data, following a database query, a profile query, a batch strain query or from a single page that allows users to view the entire MLST database. However, within mlst.net v1, the option for using comparative eBURST is not utilised when generating datasets based on querying the database.

mlst.net v2 provides the same options for using eBURST as mlst.net v1 but introduces two major additional methods:

1) Geographical Comparisons.
2) Database Query Comparisons.

3.3.3.1 Geographical eBURST comparisons

If we return to the example shown in Figure 3.7 (loading mlst.net v2, and clicking on 'Spain' to return all STs within the database) and click the eBURST TAB, we are given two options. Either we can view eBURST loaded only with STs from isolates found only in Spain or, by using comparative eBURST, we can compare those isolate STs found in Spain (the reference dataset) to those from any other country or set of countries (the query dataset). The list of countries available for comparison contains only those for which isolate details have been deposited within the MLST database (number of isolates in parenthesis).

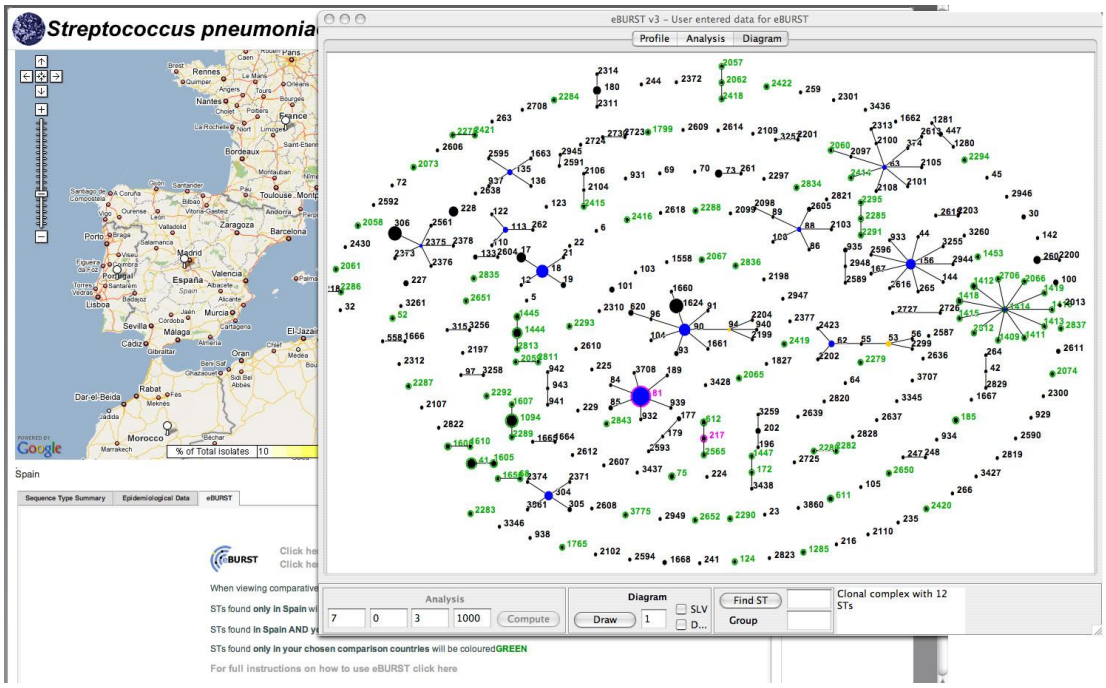


Figure 3.12 Comparative eBURST snapshot: Spain vs. South Africa.

As an example, Figure 3.12 shows a comparative eBURST population snapshot when comparing isolate STs from Spain with those from South Africa. Those STs highlighted in pink are isolated from both Spain and South Africa (i.e. ST81, as already noticed from the ST distribution map; Figure 3.8), those highlighted in green are from isolates from South Africa only and those STs in black are from Spain only.

3.3.3.2 Database query eBURST comparisons

Comparative eBURST can also be accessed from the 'Database Query'. Following on from the example shown in Figure 3.9 (serotype = 23F, order by ST). Clicking the eBURST TAB allows one to view the results of the query in eBURST (i.e. all STs of serotype 23F isolates). Additionally, using comparative eBURST, the results of the query (serotype=23F) can be compared to the results of any other database query. As an example, Figure 3.13 shows the comparative eBURST population snapshot comparing isolates with serotype 23F to isolates with 'serotype = 19F'. STs that are exclusively serotype 23F are coloured BLACK, those of both serotype 23F and 19F are coloured PINK and those exclusively serotype 19F are coloured GREEN. It can be seen that there are clonal complexes of either 23F or 19F and very few that contain both serotypes, the most notable exception being ST81, as previously indicated by pie charts in Figure 3.8.

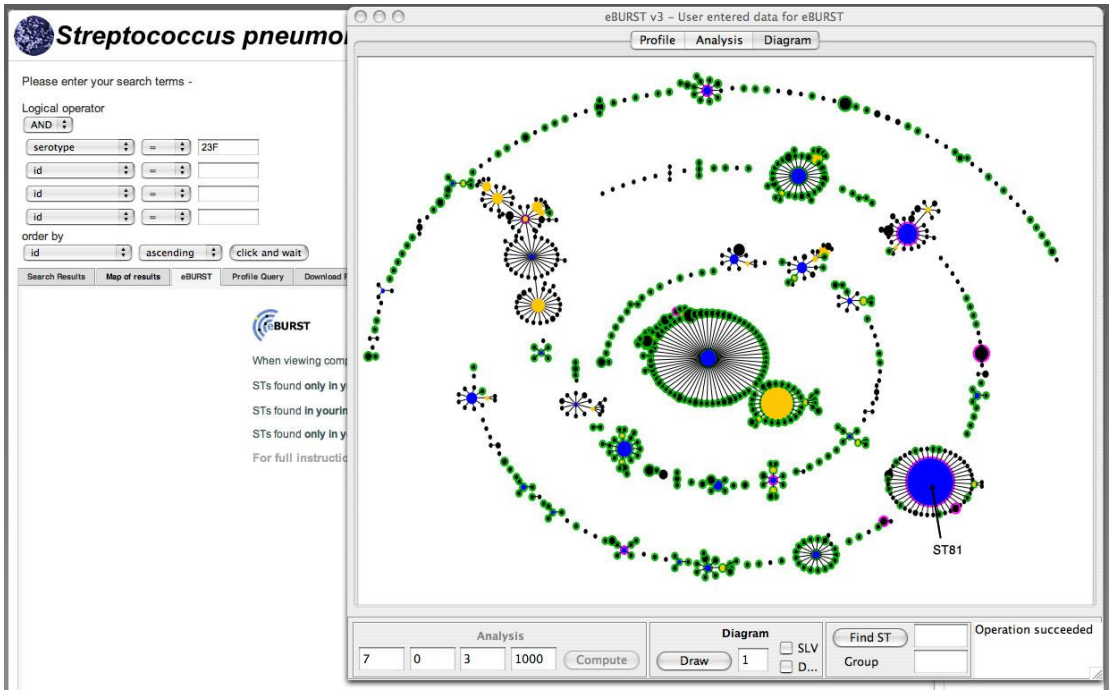


Figure 3.13 Comparative eBURST snapshot: Serotype 23F vs. Serotype 19F. Please note ST labels have been removed for clarity.

3.3.3.3 eBURST menu option

The examples of eBURST integration above allow clonal complexes to be viewed as a consequence of standard user workflows, i.e. a user has not specifically set out to solely use eBURST. The eBURST section of mlst.net v2, as well as an introduction, allows a user to access eBURST specifically or to enter their own data. The option 'Run eBURST on the entire database' presents a link to open eBURST containing all isolate STs in the database. 'Enter your own profile data' allows users to enter their own allelic profiles (that may not have been submitted to the MLST database) and either, view their profiles separately, or view their data in conjunction with either all isolate STs in the database or a list of distinct STs (one example of each ST). Finally, 'Geographical eBURST', allows a user to choose to define as the reference dataset one or more countries and compare to a query dataset of one or more countries STs.

3.3.4 Additional mlst.net v2 features

Additionally to mapping tools and increased eBURST integration, the following features are also present in mlst.net v2.

3.3.4.1 Publications

The publications option of the navigation menu provides a list of all current publications, which include the words '*Streptococcus pneumoniae'*, '*S. pneumoniae'*, 'pneumococcus' and 'multilocus sequence typing' or 'MLST'. The list is incorporated into the site as a Rich Site Summary (RSS) feed from Hubmed (Eaton, 2006; [W38]), which provides a number of accessory tools for viewing publications listed in PubMed [W39]. Abstracts can be toggled on or off and direct links to the publications are present, if available.

3.3.4.2 Download data

Links are available to download the entire database in a number of formats.
The concatenated sequences of all STs (or individual locus sequences) can be downloaded as FASTA, MEGA, CSV or TAB delimited formats. All isolate epidemiological information can also be downloaded in CSV or TAB delimited

format. This facility is also available in mlst.net V1, however has been rationalised within mlst.net V2 and links to download data in these formats are also available based on the output of a database query, as described above, allowing download of specific subsets of the database.

3.3.4.3 Organism-specific tools

As a consequence of the differing biology of organisms represented within MLST databases, tools that are relevant only to specific species may be neccessary. The navigation menu therefore includes an 'organism specific' section which, at its most basic, contains information pertinent to that particular MLST scheme (history of MLST scheme, references, loci chosen, primer sequences and PCR protocols etc.). However, further tools can be added where necessary and within the *S. pneumoniae* example of mlst.net v2, the 'Is it a pneumococcus?' facility falls into this category.

'Is it a pneumococcus?' provides the same underlying functionality as mlst.net v1 (section 3.2.4.3), however, a novel tree visualisation tool, PhyloWidget, has been incorporated in mlst.net v2 allowing trees to be viewed in a more intuitive fashion. PhyloWidget is a Java application, which is embedded within a webpage and allows communication between a webpage and Java via JavaScript (Jordan and Piel, 2008 [W40]).

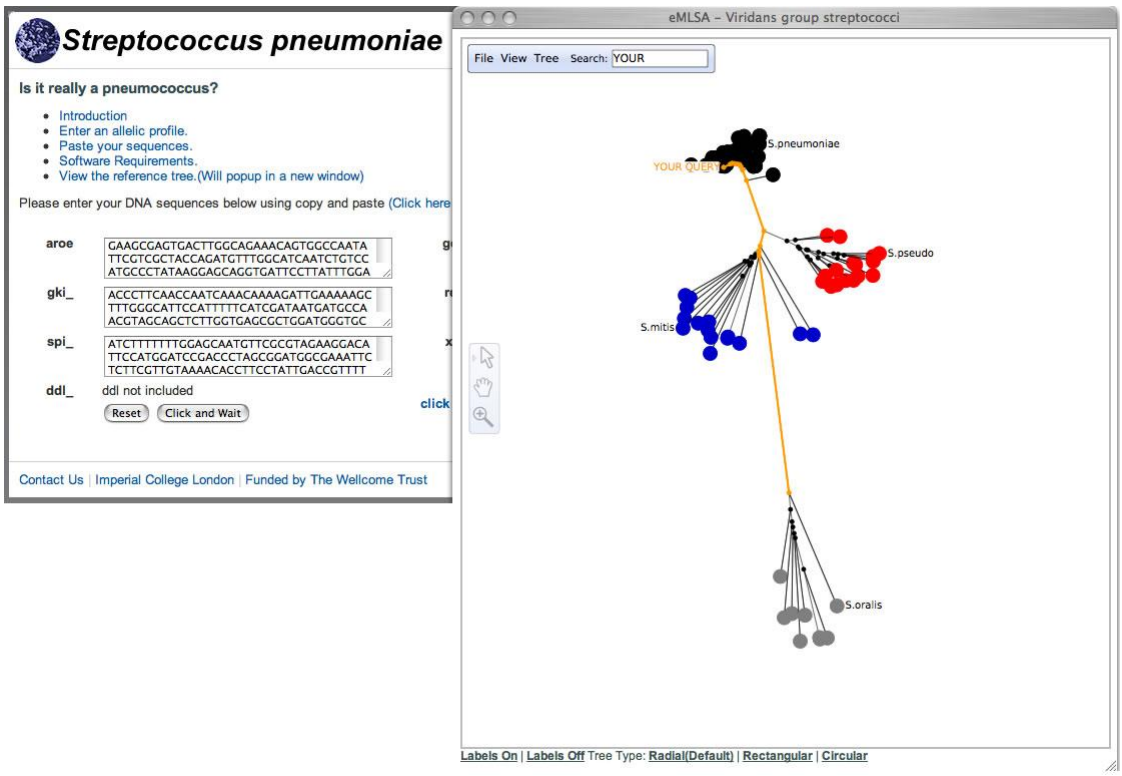


Figure 3.14 Is it a pneumococcus? mlst.net v2.

Similarly to mlst.net v1, a user can either enter their own sequences or an allelic profile, and a tree based on concatenated sequences is produced, highlighting the position of their query sequences on the tree. Phylowidget, and a web application for species comparison will be discussed in more detail in chapter 4. However, Figure 3.14 shows the output within mlst.net v2 based on a user querying the facility with a set of locus sequences, which are clearly of pneumococcal origin.

3.4 Discussion

This chapter describes the web application located at http://www.mlst.net and the tools that allow users to identify query strains, and to explore their relationship with other strains in the database. mlst.net v1 is widely used within research and public health laboratories and data submission from the communities of users continues to increase.

As databases continue to grow and web development technologies mature we have a duty as providers of research tools to increase, within such software, appropriate functionality and usability. This aim has been addressed with the development of mlst.net v2. The interface re-design and re-development (using AJAX), the tighter integration of eBURST (and comparative eBURST) and the introduction of mapping tools provide users with a much richer interrogation of the data held within MLST databases. The provision of a more desktop like user interface, with components of the page updating silently without page refresh or interruption to a user's workflow, should substantially improve download speed and improve the responsiveness of the site. Furthermore, the AJAX interface allows data entered into a page to be kept current when submitting to the server, rather than refreshed as in standard page-click-page websites (such as mlst.net v1). The advantages from a user perspective of this persistence of data will only become apparent through user feedback and it may well be the case that the page-click-page interface within mlst.net v1 is preferred. If so, the additional functionality provided by maps and eBURST can be preserved and a more traditional interface re-implemented.

The reliance of AJAX on JavaScript may potentially provide issues for some users. All current web browsers support JavaScript, however some legacy browsers may not. There is therefore a small, but real potential for a user to access mlst.net v2 from a browser which is not supported. To address this, a simple detection script (written in JavaScript), which is run on the server as soon as the initial page is loaded (to test if the function can successfully be called), will warn users that they should update, with links provided to download the latest, free versions of several browsers. A further issue may present itself should a user access the site from a location where the use of JavaScript is restricted. Some companies have, in the past, disallowed users to view sites that contain JavaScript, due to possible security issues. This has become less of

an issue over the previous years with the maturity of both the JavaScript language and security measures implemented within browsers. Furthermore, mlst.net v1 utilises JavaScript (albeit is a more basic form) for menu-based navigation between pages. Navigation in mlst.net v1 is undertaken using dropdown lists which react to a users page choice by calling a JavaScript function, which opens the selected page. Since mlst.net v1 was introduced we have received no feedback regarding problems with JavaScript and I am therefore confident that it is not an issue. However, the increased reliance on JavaScript within mlst.net v2.0 further stresses the importance of BETA testing.

The increased integration of eBURST in mlst.net v2, and in particular comparative (both database and geographical) eBURST provides users with a more intuitive exploration of an MLST database than within mlst.net v1.0. Prior to implementation, careful consideration was taken to identify where and how a user may wish to use and compare datasets within mlst.net v1. eBURST is a Java Web Start application [W44] launched pre-loaded with a particular set of STs. Once launched, it is not possible to interact further with a database without re-launching the application (loaded with another set of STs). Further interactivity with eBURST would be advantageous for users. Currently, when a population snapshot is viewed, the assignment of STs to particular clonal complexes is based on those STs loaded from the particular MLST database. The addition of subsequent STs to the population may cause clonal complexes to join together if, for example, an ST not previously sampled but potentially providing a linkage between two groups is subsequently sampled. Therefore, the clonal complex membership of a particular ST may change with time and increased population sampling. This potential for inconsistent nomenclature for assigning STs to clonal complexes precludes the inclusion of clonal complex membership directly within our MLST databases. However, it would be useful for a user to be able to interrogate MLST databases based on current clonal complexes to, for example, view antibiotic resistance data for isolates with STs assigned to a particular clonal complex. Furthermore, the ability to display directly on an eBURST diagram antibiotic resistance data, or any other epidemiological variable, would provide further benefit to users and their (and our) visualisation and understanding of the data. To further this aim, Saurabh Johri is currently re-developing eBURST within the group using a programming language focussed on integrating Java applications within a web environment.

The introduction of mapping tools within mlst.net v2 provides a simple, intuitive interface for interrogating and displaying the global distribution of STs, including epidemiological variable summaries. More importantly, it presents users with the possibility to submit and view MLST data at increasing spatial resolution, and therefore explore the geographic structure of populations further. Since the introduction of MLST-Maps (which will be described in further detail in chapter 5) user feedback has been wholly positive and this informed the decision to include mapping tools more prominently within mlst.net v2. Recently, the *Leptospira spp*. MLST database (http://leptospira.mlst.net) has been updated by the curator, Janjira Thaipadungpanit, to include point locations for a number of isolates (and corresponding STs) in and around the Udon Thani province in Northern Thailand.

The submission of increased spatial data within MLST databases raises exciting possibilities for the analysis of such information and corresponding methods will need to be introduced / developed to fully exploit the spatial clustering of STs. However, with increasing spatial resolution comes issues relating to data confidentiality (i.e. locating in space isolates of a pathogen) and caution will need to be applied. To address such issues, private databases allowing users to explore their data in such a fashion without impacting confidentially will be developed and, implemented in tandem to mlst.net v2.

Historically, the submission of epidemiological data to MLST databases is often limited to those isolates that provide novel alleles/STs. However, we aim to provide functionality to encourage users (and curators to encourage users) to submit further epidemiological data and hopefully, the additional tools provided within mlst.net v2 will further this aim.

In conclusion, the additional features implemented within mlst.net v2 are aimed at providing users with an increased toolkit to interrogate and understand their data in conjunction with the populations represented within our MLST databases. BETA testing of mlst.net v2 will be extended to all other MLST schemes held at mlst.net and the testing period will run until such time as we are confident that users are happy with the additional features.

Chapter 4: Assigning Strains to Bacterial Species via the Internet: eMLSA.net

Available: http://www.emlsa.net

Summary

The methods currently used for assigning bacterial strains to species are widely held to be inadequate. Recently, multilocus sequence analysis (MLSA), which uses the concatenated sequences of multiple house-keeping genes, has been shown to identify well resolved sequence clusters that can potentially be equated with species. An advantage of this approach is that large initial sequence databases can be generated for a species group and when combined within a web application, allows the assignment of query strains to species via the internet (electronic taxonomy). In this chapter I detail the development of a generic web application for electronic taxonomy, eMLSA.net. Using a database containing sequences from the Viridans group streptococci as an exemplar, eMLSA.net allows the sequences of a query strain to be entered and for the species assignment to be returned, according to its position within an assigned species cluster on a reference tree. Furthermore, individual gene trees allow the identification of resident or potentially foreign alleles.

The web application described has been designed to be generic, allowing the hosting of further species group MLSA schemes and strain databases at eMLSA.net

Note: The application detailed in this chapter has been described in the following publication, a copy of which is included in Appendix B:

Bishop, C.J*., Aanensen, D.M*., Jordan, G.E., Kilian, M., Hanage, W.P. and Spratt, B.G. (2009) Assigning strains to bacterial species. *BMC Biol* **7**:3

*Joint first authors

4.1 Introduction

In recent years the methods used to assign strains to bacterial species, and the definition of species have been under considerable debate. There has been little progress in defining an agreed concept with which to guide a new definition of species. However, there is agreement that the current methods of defining prokaryotic species are inadequate (Achtman and Wagner, 2008; Cohan, 2002; Fraser et al., 2007). Polyphasic taxonomy, incorporating all available phenotypic and genotypic data, is currently used to define a consensus classification for species (Vandamme et al., 1996). Genotypic definition of species is undertaken based on overall genomic relatedness. Strains which share approximately 70% or more relatedness using DNA-DNA hybridisation, with 5°C or less difference in the melting temperature between homologous and heterologous hybrids (ΔT_m) are considered as members of the same species (Wayne et al., 1987) and type strains defining species using these methods are available from organisations such as ATCC-LGC [W76] for further study. Such rigid cut-offs do not take into consideration differences in the biology of populations and may not be appropriate for all bacterial populations. Furthermore, with the widespread availability of sequence analysis methods, other methods of identifying bacterial species have been encouraged (Stackebrandt et al., 2002).

One approach is Multilocus Sequence Analysis (MLSA), which uses the concatenated sequences of multiple house-keeping genes for sequence clustering. The distribution of a large number of strains of closely related species in sequence space can be used to identify clusters of strains that are well resolved from other clusters. (Gevers et al., 2005; Hanage et al., 2006). Furthermore, if large populations of similar strains form well-resolved sequence clusters this could potentially be used as a proxy for the existence of species.

MLSA has been used successfully to explore clustering patterns within large numbers of strains assigned to very closely related species by current taxonomic methods (e.g. *Neisseria* spp., Bennet et al., 2007; *Burkholderia* spp., Godoy et al., 2003; Haemophilus spp., Nørskov-Lauritsen et al., 2005). Furthermore, the technique has been used to explore relationships between small numbers of strains within a genus (*Ensifer*, Martens et al., 2008), and within broader taxonomic groupings (Vibrios, Sawabe et al., 2007).

In a similar fashion to the assignment of isolates to strains or clonal complexes using Multilocus Sequence Typing (MLST) (Maiden et al., 1998; Chapter 3 of this thesis), MLSA should allow the assignment of unknown strains to species clusters via the internet (electronic taxonomy). To fulfil this aim, electronic taxonomy requires a large initial database (and associated clustering patterns) of multiple strains from all of the recognised species within a genus, or part of a genus. It also requires a web application allowing the sequences from the house-keeping genes of a new query strain to be entered, and for the sequence cluster (if any) within which the query strain falls to be returned, along with a species assignment for the strain. Such an approach would therefore allow web-based species recognition and also the recognition of new clusters, which can be assigned as new candidate species.

This chapter details the development and functionality of a generic electronic taxonomy web application, eMLSA.net, exemplified using sequences from the Viridans group streptococci.

Non-pyogenic streptococci are subdivided into the Mitis, Anginosus, Salivarius, Mutans and Bovis groups, of which the first three are often referred to as the viridans streptococci. Viridans streptococci (with the exception of *Streptococcus thermophilus*) are commensals of the human upper respiratory tract and, excepting *Streptococcus pneumoniae*, rarely cause disease in individuals with competent immune systems. The taxonomy of viridans streptococci is problematic (Kilian, 2005) and this is reflected by frequently changing nomenclature and significant issues with identification by phenotypic analysis and by sequencing of 16S rRNA genes (Hoshino et al., 2005; Whiley et al., 1999).

The identification of strains that fall between well-resolved sequence clusters due to inter-species recombination can lead to ‘fuzzy species’ (Hanage et al., 2005), which may prevent the identification of well-resolved sequence clusters. Recombination between closely related Mitis group species is known to occur (Dowson et al., 1989, 1993; Laible et al., 1997; Hackenbeck et al., 2001). However, studies have shown that sequence clusters within the Mitis group can be resolved using trees from the concatenated sequences of six of the seven house-keeping genes used in the *S. pneumoniae* MLST scheme (Hanage et al., 2006; Enright and Spratt, 1998; Chapter 3, this thesis). Furthermore, concatenation of the sequences of four house-keeping genes

has been shown to resolve the more distantly related species within the viridans streptococci (Hoshino et al., 2005).

The eMLSA.net web application utilising the viridans group sequences detailed in this chapter has been described in Bishop et al, (2009) (included in Appendix B) on which I was joint first author. Cynthia Bishop undertook all experimental work and prior to describing the web application, the methods for MLSA scheme definition and the sequencing and MLSA results will be detailed.

4.2 Sequencing and cluster analysis

A strain collection covering the diversity of the viridans group streptococci was established (number of strains in parenthesis)

- Mitis Group: *S.australis* (10), *S.cristatus* (5), *S.gordonii* (28), *S.infantis* (24), *S.mitis* (88), *S.oligofermentans* (1), *S.oralis* (84), *S.parasanguinis* (25), *S.peroris* (1), *S.pneumoniae* (51), *S.pseudopneumoniae* (20), *S.sanguinis* (25), *S.sinensis* (1).
- Anginosus group: *S.anginosus* (7), *S.constellatus* (4), *S.intermedius* (2).
- Salivarius group: *S.salivarius* (4), *S.vestibularis* (4), *S.thermophilus* (1).

Included in the collection were the type strains of all of the viridans group species, excepting *S. pneumoniae*.

The choice of loci for circumscribing clusters for electronic taxonomy should fulfil the following criteria:

1. Be present in all strains of species under investigation.
2. Be sufficiently conserved such that a single pair of primers can be used for amplification.
3. Should be sufficiently diverse to resolve sequence clusters.

Hoshino et al, (2005) utilised four house-keeping loci; *ddl*, *gdh*, *rpoB* and *sodA* to address species resolution within the viridans group streptococci. Primers for *ddl* and *gdh* failed to amplify sequences in a number of species while conversely *rpoB* and

sodA primers yielded amplicons from virtually all strains. Therefore, these two loci were chosen for inclusion in the viridans group MLSA study.

Additonal loci were chosen following sequence comparisons of publicly available streptococcal genomes. 138 proteins with house-keeping functions, present in all genomes, and which had at least 80% amino acid sequence identity were selected and their corresponding gene sequences assessed for suitability. Degenerate primers were designed for each selected locus, based on stretches of six amino acids, conserved in all species, which would allow an internal fragment of roughly 350bp to be fully sequenced on both strands. Primers for candidate loci were used to check their ability to amplify in a diverse subset of the viridans group strain collection until a final set of seven loci were selected (see Table 2)

Gene	Gene Product	Gene Fragment Length
map	Methionine aminopeptidase	348
pfl	Pyruvate formate lyase	351
ppaC	Inorganic pyrophosphatase	552
pyk	Pyruvate kinase	492
rpoB	RNA polymerase beta subunit	516
sodA	Superoxide dismutase	378
tuf	Elongation factor Tu	426

Table 4.1 Loci used for viridans group MLSA scheme.

This final set of primers was used for PCR amplification and to obtain the DNA sequences of loci from the strain collection. In addition, the sequences of the seven MLSA loci were extracted from publicly available genome sequences. In total there were 420 sets of sequences, corresponding to 399 obtained from viridans group streptococci (plus three extracted from *S. thermophilus* genome sequences), 17 obtained from *S. pyogenes* strains and one from *S. agalactiae.*

The sequences of the seven gene fragments from each strain were joined, in-frame, in the order *map-pfl-ppaC-pyk-rpoB-sodA-tuf,* to form a single 3063 bp concatenated sequence. Single gene trees and trees using the concatenated sequences were generated by neighbor-joining or minimum evolution using MEGA4 (Kumar et al., 2008) and the robustness of nodes was evaluated by bootstrapping (1000 replicates). This initial dataset was used to investigate patterns of sequence clustering and concordance of clusters with species defined by standard taxonomic procedures including, importantly, the location of the type strains.

Figure 4.1 shows a neighbor-joining tree based on the concatenated sequences, with the type strains highlighted. Nodes are coloured based on assigned species designation (in the laboratory of Mogens Kilian).

It can be seen that a number of nodes have high bootstrap values (≥95%) indicating well resolved clusters. *S. pseudopneumoniae* strains cluster tightly together but are not well resolved from the *S. mitis* cluster (92% bootstrap value). The *S. mitis* cluster is unusual in that it consists of a set of closely related sub-clusters arising from the branch that separates the *S. pneumoniae* and *S. pseudopneumoniae* clusters from that of *S. oralis*.

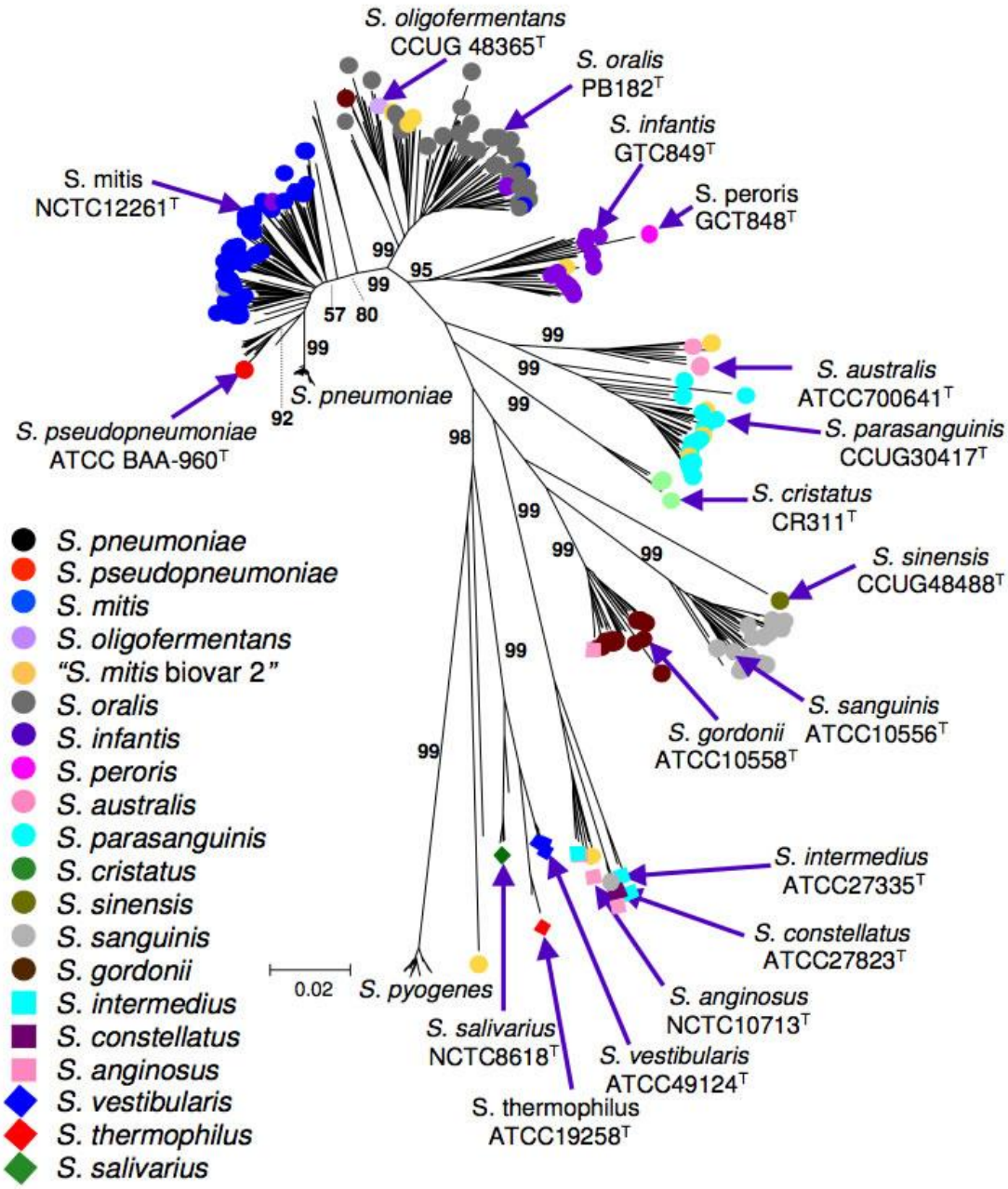


Figure 4.1 Neighbor-joining tree indicating position of type strains within sequence clusters. Mitis, Anginosus and Salivarius group species are shown, respectively, as coloured circles, squares and diamonds. Bootstrap values (%) for nodes leading to each of the Mitis group sequence clusters are shown. Reproduced from Bishop et al, (2009).

4.3 Cluster and species concordance

Almost all of the clusters contained only one designated type strain. The exceptions being the *S. peroris* type strain (within a cluster including the *S. infantis* type strain, and other designated *S. infantis* strains) and *S. oligofermentans* (within a sub-group of *S.oralis*).

In some cases, species names assigned to strains in the laboratory of Mogens Kilian and their position within sequence clusters are discordant. Strains designated as *S. mitis* fell into a number of different closely related sequence clusters, particularly those consisting of *S. infantis* and *S. oralis*. This is largely explained by strains initially assigned as '*S. mitis* biovar 2' being either part of the *S. oralis* cluster or part of the *S. parasanguinis* cluster. Additionally, single strains assigned as *S. gordonii* and *S. infantis* fell within the *S. oralis* cluster, a single strain of '*S. mitis* biovar 2' fell within the *S. anginosus* cluster and a single strain of *S. anginosus* fell within the *S. constellatus* cluster.

Despite the above discrepancies, in the majority of cases, species names assigned through standard taxonomic procedures correlated well with the major sequence clusters (which also included the type strain). As such, each sequence cluster can be used to represent a species cluster. Figure 4.2A shows a neighbor-joining tree for all 420 strains with the strains coloured to reflect their assigned species names. Different tree-building methods may have an impact on clustering and Figure 4.2B shows a tree produced by minimum evolution (based on local optimisation of an initial neighbour-joining tree). The assignment of strains to sequence clusters in the minimum evolution tree was essentially the same, the only significant difference being a slightly more marked sub-division of the *S. mitis* group into sub-clusters. Minimum evolution is a more computationally intensive tree building method and for the purposes of electronic taxonomy for the viridans group the neighbor-joining tree was chosen. The tree can be used as a reference for assigning a query strain to a viridans group species, based on its location within one of the defined species clusters. Figure 4.2C shows the tree with the Mitis group clusters collapsed to indicate better the resolution of species clusters within the Anginosus and Salivarius groups.

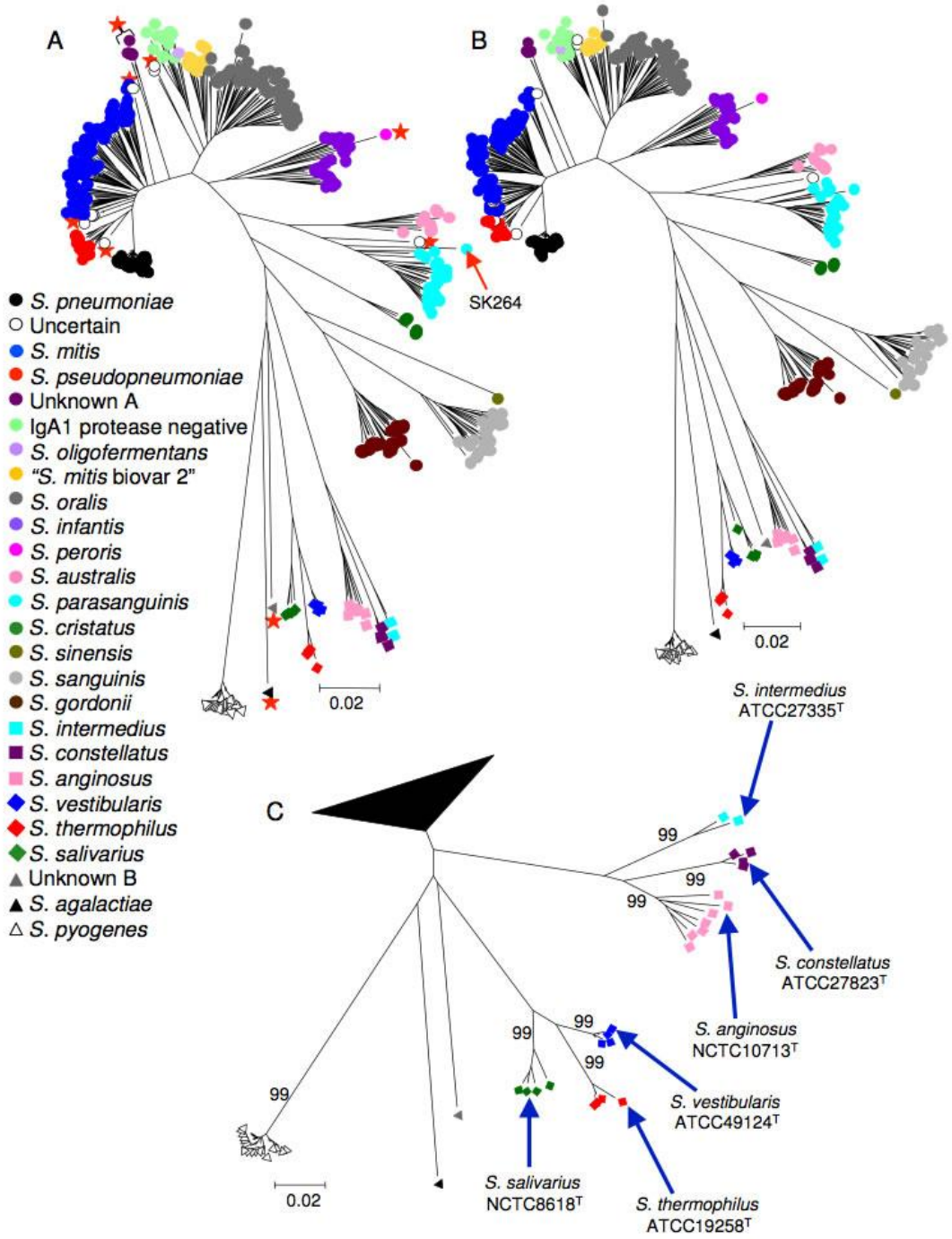


Figure 4.2 A) Re-labelling of the neighbor-joining tree from Figure 4.1 indicating that all strains within a sequence cluster were assigned to the species inferred from the positions on the tree of the type strain and other well-characterised strains. Those strains on the borders of the *S. mitis* and *S. pseudopneumoniae* clusters where assignment is unclear are shown as white circles. Red asterisk's indicates strains that were re-checked (by Cynthia Bishop) as they were outliers, or were distinct from other strains. B) The minimum evolution tree based on the same sequences. C) The neighbor-joining tree with the Mitis group clusters collapsed, indicating more clearly the separation of the Anginosus and Salivarius groups. Reproduced from Bishop et al, (2009).

4.4 Using MLSA to identify new species clusters

Using the reference tree detailed above should allow us to identify possible new species as groups of strains which fall outside of defined and demarcated species clusters. An example of such is presented by those strains labelled as 'unknown A' in Figure 4.2A, which fall between the *S. mitis* and *S. oralis* clusters. Originally assigned as *S. mitis,* these strains were isolated from members of two families included in a study of the clonal diversity of *S. mitis* within individuals (Hohwy and Kilian, 1995). Although they cannot be distinguished using phenotypic methods, they may constitute a separate species. Furthermore, clusters including the type strains of *S. mitis* and *S. oralis* both include several sub-clusters, which may also warrant recognition as separate species. Ten strains previously assigned to *S. mitis* biovar 2, mentioned above, formed a sub-cluster within the major *S. oralis* cluster. Phenotypic examination by Mogens Kilian indicated that these strains are distinct from the remaining part of the *S. oralis* cluster by being arginine hydrolysis and α-maltosidase positive, and by expressing the Lancefield K cell wall carbohydrate antigen. In addition, all strains of the *S. mitis* biovar 2 sub-cluster lacked IgA1 protease activity.

4.5 Single gene trees and identification of interspecies imports

Neighbor-joining trees constructed from the sequences of the seven individual genes from all strains, in the majority of cases, failed to identify the species clusters obtained using the concatenated sequences. Furthermore, in some cases, individual gene trees showed anomalous clustering of strains.

Approximately 96% of strains assigned to a particular species using concatenated sequences had alleles at all seven loci that were most similar to those of other strains of that species. Using the definition of species provided by the concatenated sequences provides the potential to identify inter-species allele importation and to assign sequences at each locus as resident, resident-compatible or foreign. If a strain has been assigned to species A, based on concatenated sequences, for each locus we can examine the pattern of clustering in the gene tree and if the most similar sequences are all from strains of species A the sequence at that locus can be defined as a resident allele. However, if the sequence falls into a cluster assigned to another species, the allele will be considered to be foreign, as long as species A and the

species cluster assigned for that allele are well resolved based on the individual gene tree. The allele can be assigned as 'resident compatible' where the allele sequence is within a cluster of two or more species (including species A) that are not resolved on the gene tree. Sixteen of the viridans group strains (4%) had alleles that were considered to be foreign. These criteria for allele species assignment and identification of potential imports can be utilised to identify query sequences within the web application for electronic taxonomy now described.

4.6 Development of a generic electronic taxonomy web application

To fulfil the requirements for electronic taxonomy and to allow scalability and general applicability, the design of an application should fulfil the following criteria:

i) A database should hold all sequences from a large reference dataset, including associated MLSA data and, most importantly, species designation for all reference strains.
ii) A user should be able to enter their own data for species assignment.
iii) A user should be able to view the location of their query on a tree based on the reference dataset.
iv) Overall species assignment should be based on the cluster into which the query sequence falls and on the similarity of the concatenated query sequence to the concatenated reference set.
v) Species assignment should also be available for individual loci to aid identification of resident and foreign alleles
vi) There should be a simple procedure for data deposition so that the database grows.

4.6.1 Method of tree production

Desktop software packages are traditionally accessed through a graphical user interface (GUI) which allows the processing of tasks to be undertaken through a series of menus / graphical components via the use of the mouse pointer. The vast majority of software available for the Windows and Macintosh operating systems is used in this way. GUI's make the process of using the underlying algorithms, which define the application logic of a piece of software, easy and intuitive for a user.

Many software developments in the open source community, particularly within bioinformatics software, do not rely on GUIs and provide methods to access software in either one or two ways. Firstly, access can be achieved via a set of textual menus that, through typing menu options, allow the application logic of the software to be navigated. Alternatively, and often as well as, software can be accessed and run directly via a command-line interface. The command-line interface is a mechanism for interacting directly with a computer operating system or software by typing specific commands to perform tasks. In reality, software that is generally widely used can offer all of these methods of access. For example, there are versions of Clustal, the multiple sequence alignment tool, which provide GUI access (ClustalX – available for Windows, Mac OS and Linux; Thomson et al., 2002), whereby a menu is used to open a sequence file and a sequence alignment produced. A version of Clustal also exists which provides a text menu (termed interactive mode) allowing a sequence alignment to be undertaken by typing menu options (ClustalW).

Alternatively, ClustalW allows direct command-line access and a sequence alignment can be produced by entering commands similar to the following:

```
clustalw –infile sequences.fasta –outfile alignment.out
```

As with the majority of command-line accessible software, firstly the program to be used is specified (`clustalw`). Following this a series of variables are specified. '`-infile`' is the file containing the sequence that an alignment is to be undertaken on and is set to '`sequences.fasta`', then the '`-outfile`' variable is set to '`alignment.out`' to specify the name of a file to create containing the alignment. Entering these commands passes the sequence file to clustalw, which processes and produces an alignment. In this example, default parameters are used to process the alignment (gap penalties etc.) although these can also be set via the command line. The ability to access all aspects of software functionality via the command line also allows the provision of interfaces to such software via the web.

As mentioned in the introduction to this thesis, a web interface to ClustalW is accessible at the European Bioinformatics Institute (EBI) web server [W5]. The variables that can be passed to clustalw are represented within HTML as form fields

and following submission, are processed using server side-scrips, which, via the command line, runs clustalW using commands similar to those detailed above. The '`outfile`' produced is then linked to or displayed in the results page.

This method of command line access to software is vital for use within eMLSA.net. Clustal is primarily a multiple sequence alignment program, which allows all parameters used to undertake an alignment of a set of sequences to be specified (gap penalties, word size etc.). Clustal also provides a representation of a sequence alignment as a dendrogram. Two methods for tree building are provided; neighbor-joining and UPGMA. In many cases a user may have more specific questions, or wish to use a different tree-building method (such as minimum evolution) and a different model of sequence evolution. In these cases, sequence alignments produced by Clustal can be downloaded and used within other programs that provide more comprehensive functionality.

MEGA is a software package (mentioned in chapter 3) that is specifically focussed on evolutionary genetic analysis. MEGA provides multiple tree production methods and access to multiple evolutionary models, which can be used when inferring relationships within sets of sequences. MEGA uses an aligned set of sequences (produced by Clustal or similar alignment tools) and as such provides tree production methods far superior to those included directly within Clustal.

For the purposes of eMLSA.net we initially wish to produce a neighbor-joining tree, and for this purpose Clustal would suffice (and scripts for web access are freely available). However, the scope of eMLSA.net is to provide generic software, which can be used for further species group schemes. As such, in the future, we may need to provide different tree building methods and associated parameters. Therefore, MEGA provides a more attractive set of functionality for initial usage. With respect to sequence alignment, the lengths of sequences within the viridans group (and generically for MLSA) are the same for each locus and contain no indels, negating the necessity for a full sequence alignment. A server-side version of MEGA, kindly provided by Sudhir Kumar allows, via the command line, a tree description to be produced based on passing the name of a sequence file and specifying the tree building method to be used.

Both Clustal and the server-side version of MEGA output text files. When calculating the topology of a tree a description is produced, in plain text, in Newick tree format (or Newick notation). Initially defined by Joseph Felsenstein and colleagues (with the final specification defined whilst at Newick's restaurant in Dover, New Hampshire, USA), the Newick format provides a textual description of the edges and nodes of a tree using parenthesis and commas. The visualisation of such a description is generally undertaken by, in the case of Clustal, a separate program or, in the case of the desktop version of MEGA, inbuilt graphical rendering tools.

4.6.2 Method of tree visualisation

Ideally, for the purposes of eMLSA.net we would like a user to be able to view, within the web application, the same tree that was produced in Figure 4.2A. Each of the species clusters should be coloured differently and the position of their query sequence should be highlighted. Furthermore, the option to collapse parts of the tree should be available to, for example, collapse the Mitis group to highlight sub-clustering within the Anginosus and Salivarius groups as in Figure 4.2C. These features, combined with the variety of tree building options provided by MEGA would maximise generality for the addition of future MLSA schemes.

A number of software applications, focussed solely on tree visualisation and editing (taking as input a Newick tree description) are currently available. Of these, the majority exist purely as desktop, GUI-based applications (eg TreeView, Page, 1996; NJPlot, Perriere and Guoy, 1996 and TreeDyn, Chevenet et al., 2006) and are therefore not applicable for web application development. Of those that are applicable for use within a website, ATV (Zmasek and Eddy, 2001) and Dendroscope (Huson DH et al., 2007) are the most widely used. Both ATV and Dendroscope are available as Java applets and a Newick file can be specified as a parameter on launching. ATV renders the Newick description as a simple horizontal or vertical dendrogram, which can be zoomed in an out. Dendroscope offers a number of extra rendering options such as radial, circular and slanted trees. While both are very good for representing tree structure, the customisation options are quite limited. ATV allows only text labels to be coloured, whereas Dendroscope allows individual terminal nodes and labels to be differentially coloured. However, this needs to be undertaken by a user within the

software and cannot be specified using parameters upon initial launch. Both ATV and Dendroscope also suffer from a lack of two-way interaction with a web application, which was discussed briefly with regards to the current implementation of eBURST in the discussion to chapter 3.

These issues have been addressed to some extent within the software PhyloWidget (PW) (Jordan and Piel, 2008). The development of PW aimed to provide an intuitive, highly configurable tool, tightly integrated within the web environment. Written by Greg Jordan at the EBI, PW uses the programming language Processing [W45]. Processing is an open source language, aimed at programming images and animation with an emphasis on interactivity. Based on Java it provides multiple graphical libraries and network capabilities that allow it to be controlled via JavaScript, making it ideal for integration with web applications.

PW provides a click-drag-zoom interface for tree navigation, multiple options for tree rendering (radial, horizontal, circular etc.) and highly configurable options for the colour of nodes and text. Through the provision of an API, underlying functionality within PW can be accessed via JavaScript, allowing aspects of tree rendering to be controlled directly via HTML. However, the initial release of PW does not provide enough control to specify sets of terminal nodes to be coloured (or indeed symbol-based node representation) and furthermore, does not contain a method for node collapsing, both of which are required to fulfil the aims of eMLSA.net.

Through a collaboration with Greg Jordan, these issues were addressed and a specific version of PW produced for eMLSA.net which provides such functionality. To achieve the aim of labelling we defined a basic vocabulary which, when added to terminal node labels within the Newick format produced by MEGA, would allow the shape, size, colour and initial text labelling of terminal nodes to be specified and subsequently used for display. The ability to collapse branches of the tree by selecting a specific node was added and, furthermore, additional parameters such as initial tree type (circular, radial etc.) could be defined and subsequently changed directly from JavaScript.
Using the server-side version of MEGA we can produce a Newick description of a tree based on a set of sequences, with flexibility for different tree building methods and models of sequence evolution to be specified. In conjunction, the purpose built

version of PW provides an extremely flexible interface for the visualisation of trees produced by MEGA and provides a solid basis for developing a generic eMLSA web application.

4.6.3 eMLSA.net database schema

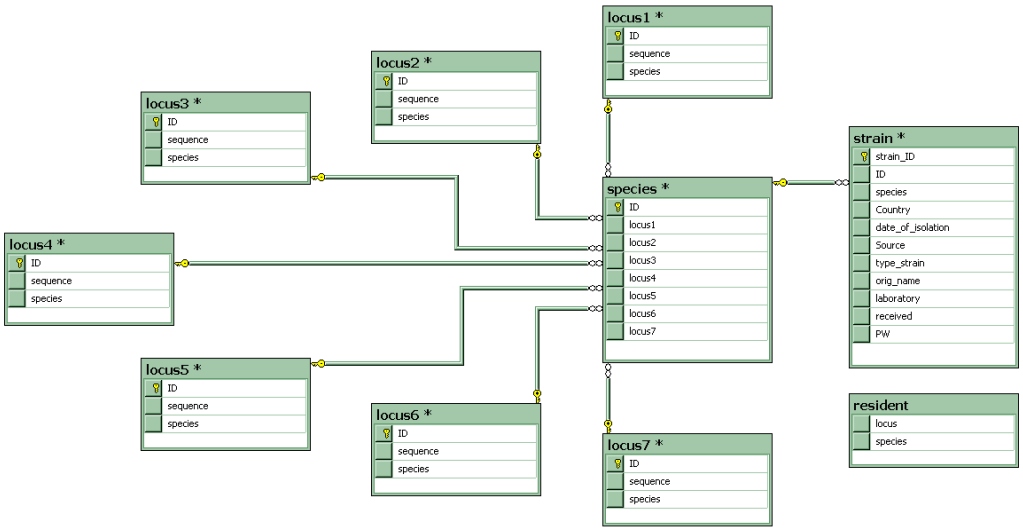


Figure 4.3 eMLSA.net database schema.

The database schema used for eMLSA.net (Figure 4.3) is, due to the nature of the methodology, very similar to that utilised within databases held at mlst.net described in chapter 3. Each of the loci is represented by a separate table, which contains unique examples of each sequence (alleles) and an ID. This ID is referenced in a column of the 'species' table, which contains an allelic profile (and corresponding unique ID) for each unique example of a set of seven allele sequences. The species 'ID' is linked to the 'strain' table, which contains, along with the species name and strain name, the date of isolation, country of origin and source of isolation. The strain table also contains a column 'PW' which details for each strain the codes used within PW to represent the colour and size of the terminal node and whether or not the label should be displayed. The 'resident' table is a look-up, which lists, for each locus, those species that could not be resolved on the gene tree for that locus. This table is used to assign the resident/foreign status of a user's sequence as described previously (resolution and lack of resolution is based on bootstrap values and a cluster is defined

as resolved if the bootstrap value of the node separating if from the neighbouring clusters is ≥ 80%).

This schema has been designed to minimise redundancy and allows the data held within tables to be extracted and joined together using server side scripting.

4.6.4 Definition of application logic

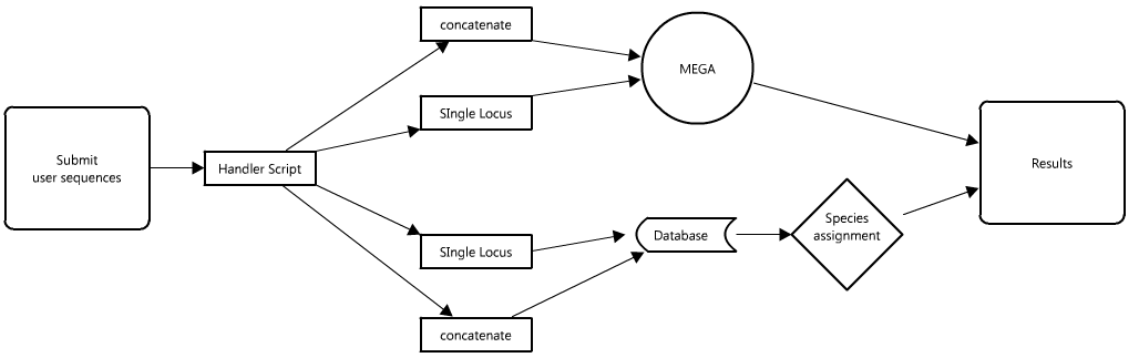


Figure 4.4 Application logic: eMLSA.net

Following database schema design and definition of tree building and visualisation methods, the application logic for eMLSA can be been defined as in Figure 4.4.

After submission of a set of query sequences (representing each locus), a single script will produce a set of text files. Firstly, the concatenated query sequences in conjunction with all concatenated database sequences will be produced in a single text file. Secondly, seven further text files corresponding to the individual locus query and database sequences will be produced. These text files will be used as input to the server-side version of MEGA for the production of the tree description in Newick format. In tandem, the same script will, based on the user query sequences, look up the closest matches to the concatenated database sequences and furthermore, each individual locus query sequence will be searched against all other alleles for that locus within the database. This two-pronged approach is necessary because the server-side version of MEGA only produces a Newick file, which by itself does not include any indication of % sequence identity. The species designation of the top scoring concatenated database match will be assigned as the concatenated query species and the top five database matches to the query locus sequences will be used to determine

the proposed resident / foreign status of the allele using the criteria described previously.

Following tree production, and species and locus status assignment, results will be presented to the user in a single webpage within which PW is embedded pre-loaded with the Newick file produced by MEGA. The results page should also allow users to access individual gene trees and furthermore, provide a generic text search of the contents of the strain database and the ability to download all data for further use.

4.7 Walkthrough of eMLSA.net

4.7.1 Portal homepage

In a similar fashion to mlst.net, eMLSA.net acts as a central portal to species-group specific sub-domains defined in the format: http://*speciesgroup*.eMLSA.net.

The eMLSA.net homepage, as seen in Figure 4.5, includes links to comprehensive instructions for using the site, background information on electronic taxonomy and MLSA, acknowledgements and references. The left hand side of the page contains links to individual species group databases. eMLSA.net has been optimised for viewing with most modern browsers, however PW has been reported to display unpredictable behaviour on certain browser/Java version combinations. We therefore recommend that the site is viewed using Mozilla Firefox (a free open source web browser) and that users have at least Version 1.5 of Java installed.



Figure 4.5 eMLSA.net homepage.

As a consequence of potential browser conflicts, within the eMLSA.net homepage, several JavaScript functions are used to check which operating system, browser and Java version a user is accessing the site with. Should the configuration fall below or differ from the recommendation warnings are given. For example, it can be seen that whilst taking the screenshot for Figure 4.5, I was viewing the eMLSA.net homepage using the Safari web Browser on Mac OS X. Furthermore, my version of Java is not the latest. Therefore, it is (strongly) suggested that I download and install Firefox (link provided) and visit the Java website to download the latest version (link also provided). In reality, most browser / recent Java versions are adequate to use eMLSA.net however, it is common practice to suggest specific requirements, especially when using Java or novel functional code within a site, which may disrupt even a small subset of individuals usability.

4.7.2 Viridans group sub-site: viridans.eMLSA.net

The viridans.eMLSA.net sub-site presents the user with a number of options as seen in Figure 4.6. The 'Introduction' section details the specifics of the viridans group eMLSA scheme, and is intended, generically, to act as a place for listing details which may be specific to a species-group MLSA scheme including; explanation of current species designation, loci utilised, primer sequences and PCR conditions, curator details and any references and acknowledgements.



Figure 4.6 Viridans.emlsa.net homepage.

The 'View tree of strains in the current database.' option presents simple access to viewing the tree based on the current contents of the database. The final two options provide access to user tree production. A blank form containing seven locus text boxes is presented when clicking 'To enter your own sequences'. However, for the purposes of demonstrating the use of the application, the penultimate option 'To use the tree building facility pre-loaded with an example sequence' presents a form with demonstration sequences pre-loaded and can be seen in Figure 4.7.

4.7.3 Querying the database with sequences

Additional to seven locus text boxes, a check box is present which allows a user to choose to add sequences obtained from publicly available genomes to those in the database based on the strain collection (Figure 4.7).

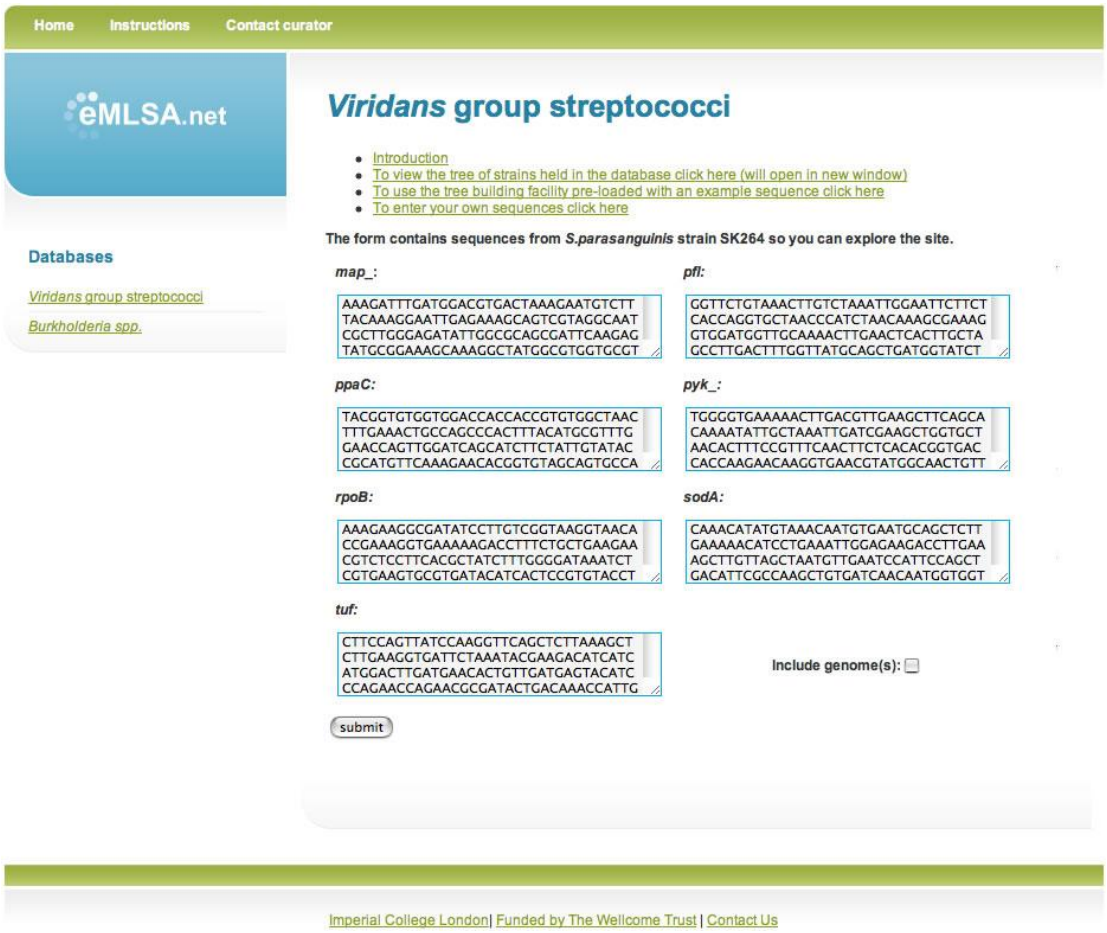


Figure 4.7 Demonstration locus sequences for viridans.emlsa.net

The strain sequences used for demonstration purposes are from *S. parasanguinis* (SK264) and consequently, the database sequences of this strain are not included when processing the demonstration query. Clicking 'submit' passes the sequences, via HTTP, to the web server and the text file production described in section 4.6.4 is undertaken. An interim page is presented to the user with a 'Click to view tree' option, and following this, the application flow as detailed in section 4.6.4 is undertaken. During processing by MEGA, and whilst the sequence searches are undertaken on the database, 'loading' icons are displayed to indicate to the user that processing is underway. Once complete, PW, embedded in the resultant page is launched and the Newick output from MEGA is loaded.

4.7.4 Viewing the species tree: The tree view

The results of the demonstration query can be seen in Figure 4.8. Additional to links to eMLSA subsites, a navigation menu is present to the left of the tree and the 'Reference Tree View' is displayed by default.

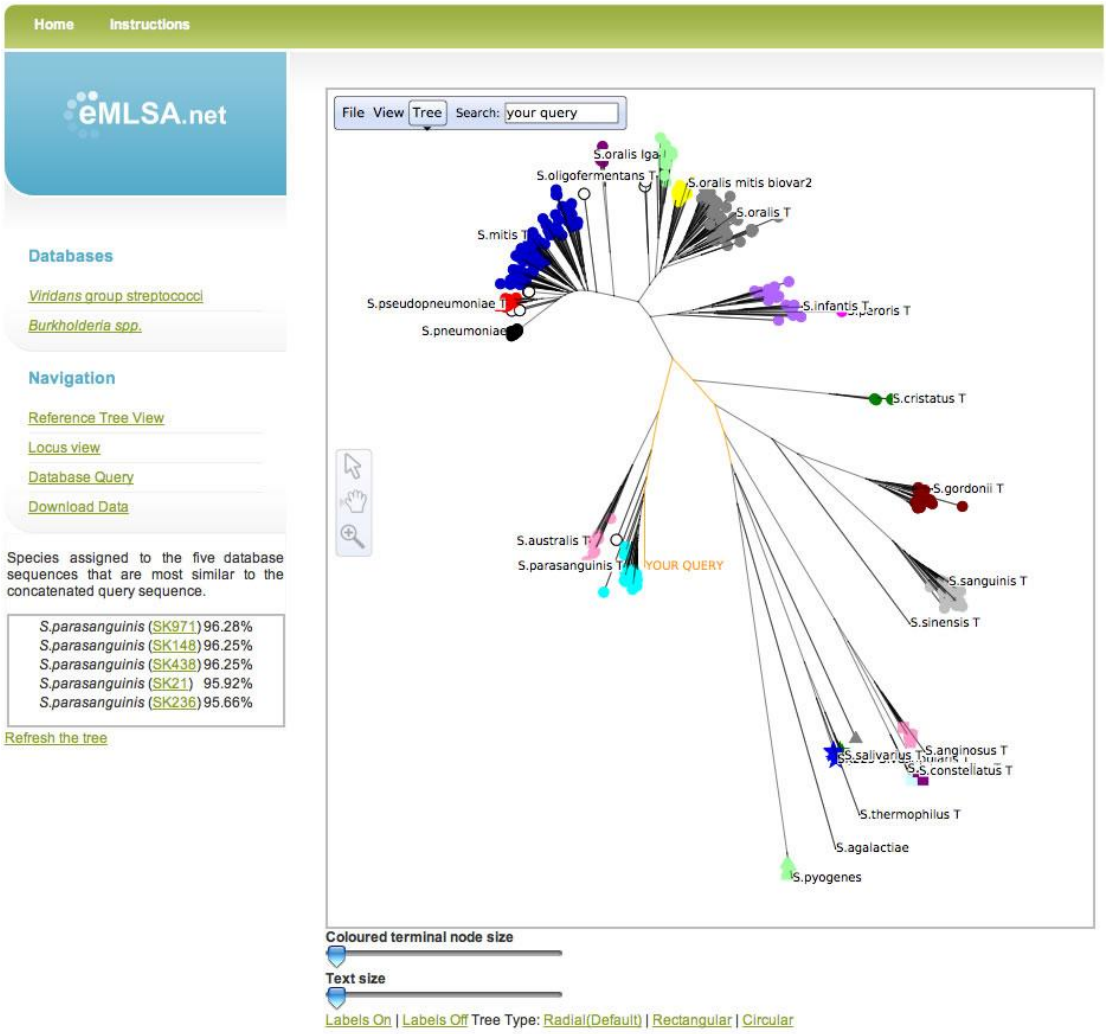


Figure 4.8 Sequence clusters and species assignment for viridans.mlst.net demonstration sequences.

Each terminal node on the tree represents a strain and each of the symbols used to represent the species clusters are coloured differently, in accordance with the reference tree (Figure 4.2A). Each of the type strains is labelled with the name of the species designation to make understanding patterns of clustering more intuitive for the user. The default tree type displayed is an unrooted radial tree and the query sequence is highlighted and labelled 'YOUR QUERY'. The PW menu allows various manual options, such as size of terminal node, type of tree etc. to be changed.

Within the PW window, to the left of the tree, three tools are present in a ‘tool panel’. These allow, from top to bottom, pointing and clicking to select nodes, clicking and dragging, and zooming in and out.

Below the PW window, HTML elements are present which allow the resizing of terminal node symbols and text size (by moving sliders), turning labels on and off and changing the type of tree displayed. These HTML elements all trigger JavaScript functions, which communicate with PW through the API to achieve the desired action. The list of species assignments of the five strains in the database with the most similar concatenated sequences, and their sequence similarity to the query sequence are shown below the main navigation menu. In Figure 4.8, the query strain is assigned as *S. parasanguinis* (as expected) since the concatenated sequences of the seven MLSA loci are located within this species cluster, and the top five most similar concatenated sequences are from strains assigned to this species.

Changing the tree type to a circular tree produces a view as in Figure 4.9 and similarly, changing the tree to rectangular provides a rectangular view of the tree. By zooming the tree can be explored in more detail as in as in Figure 4.10.

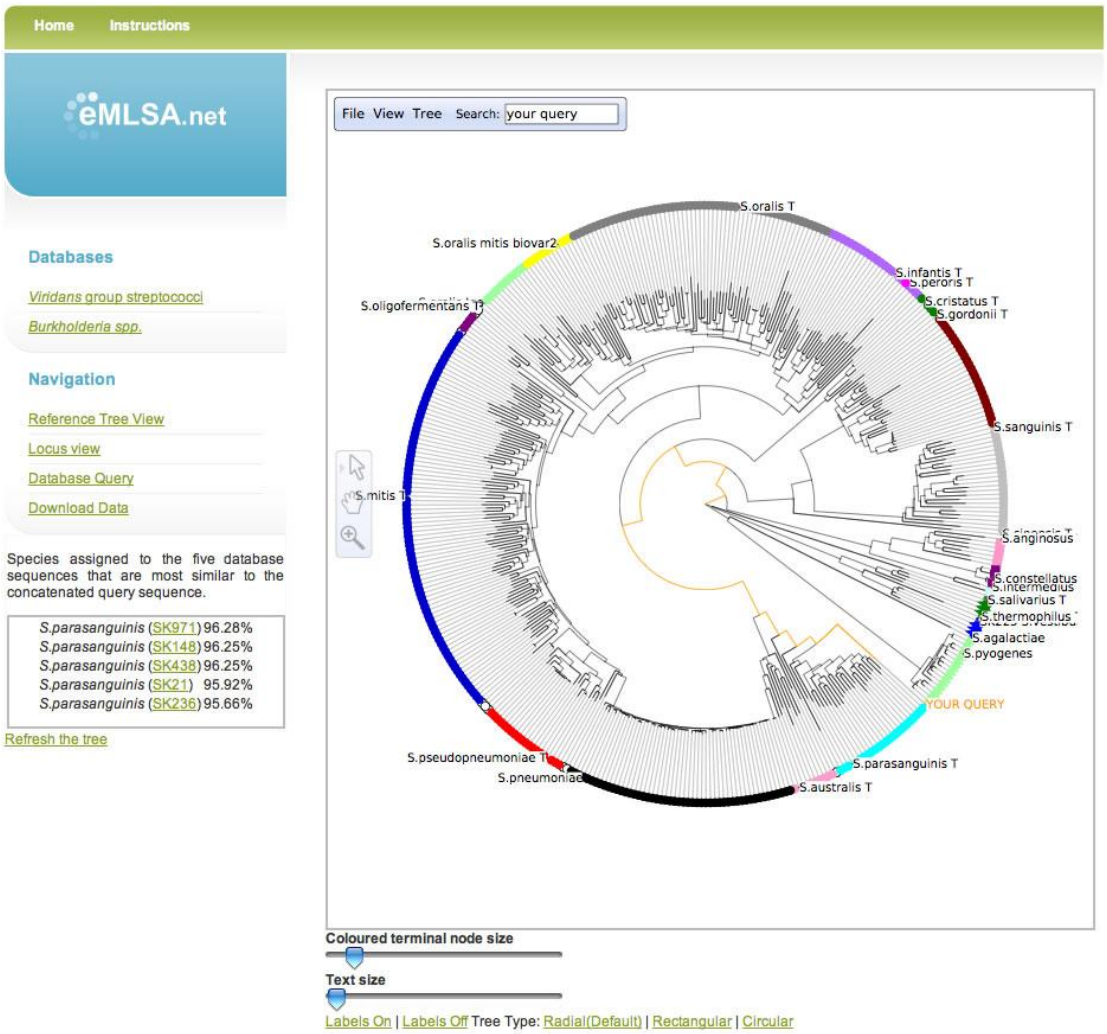


Figure 4.9 Circular tree type display.

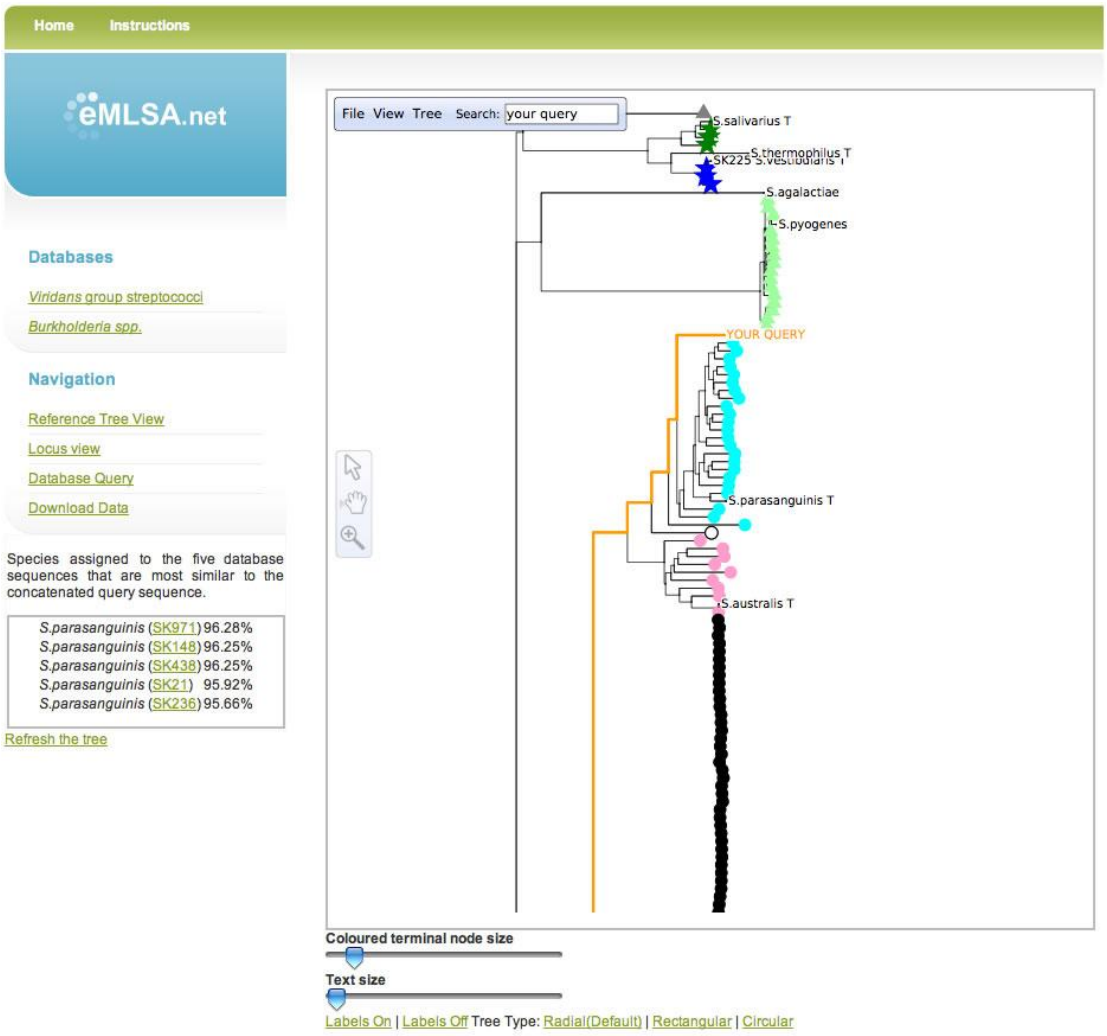


Figure 4.10 Rectangular tree type dislay and zooming in to query sequence.

4.7.5 Collapsing nodes on the tree

The ‘pointer’ tool within PW can be used to select individual nodes on the tree and allows the collapsing of branches to view in further detail sub clusters of the resultant tree. The collapse option allows a user to view a version of the tree with the Mitis group collapsed to highlight sub-clustering within the Anginosus and Salivarius groups as suggested previously and seen in Figure 4.11.

Other options in the menu displayed when selecting a node allow, cutting branches of the tree, copying the Newick format for a particular branch, flipping branches and rooting the tree based on the node selected. The options available are defined by an XML file, which is loaded into PW when initially viewed. This presents the possibility to define new node-specific functions and methods should they be needed in the future.

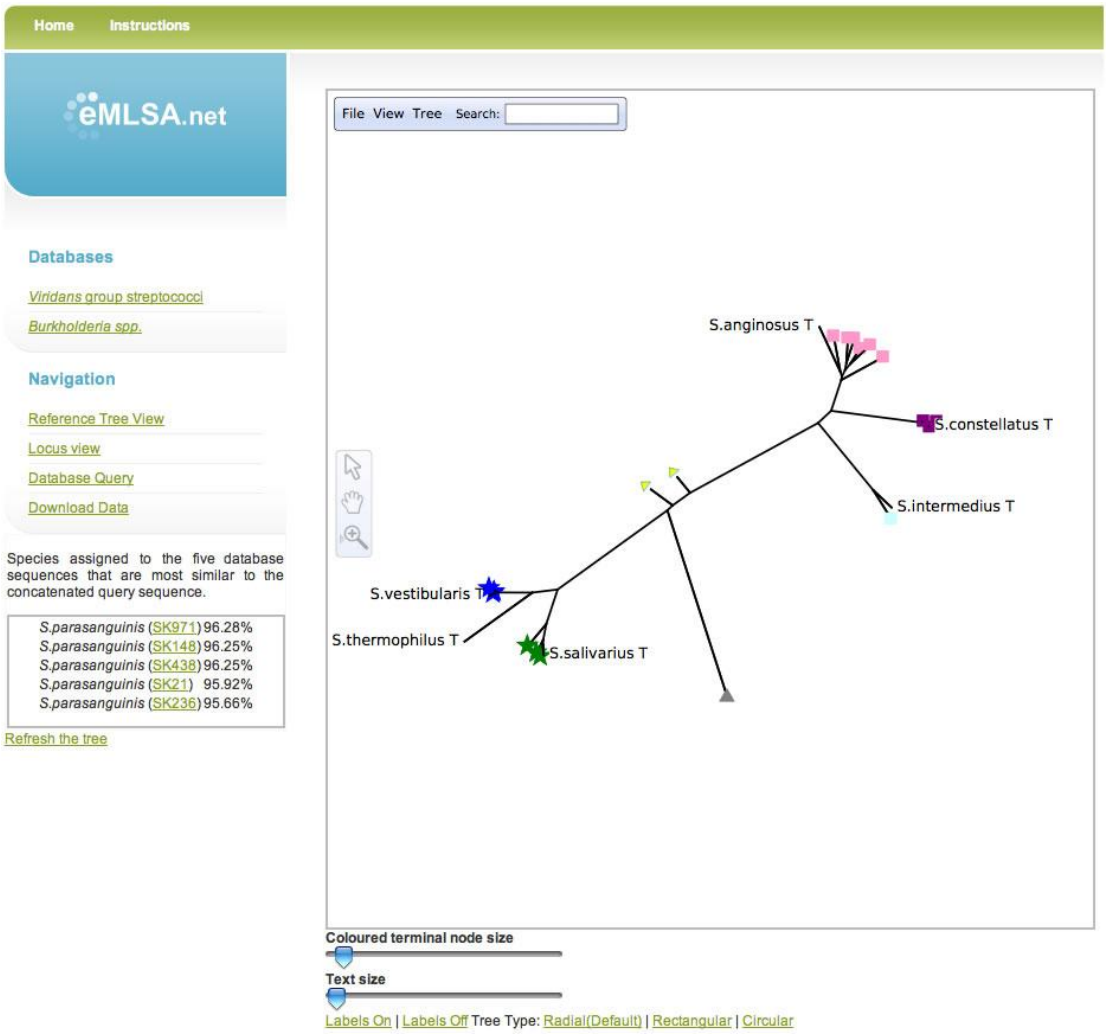


Figure 4.11 Collapsing the Mitis group branch of the tree to display sub-clustering within the Anginosus and Salivarius groups.

4.7.6 Viewing individual locus species assignments and trees: The locus view

Similarly to the AJAX design principles used in mlst.net v2 described in chapter 3, the eMLSA results page is composed of a series of DIV elements, which are contextually hidden or displayed depending on navigation choice. When clicking the 'Locus View' option in the Navigation menu, the 'Reference Tree View' DIV is hidden and the 'Locus View' DIV is displayed (Figure 4.12). The Locus View displays the resident, resident compatible or foreign status assigned to each of the individual alleles of the query strain. In the example, five of the individual loci are most similar to sequences from *S. parasanguinis* (resident alleles) or are resident compatible (for example, *ppaC*). However, the *pyk* query allele is most similar to a number of different species and is present within a sequence cluster that is well

resolved from the *S. parasanguinis* cluster (as recorded in the resident table of the database).

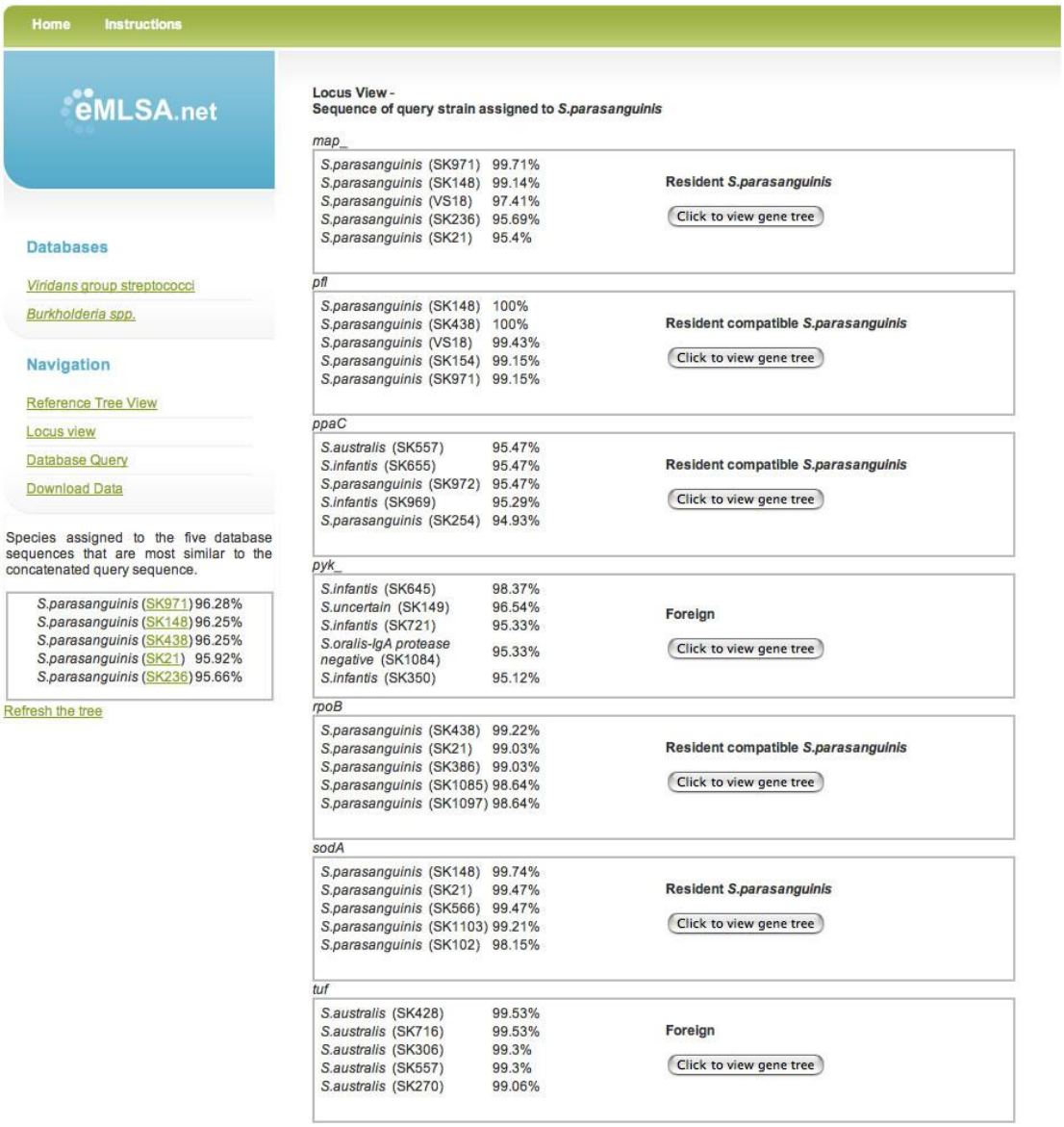


Figure 4.12 viridans.emlsa.net: The locus view.

Similarly, the query *tuf* allele sequence is most similar to alleles from strains assigned to the species cluster of *S. australis*, which is also well resolved. These two alleles are therefore assigned as 'foreign'. An option is provided to view the individual gene tree for each locus and Figure 4.13 shows the *pyk* gene tree zoomed in to highlight the position of the query sequence. The presence of two foreign locus sequences within the query may explain the position of the concatenated query sequence on the tree in Figure 4.8 as an outlier within the *S. parasanguinis* cluster.

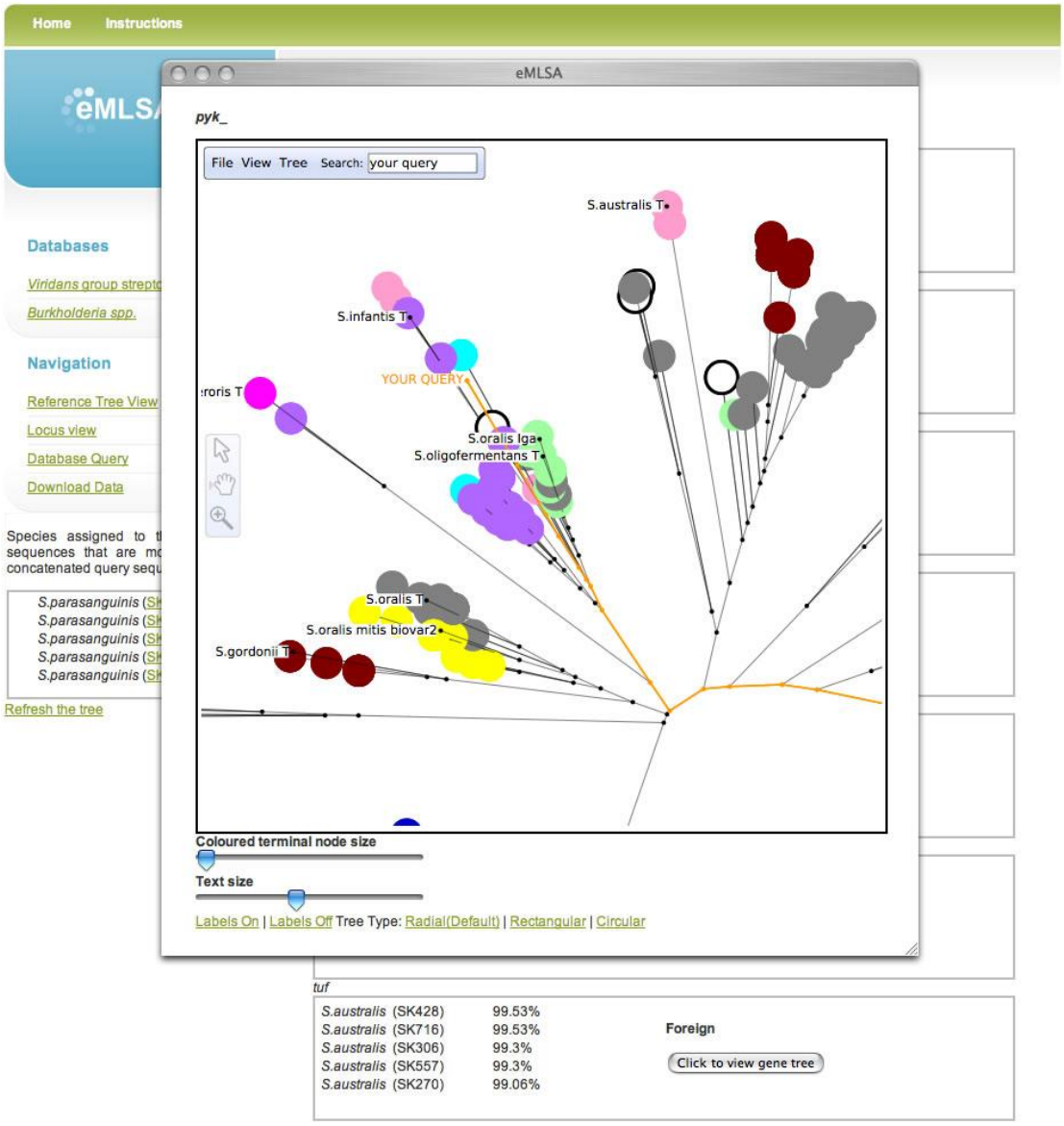
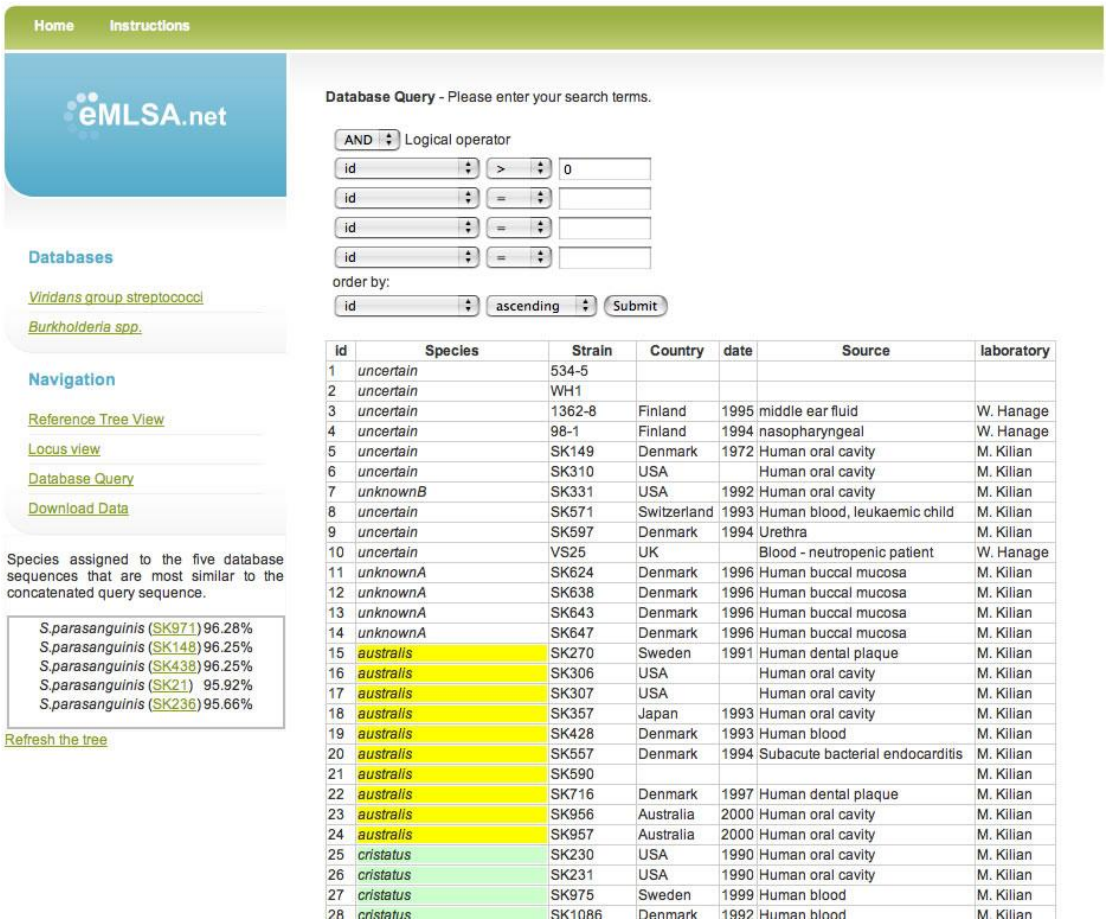


Figure 4.13 Location of foreign *pyk* query sequence within single locus tree.

4.7.7 Other options

Two further options are available within the Navigation menu. The ‘database query’ allows a standard form based query to be made on the underlying MLSA database, which retrieves and displays strain details (Figure 4.14). The ‘Download Data’ option provides links to download all data contained within the database (Figure 4.15). Sequences for all alleles of each locus or all concatenated sequences of all strains are available to download in common sequence formats (FASTA or MEGA) as well as comma or TAB delimited files. Furthermore, the entire database can be downloaded as an Excel file containing multiple worksheets containing species details as well as

locus sequences. Providing all data for download in this manner is common practice and allows users to undertake further analysis easily.



id	Species	Strain	Country	date	Source	laboratory
1	*uncertain*	534-5				
2	*uncertain*	WH1				
3	*uncertain*	1362-8	Finland	1995	middle ear fluid	W. Hanage
4	*uncertain*	98-1	Finland	1994	nasopharyngeal	W. Hanage
5	*uncertain*	SK149	Denmark	1972	Human oral cavity	M. Kilian
6	*uncertain*	SK310	USA		Human oral cavity	M. Kilian
7	*unknownB*	SK331	USA	1992	Human oral cavity	M. Kilian
8	*uncertain*	SK571	Switzerland	1993	Human blood, leukaemic child	M. Kilian
9	*uncertain*	SK597	Denmark	1994	Urethra	M. Kilian
10	*uncertain*	VS25	UK		Blood - neutropenic patient	W. Hanage
11	*unknownA*	SK624	Denmark	1996	Human buccal mucosa	M. Kilian
12	*unknownA*	SK638	Denmark	1996	Human buccal mucosa	M. Kilian
13	*unknownA*	SK643	Denmark	1996	Human buccal mucosa	M. Kilian
14	*unknownA*	SK647	Denmark	1996	Human buccal mucosa	M. Kilian
15	*australis*	SK270	Sweden	1991	Human dental plaque	M. Kilian
16	*australis*	SK306	USA		Human oral cavity	M. Kilian
17	*australis*	SK307	USA		Human oral cavity	M. Kilian
18	*australis*	SK357	Japan	1993	Human oral cavity	M. Kilian
19	*australis*	SK428	Denmark	1993	Human blood	M. Kilian
20	*australis*	SK557	Denmark	1994	Subacute bacterial endocarditis	M. Kilian
21	*australis*	SK590				M. Kilian
22	*australis*	SK716	Denmark	1997	Human dental plaque	M. Kilian
23	*australis*	SK956	Australia	2000	Human oral cavity	M. Kilian
24	*australis*	SK957	Australia	2000	Human oral cavity	M. Kilian
25	*cristatus*	SK230	USA	1990	Human oral cavity	M. Kilian
26	*cristatus*	SK231	USA	1990	Human oral cavity	M. Kilian
27	*cristatus*	SK975	Sweden	1999	Human blood	M. Kilian
28	*cristatus*	SK1086	Denmark	1992	Human blood	M. Kilian

Figure 4.14 viridans.emlsa.net: Database query view.

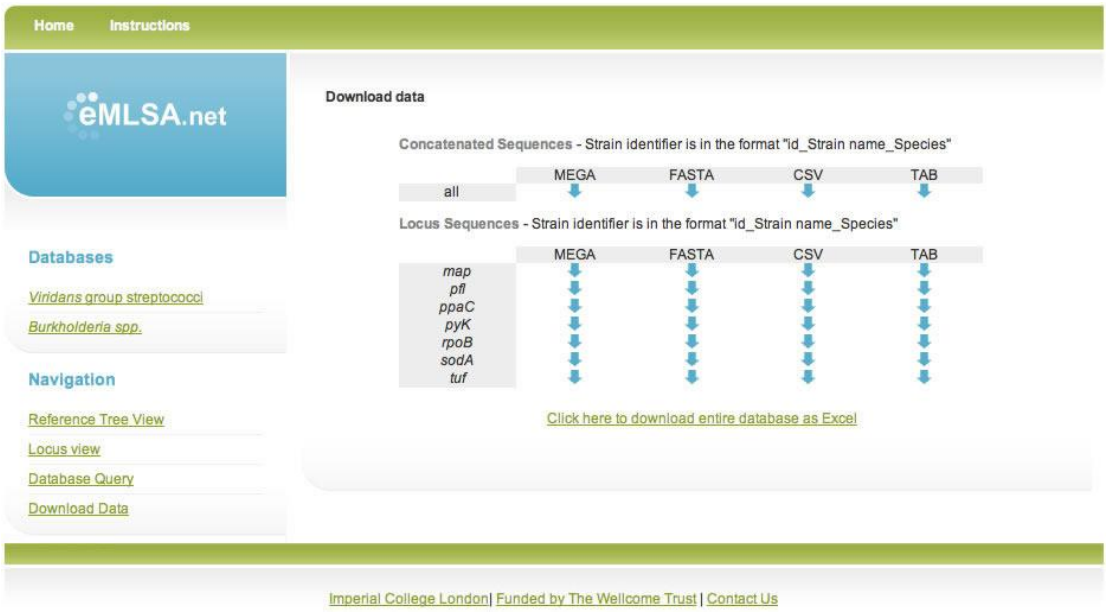


Figure 4.15 viridans.emlsa.net: Download data view.

4.8 Discussion

The generic eMLSA web application described has been developed to allow the assignment of bacterial strains to species via the internet. The viridans group sequences provided a particularly difficult dataset with which to prove the concepts of electronic taxonomy as current methods used for assigning strains to species within this group are problematic.

Using eMLSA.net, a user can query the database with their strain sequences and, based on the concatenated query matches, discover a species designation for their strain by investigating the position of the query on a tree. Furthermore a user can view the closest database matches based on sequence similarity. However, it is important to note that the position within, or closely allied to, a species cluster is the main criterion of species assignment rather than sequence similarity to a match in the database. There may be instances where a user's query falls mid-way between two clusters. This may constitute inclusion as a new species, or may simply be an indication that not enough diversity has been sampled within the initial MLSA dataset for a species designation to be apparent.

Single gene trees, in most cases, fail to identify sequence clusters (most likely due to lack of informative sites at a single locus or a history of interspecies recombination). However, the position of a query allele on a single gene tree does allow further investigation of the history of interspecies recombination between strains.

Only 16 strains clearly had foreign alleles within the viridans group MLSA dataset (Bishop et al., 2009). Such events could lead to a strain clustering on the tree, based on concatenated sequences, away from other strains of the species if the recombinational replacement is from an unusually diverse donor, or there is more than one locus that has received a recombinational replacement. This potential problem can be recognised within eMLSA.net by examination of the individual gene trees and the resident or foreign status of the alleles.

A server-side version of MEGA was used to provide a neighbor-joining tree for the viridans group sequences. For differing species group MLSA schemes, different tree drawing methods may be more applicable (e.g. minimum evolution) and MEGA

allows such methods to be provided. PhyloWidget represents a novel tree visualisation application, which has been key for the provision of the eMLSA framework.

The recognition by the developer of PhyloWidget (Greg Jordan) that such tools offer greater benefit when tightly integrated within a web environment should be applauded. The provision of an API for direct manipulation of the software directly from JavaScript (via AJAX), in combination with the development of a simple vocabulary for specifying precise graphical rendering of nodes has allowed the eMLSA.net application to offer a facility that mirrors to some extent the desktop analysis undertaken to initially define the eMLSA scheme. This combination of MEGA and PhyloWidget embedded within an AJAX web application offers a scalable, generic framework for the further inclusion of other species group MLSA schemes within eMLSA.net.

The viridans group MLSA scheme utilised seven loci, however the number needed for further schemes may vary and resolution of clusters may improve with increasing numbers of loci. The number is likely to differ depending on the extent of sequence diversity within the group and the extent of interspecies recombination. The eMLSA application has been designed to function with an unspecified number of loci, however, the time taken to produce trees with many more loci and many more strain sequences will increase and it may be necessary to alter the method of tree production to cope with ever increasing datasets.

One possible solution is to only include variable sites when passing the concatenated sequences to MEGA. Within the viridans group scheme there are 1596 of 3063 sequence positions that vary between all strains. This would immediately almost halve the amount of sequence data utilised while retaining all of the informative sites for tree production. Such a solution may become necessary with increasingly large datasets.

The major advantage of the MLSA approach for electronic taxonomy is that species clusters can be identified and circumscribed by visual inspection of a tree, without regard to preconceptions about the extent of diversity a bacterial species should have.

The process of tree production, visualisation and the assignment of alleles as resident, resident compatible or foreign is undertaken automatically.

However, although the eMLSA process is automatic, the assignment of strains to known species and the identification and acceptance of new species cannot realistically be automated as it requires the experience, knowledge and judgement of taxonomists. eMLSA.net provides a community tool with which those interested in a particular taxonomic group can share their experience and knowledge to provide a consensual approach to deciding whether new sequence clusters should be assigned as new species.

Developers of MLSA schemes for other closely related bacterial or other microbial species can host their data at eMLSA.net to take advantage of the generic nature of the web application. The setup of a new scheme is simply a matter of creating and populating a new database with sequences, species assignments and an indication of which species are not resolved on individual gene trees.

Currently, a scheme for *Burkholderia pseudomallei* and similar species is hosted at eMLSA.net and we are setting up a scheme for the Genus *Bifidobacterium* with data provided by The Pasteur Institute (Sylvain Brisse). Furthermore, the *Borrelia* spp. MLST scheme (Margos et al., 2008) utilises MLST loci that can be used for both MLST and MLSA approaches and therefore a dual functioning implementation of the software is planned, which utilises the MLST allele/ST assignment and interrogation methods detailed in chapter 3 in conjunction with the MLSA methods detailed in this chapter.

Chapter 5: Spatialepidemiology.net: Incorporating Web Mapping tools with Molecular Epidemiological Databases

Available: http://www.spatialepidemiology.net

Summary

The mapping tools integrated within mlst.net, described in Chapter 3, provide a simple means of interrogating the data held within MLST databases in the geographical context they were collected. Such mapping tools are generic and applicable to any molecular epidemiological database that contains a level of geographic resolution.

In this chapter I detail a number of developments that explore such potential. Firstly, the MLST-specific portal MLST-Maps, mentioned briefly in Chapter 3, is described. Secondly, an application produced for a European-wide *Staphylococcus aureus* genotyping project is detailed, the aim of which is to allow participants from 26 European countries taking part in the study, a simple means to view and analyse the distribution of particular genotypes. Thirdly, I detail an application being used to catalogue the world-wide prevalence of disease caused by the emerging amphibian fungal pathogen, *Batrachochytrium dendrobatidis* which, although not a molecular epidemiological database, has allowed the development of a simple framework for handling large datasets for display and interrogation using web mapping tools. Finally, and with the aim of encouraging users of databases such as mlst.net to submit higher geographical resolution with their data, I detail a facility for geocoding and basic map production for users who are unfamiliar with such concepts.

I demonstrate that such frameworks are flexible and, through the use of AJAX, highly extensible. Furthermore, such tools offer the potential to provide users of web-based molecular epidemiological databases with functionality more usually associated with dedicated Geographic Information Systems.

Note: Two publications that include descriptions of tools in this chapter are currently in submission and details are given within the text.

5.1 Introduction

Geographic Information Systems (GIS) such as the proprietary software ArcGIS [W77] or open source alternatives such as GRASS GIS [W78] are traditionally used to undertake the spatial display and analysis of data. Such software is installed on a user's computer and provides varying toolsets for the management, visualisation and analysis of geospatial data. GIS systems are used by individuals, or groups of individuals, to investigate relationships between data variables within a dataset. The aim of such study is to tease out and understand the influence that geography may have on the system under investigation.

Although very powerful, GIS systems are often limited to a single user on a single machine, and as such are not immediately accessible to large numbers of users without direct access to GIS software and expertise in its use. To investigate the spatial structure within web-based molecular epidemiological databases (e.g. mlst.net; Chapter 3), we require dynamic mapping tools that allow the display and analysis of the genotypic and phenotypic characteristics of a population, within the geographical context the data were collected, via the web.

'Web Mapping' is a term that has gained widespread acceptance over the preceding years and encompasses the process of designing, implementing and delivering maps on the web. Often used interchangeably with the term 'Web GIS' to denote the potential for analytical mapping, the use of web mapping is seen as a major advancement in cartography and presents new opportunities for real-time maps, cheaper dissemination, personalised map content, distributed data sources and sharing of geographical information (Neumann A, 2008). A myriad of techniques are now available to display data within web maps which broadly fall somewhere on a spectrum increasing in sophistication from static to dynamic (Kraak, 2001).

Static web maps are images displaying point data or thematic maps displaying aggregated data. Such images are produced using desktop software (for example GIS systems) and then made available for viewing via a website. As such they provide a simple geographic snapshot representing a statistic or set of statistics within a dataset. To progress beyond simple geographic snapshots and to allow interactivity with a map (and underlying data), methods that take advantage of 'page-click-page' web

navigation have been developed. In the same way that a series of webpages are linked together and navigated one by one, sets of static maps can be linked together to provide exploration of a dataset. Such an example is provided by the Los Alomos HIV Sequence Database, which provides a geographic search interface [W79]. Starting from a 'World view' with pie charts placed on major world regions (indicating HIV subtypes with sequences represented in the database), a user can 'drill-down' into the map by clicking to view Continent, then Country views. Sequences within the database from the chosen region can then be extracted and downloaded for further use.

Such methods provide a simple level of geographic interactivity, however, they are not, as such, 'dynamic' as the data is static and the images are simply linked together to provide an interface (the Los Alomos images are time-stamped to indicate date of production). Any data added to the database since the image was produced would not be immediately available for access.

Recently, truly dynamic web mapping has become a reality with the introduction of a number of tools, from a number of providers, allowing the production of customisable web maps that can be incorporated and further developed within web applications. These tools provide the potential for tight integration of maps with molecular (or purely) epidemiological databases to allow the geographic visualisation and analysis of genotypic, phenotypic and epidemiological data held within.

The development of web applications of this kind are the subject of this chapter, and can be accessed at http://www.spatialepidemiology.net. However, firstly the principles of dynamic web mapping and a further explanation, and justification, of the techniques used is warranted.

5.1.1 Dynamic mapping within Web Applications

Dynamic web mapping tools are available from a number of providers, the most well known being Google Maps [W50], Google Earth [W51], Microsoft Bing Maps [W52], Yahoo Maps [W53] and Open Street Map [W54]. These tools also provide application programming interfaces (APIs) for web developers to embed such maps within their own applications. Each API offers similar core functionality and the choice of which one is used is often solely based on a developer's preference. Within the web applications detailed in this chapter I use the Google Maps API and the whole earth browser, Google Earth which, although not strictly API based, does provide an XML vocabulary allowing content to be generated dynamically. This may seem unnecessarily Google-centric, however, careful consideration was made as to the best and most flexible mapping tools to use. The Google versions (Maps and Earth) are currently the most advanced and have the largest and most active development communities. However, as shall be seen in section 5.1.4.3, should other APIs become functionally more attractive or further developments demand a different provider, provisions have been made. In addition to mapping tools, a number of additional, externally provided, JavaScript frameworks are utilised to provide additional functionality within the developments described. The following section details these after a closer inspection of the principles of web mapping using the chosen Google tools.

5.1.2 **Google Maps** [W50]

Google Maps is an interactive web mapping application that allows a user, through an intuitive 'click-drag-zoom' interface, to navigate a map of the world. Google Maps provides a graphical representation of the earth using the Mercator projection, a flattened 2-D cylindrical view of the earth with the linear scale constant in all directions around any point. This preserves the shapes of small objects, however large objects are slightly distorted as the scale increases from the equator to the poles. Google Maps provides a set of eighteen pre-defined zoom levels, which allow a user to zoom from an initial view of the entire world to very fine scale geographical resolution. The key to this is the splitting of the different zoom level views of the earth into sets of images (each 256 x 256 pixels in size and often referred to in the context of web mapping tools as 'tiles'). The subset of images displayed is based on

the level of zoom chosen by a user. For example, at zoom level 'zero', a single tile covers the earth. At zoom level 'one', four tiles cover the earth; zoom level 'two', sixteen tiles etc. As a consequence, for any view, of any region in the world, only a small number of images are ever displayed. Google Maps is an AJAX application, which can 'react' to a user's actions (such as moving from one zoom level to another, or by clicking and dragging the map from left to right). Following an action, a request is sent to the Google web servers, asynchronously, using the XHR object (see thesis introduction) and an updated set of images are retrieved and displayed. This allows the provision of a fast, lightweight method of accessing images of any location, at any level of resolution, to provide a smooth and responsive user experience.

Google Maps also provide an API [W19] that allows developers to embed the application into their web pages. The following code demonstrates the incorporation of a simple map into a webpage:

```
1  <HTML>
2    <HEAD>
3      <script src=http://maps.google.com/maps?file=api&amp;v=2&amp;key=demo></script>
4      <script type="text/javascript">
5
6      function initialize() {
7        if (GBrowserIsCompatible()) {
8          var map = new GMap2(document.getElementById("map_canvas"));
9          map.setCenter(new GLatLng(53.727, -3.208), 5);}}
10
11      </script>
12    </HEAD>
13    <BODY onload="initialize”>
14      <div id="map_canvas">  </div>
15    </BODY>
16  </HTML>
```

Similarly to the inclusion of a separate JavaScript or CSS file in the HEAD of a HTML document (see thesis introduction), the Google Maps API is referenced in the document HEAD (Line 3). Using the DOM, the map is going to be placed in a DIV element in the BODY of the document, in this instance called '`map_canvas`' (Line 14). When the document loads in a browser, the '`onload`' DOM method (Line 13) calls the JavaScript function '`initialize()`' (Lines 6-9). Firstly, JavaScript checks if the browser being used to view the page is compatible (all current browsers are but legacy browsers may not be). If so, the location for the map ('`map_canvas`') is set (Line 8) and the initial centre point and zoom level are specified (Line 9). The initial view is defined using three values; the latitude and longitude (see below), which indicate where to centre the map (in this case the centre of the UK), and the zoom level (an integer between 0 and 18). Loading this into a web browser displays a map as in Figure 5.1.



Figure 5.1 Simple Google Maps API example.

The controls at the top left of the map allow navigation and zooming (as well as clicking and dragging) and the options at the top right allow the type of map to be changed. Each type of map represents a different set of tile images. For example, clicking 'satellite' will stream the relevant satellite tiles to the browser based on the current map view.

Individual point locations on the map are specified using latitude and longitude. Google Maps uses the World Geodetic System version 84 (WGS84) co-ordinate system provided by the US National Geospatial Intelligence Agency [W56]. WGS84 provides a standard co-ordinate frame for the Earth and a standard spheroidal reference surface for raw altitude data. Any point on earth can be represented by a latitude and longitude and its corresponding altitude extracted from the reference surface. WGS84 is a standard used in cartography, geodesy (a discipline concerned with the measurement and representation of the earth) and navigation, although other systems do exist.

In the simple example above, a latitude and longitude (and zoom level) was used to specify the initial view displayed on the map loaded into a web page. Latitudes and longitudes can be used for highlighting points on the map using 'placemarks', often also termed 'markers'. The Google Maps API provides a simple method for placing markers on a map as follows:

```
1  <HTML>
2    <HEAD>
3      <script src=http://maps.google.com/maps?file=api&amp;v=2&amp;key=demo></script>
4      <script type="text/javascript">
5
6      function initialize() {
7        if (GBrowserIsCompatible()) {
8          var map = new GMap2(document.getElementById("map_canvas"));
9          map.setCenter(new GLatLng(53.727, -3.208), 5);
10           var marker = new GMarker(new GLatLng(51.50, -0.126));
11           var html = 'Simple example 2'
12           map.addOverlay(marker);
13           marker.openInfoWindowHtml(html);}}
14
15     </script>
16   </HEAD>
17   <BODY onload="initialize">
18     <div id="map_canvas">  </div>
19   </BODY>
20 </HTML>
```

The marker to be displayed is defined by a specific latitude and longitude, in this case the centre point of London, UK (Line 10). Some simple HTML is defined to display text in the marker's Info Window (Line 11). The marker is then displayed on the map (Line 12) and the Info Window opened (Line13), as in Figure 5.2.



Figure 5.2 Displaying a marker and associated Info Window using the Google Maps API.

The Google Maps API contains several other methods that can be used by a developer to control how markers are displayed (e.g. using different images, allowing the marker to change when placing the mouse over and removing the marker). Further API methods allow lines to be drawn on the map (by specifying a start and an end latitude and longitude), overlaying images on the map (by specifying the latitude and longitude of all corners (the bounding box) of the image), and placing JavaScript 'listeners' on the map that react to a users input (for example detecting and responding when a user clicks a particular point).

Complex HTML can be rendered in a marker's Info Window including images and/or formatted text and hyperlinks. As the map is embedded in a webpage located on a developer's server, these hyperlinks can be used to call custom written JavaScript functions which could, for example, remove the current marker (or set of markers) and display a different one (or many). JavaScript could also be used to call server-side scripts to, for example, update other sections (DIVs) within the existing page (using the DOM) with the results of a database query. This general framework provides the tools for developer's to produce web applications that combine the power of a fully controllable map interface with their own data and functionality.

5.1.3 **Google Earth** [W51]

Google Earth (GE) is a desktop application presenting a user with a 3D representation of the earth. In a similar fashion to Google Maps it utilises a number of images (satellite and aerial photography) that, dependent on the zoom level and location, are placed in geographical context onto a virtual globe. These images are requested from Google's servers during navigation and as such GE is closely tied to the web for the delivery of content. For this reason, GE is often referred to as a 'Whole Earth browser'. This metaphor is extended further by the fact that GE allows a large amount of additional content to be loaded from the web through 'Data layers', some of which are permanently accessible from the 'Layers' panel (e.g. road names or weather conditions) and can be turned on or off as required (Figure 5.3). Multiple data layers can be loaded at one time allowing disparate sets of data to be viewed simultaneously.

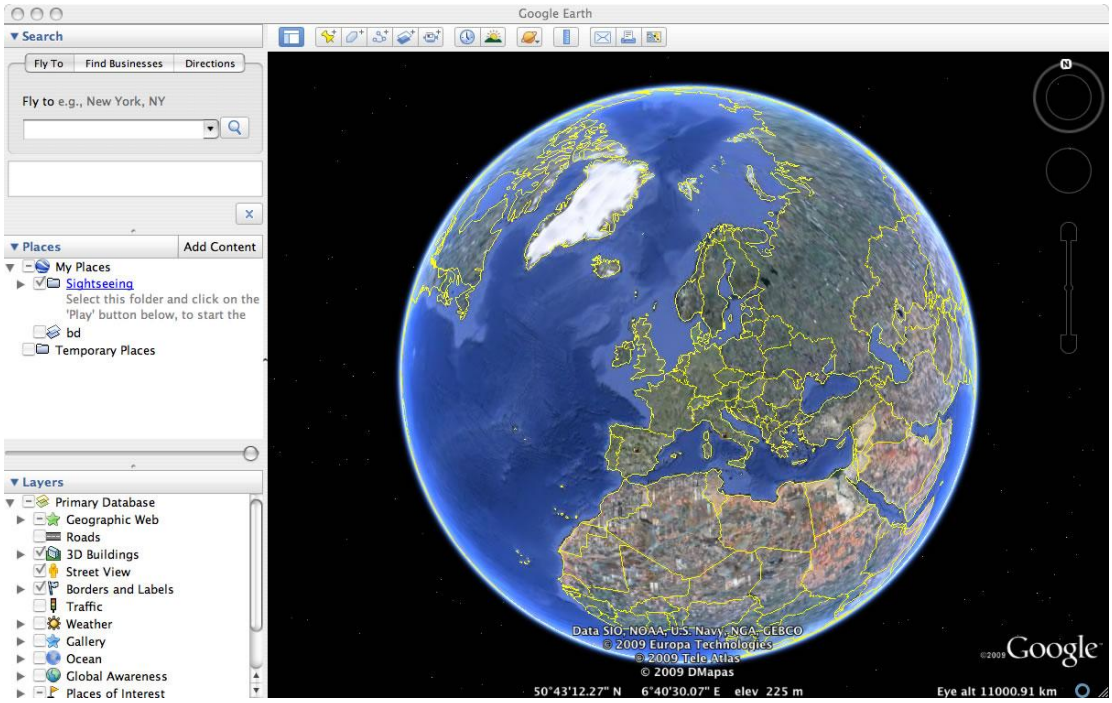


Figure 5.3 The 'Whole Earth browser' Google Earth.

A key feature of GE is the ability for users to create, save and load their own content. This can be done in two ways. Firstly, using the menus in GE, a user can place markers (when referring to GE these are termed 'Placemarks') on the globe simply by clicking a location. Content can be entered (text, images etc.) which is displayed within an Info window when a placemark is clicked, in a conceptually similar manner to markers on Google Maps. GE also provides the ability for a user to create 'Paths' (a term denoting a line or series of lines connected together), polygons (multi-edged shapes which can be filled with colour) and overlay images. As the representation of the earth is rendered in 3D, polygons and lines can also be 'extruded' by specifying a height (in metres), or altitude, at which they should be placed above the 'ground'. Once a user has created their content, it can be saved as a 'KMZ file', which can be re-loaded subsequently. KMZ files can be opened in any copy of GE and can therefore be made available via the internet for sharing. The second method for a user to produce content is by generating KMZ files outside of Google Earth.

In reality, a KMZ file is just a zip (compressed) file with a different extension (changing the extension from '.KMZ' to '.zip' allows opening with standard decompression software such as Winzip [W57]). Within the compressed file is a '.KML' file which is a standard text file containing data, formatted in a fashion understandable by GE. KML stands for Keyhole Markup Language and, like HTML

and XML, is based on a vocabulary of TAGS used to format plain text and can easily be created in a simple text editor:

```
1       <kml xmlns="http://www.opengis.net/kml/2.2">
2         <Placemark>
3         <name> Simple KML example </name>
4         <description> This is displayed in the Info Window </description>
5               <Point>
6                <coordinates>53.727, -3.208,0</coordinates>
7               </Point>
8         </Placemark>
9       </kml>
```

The simple KML description above has a root opening and closing TAG (Lines 1 and 9). The 'Placemark' TAGS define that the content within is to be rendered as a single point placemark (Lines 2 and 8). The placemark's name is defined (Line 3), and a description specified which is displayed in the Info Window when the placemark is clicked (Line 4). The latitude and longitude coordinates for where to position the point are then defined (Lines 5-7). When this is loaded into GE it renders as in Figure 5.4.

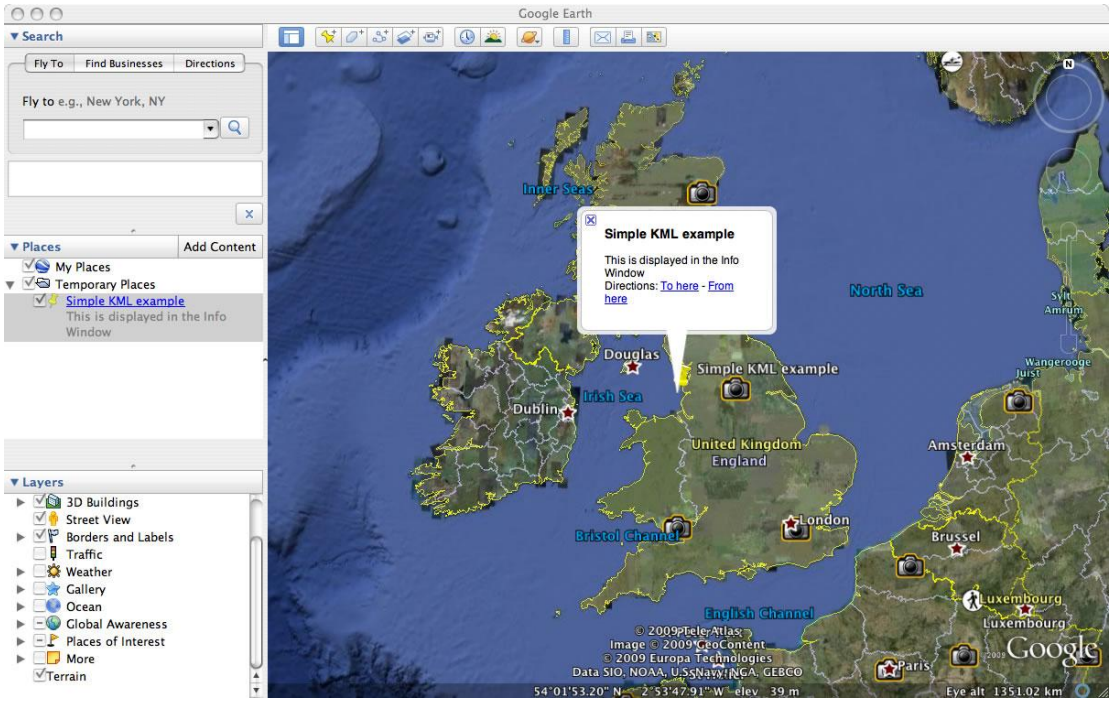


Figure 5.4 Simple KML example displayed in Google Earth. The 'Places' section of the interface (to the left of the map) contains a reference to the KML file, allowing multiple files to be loaded and switched on / off as required.

The KML specification [W58] defines the vocabulary of TAGS, allowing user-generated content to be created outside of GE. Furthermore, as KML is simple text, its generation can be undertaken within a web application (using server-side scripting) allowing the provision of data, viewable in GE, dynamically.

5.1.4 JavaScript libraries

A JavaScript library contains pre-written JavaScript functions that can be incorporated within AJAX applications. Rather than re-writing functions each time a new application is built, a library can be linked in the <HEAD> of a HTML document providing access to commonly utilised functions. A number of developers have produced libraries and made them available for others to use. Each available library tends to include functions that are focussed on a particular area of development. Within spatialepidemiology.net I make use of the following open source libraries.

5.1.4.1 **Prototype –** [W59]

The Prototype library, created by Sam Stephenson, provides an AJAX framework and other utilities. Prototype provides a set of commonly used functions ranging from programming shortcuts to major functions dealing with *XMLHttpRequest* (XHR). Prototype also provides libraries to support classes and class-based objects (allowing object-oriented programming), something that the JavaScript language does not natively have.

5.1.4.2 **Script.aculo.us** [W60]

Script.aculo.us is a set of JavaScript libraries focussed on enhancing the user interface of web applications. It provides a visual effects engine, a drag and drop library (including sortable lists), visual controls (AJAX-based auto-completion for form fields, in-place editing and sliders) and several other effects based controls. It is most often used in conjunction with Prototype as many of the functions build on novel classes provided by Prototype.

5.1.4.3 **Mapstraction** [W61]

As mentioned previously, the majority of web mapping tools (Microsoft, Yahoo, Google etc.) provide similar functionality and the description in section 5.1.2 could quite easily be amended to refer to any / each of them. I am conscious that, particularly within academic research or for public health applications, there may be concerns regarding using tools provided by Google (a commercial company) to

display data. I am not convinced this is a real issue as when embedding a map within a webpage, data is only ever requested from Google (i.e. map tiles) and not sent (and this can be observed by monitoring web server activity). However, to pre-empt any such concerns and to make the mapping frameworks as flexible as possible I have incorporated the Mapstraction JavaScript library.

Mapstraction provides a common API for other JavaScript mapping APIs. This allows developers to code their applications once, and then easily change the mapping provider based on project needs, terms and conditions and required functionality.

As a demonstration, a single page displaying the same data as in figure 5.2, using the Google Maps API and Mapstraction can be accessed at the following URL:

http://www.spatialepidemiology.net/thesis/ms.html?maptype=google

The initial API provider is set using the 'maptype=' part of the URL, which is read by a custom JavaScript function and defines which one to use. If we change the 'maptype=' to yahoo the same point and location are displayed, however this time using the Yahoo Maps API (see Figure 5.5). The API specified can also be change to 'maptype=microsoft' to use the Bing Maps API or 'maptype=openstreetmap' to use the Open StreetMap API.

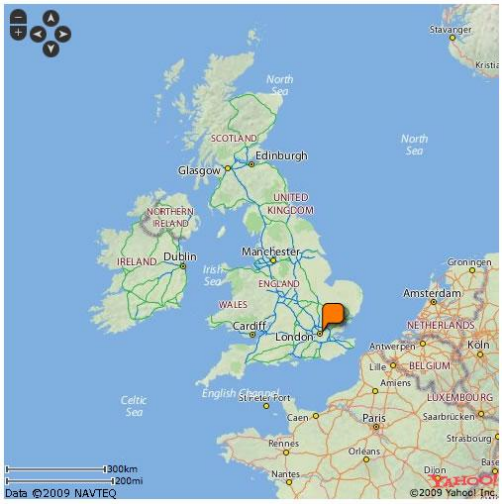


Figure 5.5 Simple map example displayed using Mapstraction and Yahoo Maps. Available: http://www.spatialepidemiology.net/thesis/ms.html?maptype=yahoo

As mentioned, each API offers a similar set of core functions, however there are differences within the accessory functions offered [W85] and I have chosen initially to use the Google API due to its enhanced functionality (and larger development community). However, incorporating Mapstraction into the applications allows the trivial switching of map providers should this be warranted. Furthermore, should functionality unique to a particular API be needed, combinations of map types could be used within the same webpage (Figure 5.6).

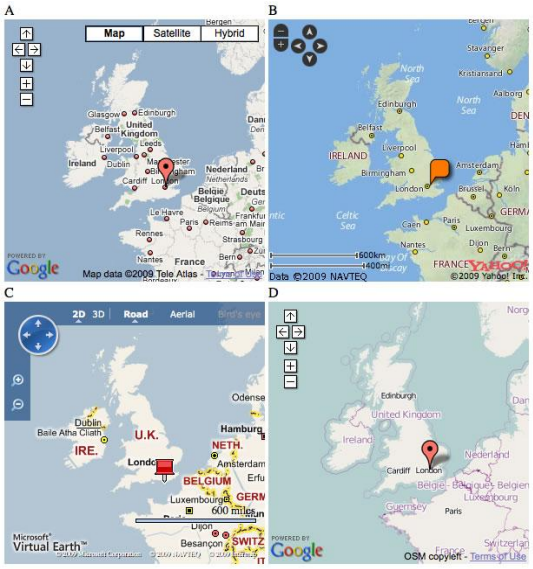


Figure 5.6 A single webpage utilising multiple web mapping APIs. A) Google Maps API. B) Yahoo Maps API. C) Microsoft Bing Maps API and D) OpenStreetMaps API. Available: http://www.spatialepidemiology.net/thesis/ms2.html

5.2 Spatialepidemiology.net

The coding frameworks described in the introduction provide the necessary tools to develop web applications that allow the mapping and interrogation of any epidemiological database that contains any level of geo-located data (from continent down to latitude and longitude). spatialepidemiology.net was developed to act as a central portal to the map-based developments which have been undertaken and described in this chapter. It was envisioned that each of the applications would be stand-alone and address the needs of a specific epidemiological database (or databases) and user group (or groups). Furthermore, the provision of a portal allows the development and addition of generic tools that are not database or user specific.

The following sections detail three stand-alone applications and a generic map making and geocoding tool aimed at giving access to basic geographic data, such as latitudes and longitudes, for users who may be unfamiliar with such concepts.

5.2.1 Multilocus Sequence Typing Maps: MLST-Maps

Available: http://maps.mlst.net

5.2.1.1 Introduction

The additional developments characterising version 2 of the software located at mlst.net, detailed in chapter 3 of this thesis, include the ability to interrogate the *Streptococcus pneumoniae* MLST database using Google Maps. mlst.net v2, as mentioned, is currently in BETA testing, however, this map-based method of viewing data was introduced (in an albeit more limited form) to all databases hosted at mlst.net some time previously (June 2007). Within each organism-specific site a separate page, which is linked to the home page, provides the ability to view the mapping interfaces (e.g. http://saureus.mlst.net/earth/). This provided a way to gently introduce the communities of users to the geographic browsing of MLST data. Concurrently, a central 'map-centric' MLST portal was introduced under the banner of 'MLST-Maps' (http://maps.mlst.net and linked to from spatialepidemiology.net). MLST-Maps, like the central 'eBURST-centric' portal for MLST datasets (http://eburst.mlst.net), offers a tool-specific view of multiple MLST datasets, easily accessible in one place.

The MLST-maps homepage provides a set of instructions, and links to view map interfaces to MLST databases. Clicking the 'View MLST maps' link from the menu on the homepage lists those databases that can be viewed (Figure 5.7).

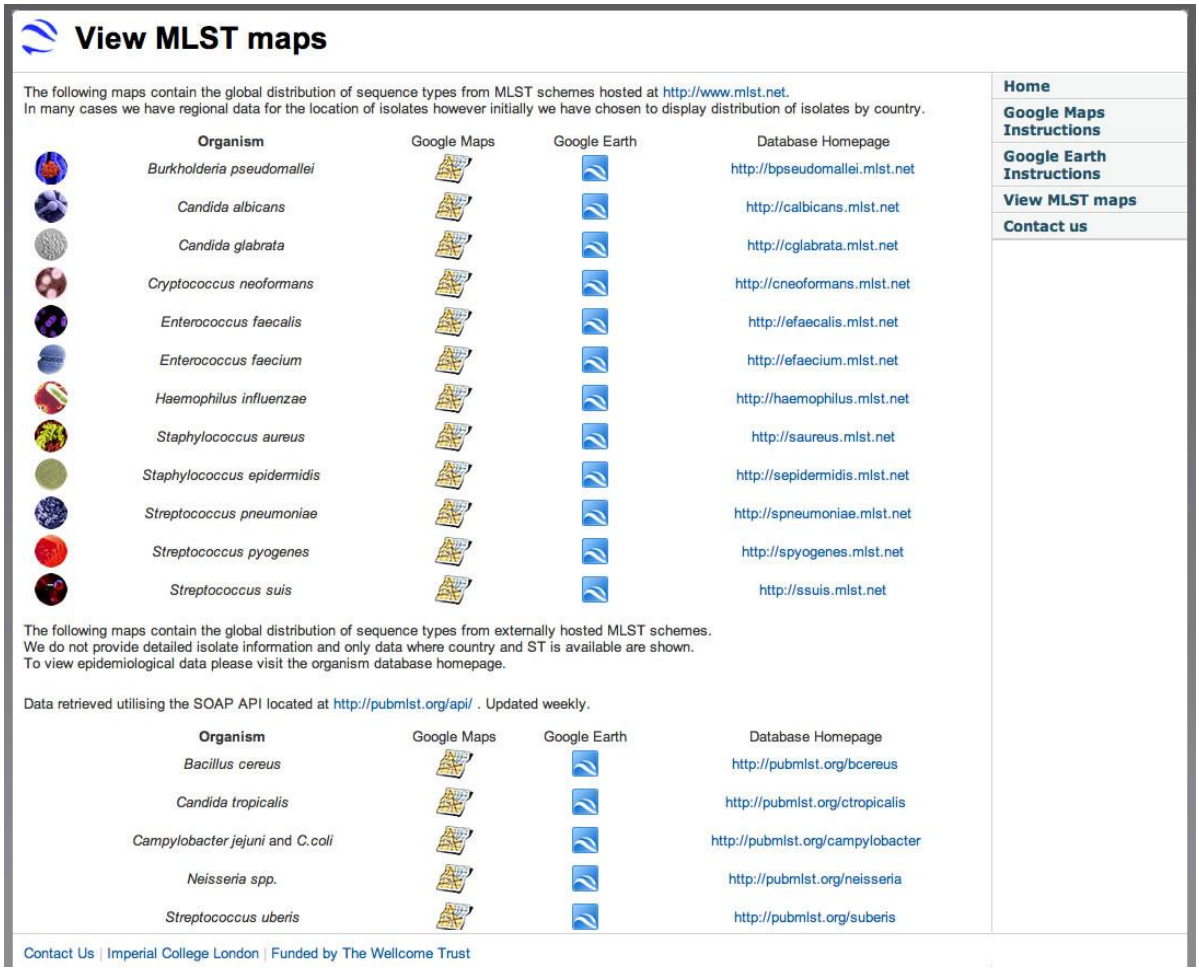


Figure 5.7 MLST-Maps: Database access page.

For each database, links to view using Google Maps or Google Earth are provided along with a link to the database homepage. In addition to the databases hosted at mlst.net, access to several databases hosted at pubMLST.org is provided. Only those pubmlst.org databases which contained ≥ 500 isolate records at the time MLST-Maps was released were included, excepting the *Helicobacter pylori* database which contains almost twice as many STs as isolate records and was not included due to lack of any geographical data. Data is retrieved from pubmlst.org via the Simple Object Access Protocol (SOAP; [W62]) server provided by the manager, Keith Jolley [W63]. A SOAP server, in a similar fashion to a JavaScript API, allows the retrieval of data programmatically from a data provider. The difference between SOAP servers and JavaScript APIs is that SOAP is generally focussed on server-side and desktop programming languages such as PHP/Perl, C++, Java etc., and is not as readily amenable for web development techniques such as AJAX. However, using PHP/Perl it is possible to retrieve the data and format it into a data structure that is compatible with AJAX. Using the pubmlst.org SOAP server it is possible to retrieve allele

sequences, STs and a list of corresponding isolate IDs. It is also possible to retrieve basic epidemiological data for each isolate, including country of origin, by querying the SOAP server using an isolate ID, one at a time but not in bulk.

There are valid reasons for limiting the data access in such a way. Each pubMLST.org database has a species-specific 'Data Policy' document stating that the data within is owned by the submitters and in some cases may contain unpublished data. To protect the data owners, full copies of the databases can only be obtained by emailing the Webmaster (Keith Jolley) who then requests permission(s). Therefore, for the purposes of MLST-Maps, for each isolate in each database used, only the id, species, ST and country of origin is accessed allowing similar functionality to that offered for the mlst.net databases.

5.2.1.2 Google Map interface

Clicking the 'Google Map' link for a database presents a similar interface to that provided in mlst.net v2 (Figure 5.8). Every country where isolate data has been deposited in the corresponding MLST database is represented with a marker on the map. The colour of the marker is based on the percentage of all isolates present from each country. In Figure 5.8, it can be seen that a large proportion of isolate information submitted to the *S. aureus* database originates from England (48%: 632 of 1314 isolates) highlighting geographical sampling biases. The key to the right of the page can be used to uncheck and remove markers from the map to explore further.

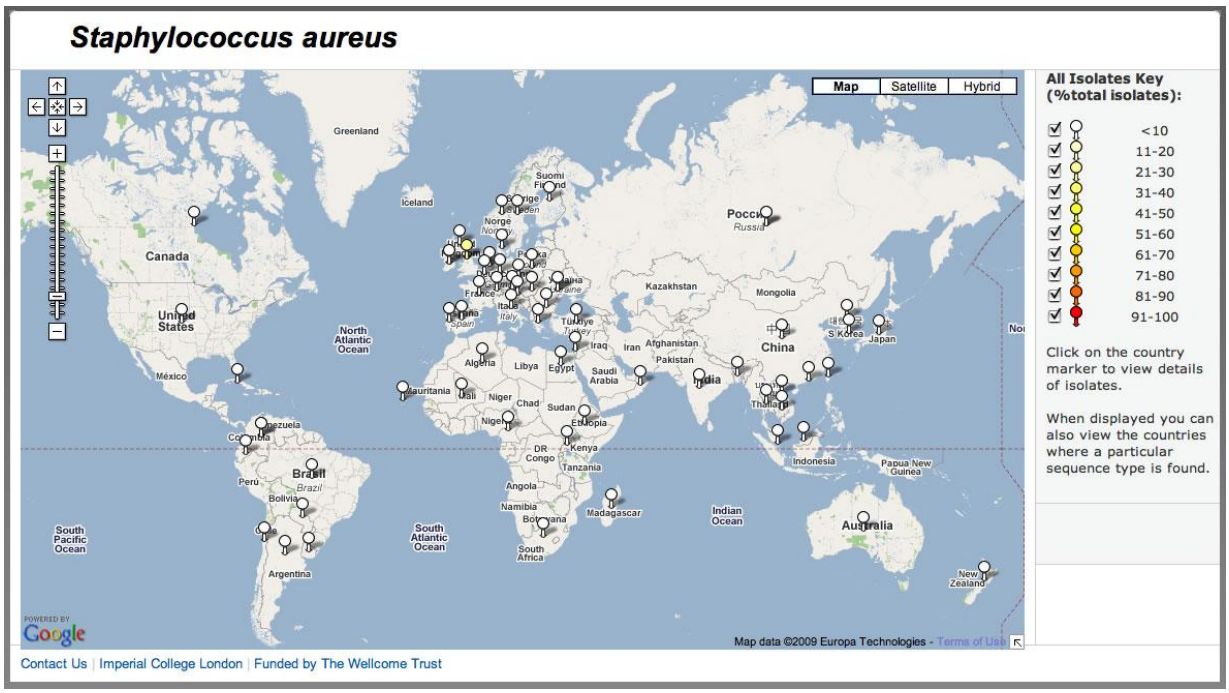


Figure 5.8 The *Staphylococcus aureus* MLST database viewed using Google Maps at MLST-Maps.

Clicking a marker opens an Info Window and provides a link to 'View STs'. Clicking centres and zooms the map to the country of choice and loads the ST data into a hidden DIV element below the map. As with MLST.net v2, a list of STs is shown, together with their allelic profiles and the number of isolates (and number of countries) found elsewhere and a link to view the global ST distribution. Clicking the link zooms the map to the global view and the distribution of the ST is shown as a series of markers. Figure 5.9 shows the global distribution of *S. aureus* ST 239 based on an initial query of Thailand. The other countries where ST 239 is found (and the number of isolates in each country) are displayed to the right of the map for clarity. eBURST can also be launched containing just the country data chosen or, using comparative eBURST, comparing STs from the country chosen to the entire MLST database.

A link to view epidemiological information deposited for isolates form a country is present below the map, however, only for those databases hosted at mlst.net.

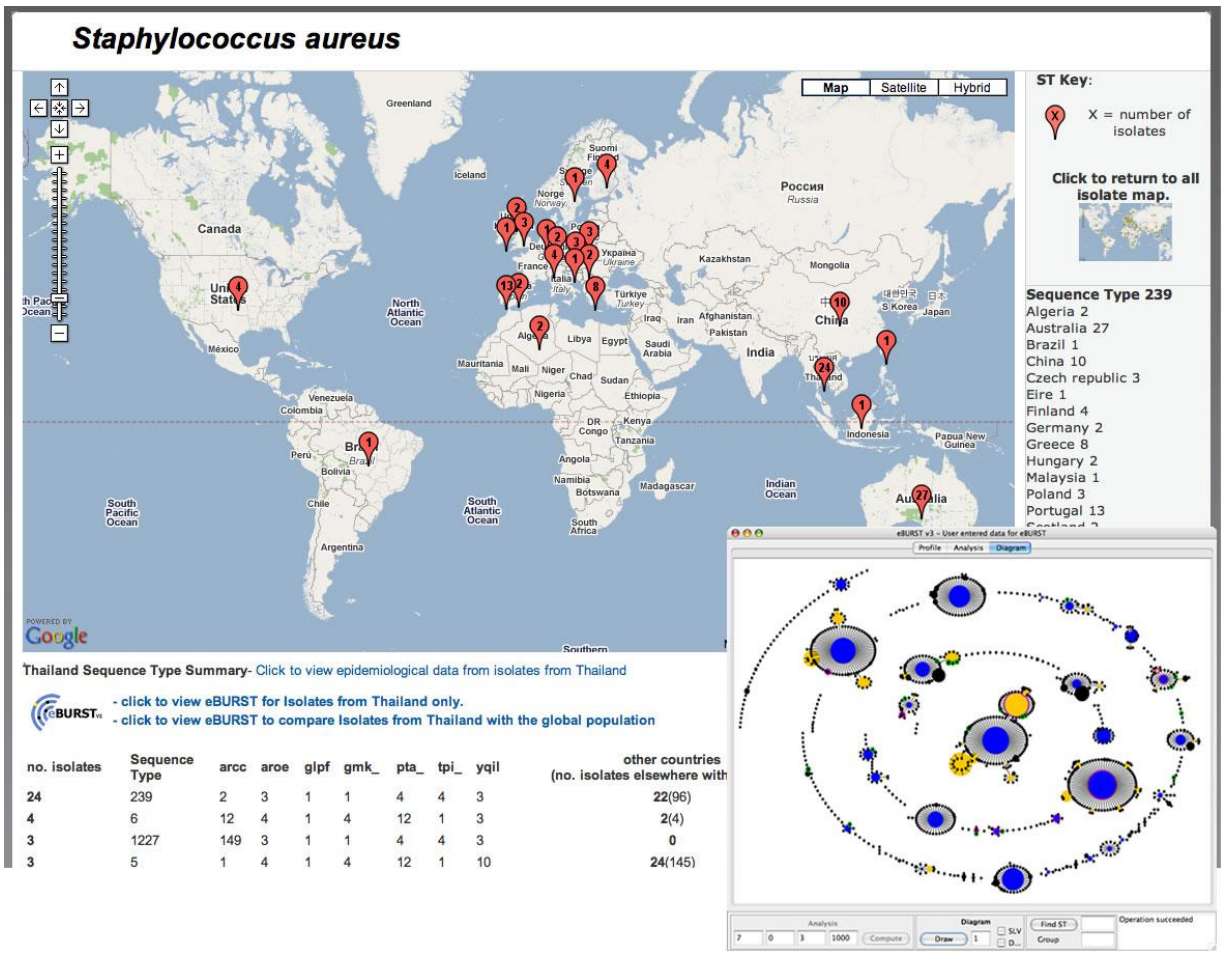


Figure 5.9 Global distribution of *Staphylococcus aureus* Sequence Type 239 and eBURST population snapshot.

5.2.1.3 Google Earth interface

MLST-Maps allows MLST databases to be viewed within Google Earth. When the relevant link is clicked (Figure 5.7) a KMZ file is generated (using server-side scripting) which contains a current snapshot of the data held in the corresponding database. Each country from where isolate data has been deposited is highlighted with a placemark, located at the centre point of that country (the 'centroid'). Clicking on the placemark opens an Info Window detailing the number of isolates from the country and a link to view STs (Figure 5.10). Unlike Google Maps, which is embedded in a DIV within a webpage, and subsequently, through AJAX, can communicate and update another DIV with query results (a list of STs), Google Earth, as stand-alone software, must launch a web browser to view webpages (Figure 5.11).

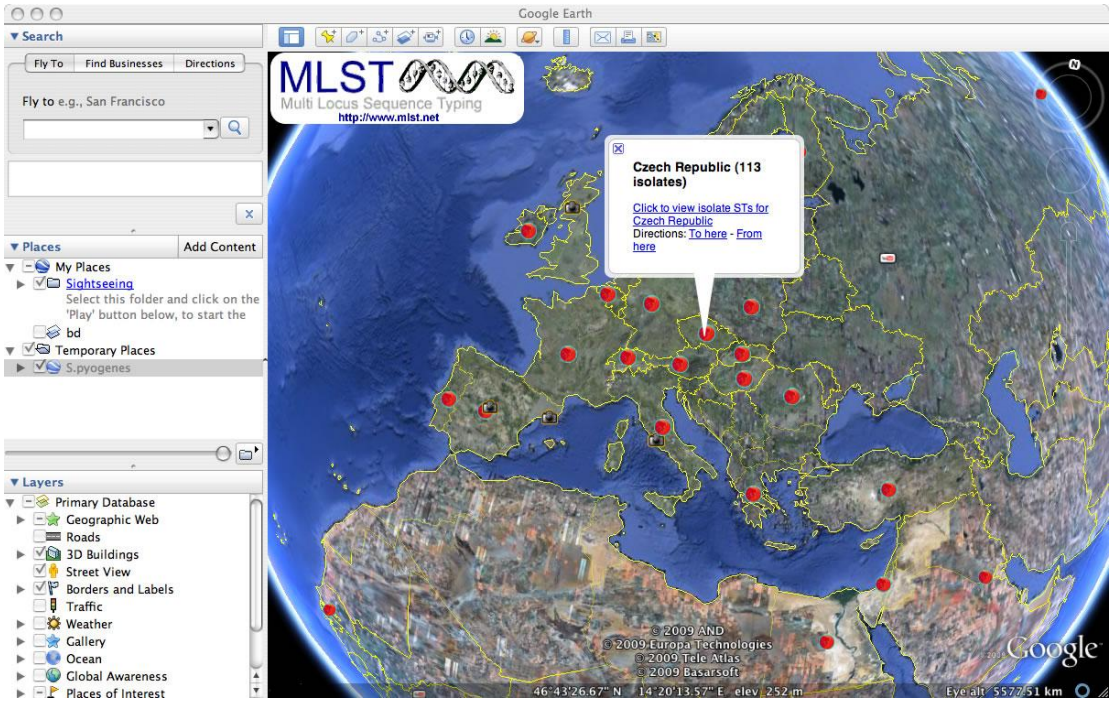


Figure 5.10 *Streptococcus pyogenes* MLST database viewed within Google Earth.

For each ST, the number of isolates from the country are listed along with the allelic profile and the number of isolates found elsewhere (and number of countries) with that ST. Furthermore, eBURST can be launched pre-loaded with all STs from that country or comparative eBURST can be used to compare the country STs to the entire database.

To view the global distribution of a particular ST, a user must use the Google Maps interface highlighting a lack of functionality within the Google Earth view. Currently, the potential for two-way communication between a web application and Google Earth (as stand-alone software) is limited. However, KML has an active and enthusiastic user and development community and it is expected that functionality will increase in the future. At present we recommend the use of the Google Maps interface for viewing MLST databases.

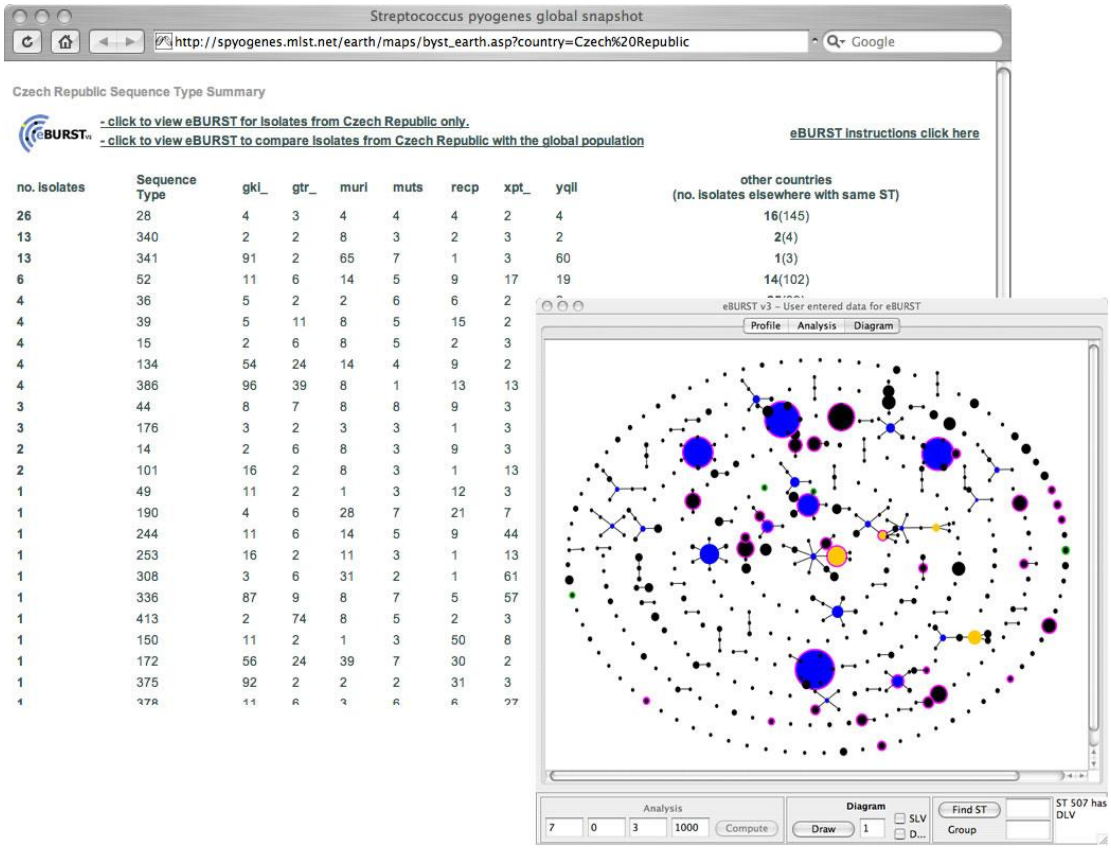


Figure 5.11 *Streptococcus pyogenes* Sequence Types for isolates from the Czech Republic, launched from within Google Earth and subsequent eBURST snapshot.

In comparison to the BETA version of mlst.net v2 (*S.pneumoniae*), the functionality within MLST-Maps is somewhat restricted. However, when the BETA version becomes more widely implemented, the extra functionality will be introduced to other databases and MLST-Maps will continue to act as a central point for viewing multiple MLST databases (including those not hosted at mlst.net) geographically.

5.2.2 Staphylococcal Reference Laboratory Maps: SRL-Maps

Available: http://www.spatialepidemiology.net/SRL-Maps

The SRL-Maps web application and its utilisation has been described in the following manuscript, which is currently in submission:

Grundmann, H., Aanensen, D.M., van den Wijngaard, C.C., Spratt, B.G., Harmsen, D and Friedrich, A.S. for the European Staphylococcal Reference Laboratory Working Group. (2009) Geographic distribution of *Staphylococcus aureus* causing invasive infections in Europe.

SRL-Maps will be released following publication, however currently, access requires the following username and password: Username: Europe ; Password: catechol

5.2.2.1 Introduction

The Staphylococcal reference laboratories Maps application (SRL-Maps) is a collaboration with Hajo Grundmann, leader of the European Antimicrobial Resistance Surveillance System (EARSS) [W65] and Staphylococcal reference laboratory (SRL) representatives from the EARSS network. EARSS collates antimicrobial resistance data for a number of pathogens and a number of antimicrobial compounds from over 28 European countries through a network of national reference laboratories. This network of laboratories was utilised for a focussed project aimed at typing and characterising the European diversity and distribution of *Staphylococcus aureus* causing invasive disease. Before detailing my input, I shall introduce the project in more detail.

S. aureus is one of the most important human pathogens and its methicillin-resistant variant (MRSA) is a major cause of hospital and community-acquired infection. It is estimated that between 25-35% of healthy human individuals carry *S.aureus* either on their skin or mucous membranes (Wertheim et al., 2005). Spread from the skin to the bloodstream can occur through disruptions to normal skin tissue (cuts, abrasions etc.), during surgical procedures, or in patients with a depressed immune system and associated infections. The range and severity of infection is largely due to the virulence of the strain and the status of the individual's immune system (Croze et al.,

2009; Tacconelli et al., 2006). *S. aureus* occasionally acquires virulence determinants and resistance to antimicrobials through horizontal gene transfer. Clones of *S. aureus* are relatively stable, with diversification at house-keeping genes generally occurring through point mutation, which allows these mobile genetic elements to persist within a population. Molecular typing can be utilised to discern the clonal lineages within a population and the importance of strain identification for understanding the evolution and population biology of an organism has been discussed previously in chapter 3. Methicillin-susceptible *S. aureus* (MSSA) provides the genetic reservoir through which MRSA has emerged due to the acquisiton of genetic elements containing methicillin resistance genes (the SCCmec element which includes the *mecA* gene), as demonstrated using MLST data and the BURST algorithm (Enright et al., 2002).

Structured geographical surveys of strain distribution (both MSSA and MRSA) on a continent-wide scale are scarce, generally due to lack of survey strategies and agreement on a consistent, standardised and appropriate typing approach.

The study undertaken provided:

i) A representative and contemporary snapshot of the genotypes of *S. aureus* (based on *spa* sequence typing) that cause invasive infection in the European region.
ii) An insight into the geographical distribution of different genotypes among MSSA and MRSA on a continent-wide scale.
iii) A web-based mapping tool that provides interactive access and an intuitive illustration of the results generated by such a large scale typing initiative.

My part in the project was the design and development of SRL-Maps (iii)). Prior to further details, I outline the methods chosen by SRLs and utilised for i) and ii).

Spa typing is a single locus typing method that involves DNA sequencing of the repeat region of the *Staphylococcus* protein A gene (*spa*) (Frenay et al., 1996). The region consists of a frequently varying number of nucleotide tandem repeat sequences. Each unique repeat sequence is identified as an integer and the number and composition of these repeat sequences is used to define the *spa* type (assigned as an integer in the format ‘t082’ where 082 is the type). In a conceptually similar

fashion to the assignment of sequence types using MLST, *spa* type designations are assigned centrally via a web server curated by SeqNet.org at University Hospital Münster, Germany [W66]. The commercial software 'StaphType^{TM}' from Ridom Bioinformatics, allows DNA chromatogram editing and automatic assignment of *spa* types by synchronising with the central *spa* server [W67].

During workshops organised for technical personnel from 28 European SRL's, participants received training in *spa* typing and analysis using StaphType. Proficiency testing was carried out by supplying each SRL with five isolates of well-characterised *S. aureus* and five sequence chromatograms of known *spa* types. Only those laboratories performing to fully quantifiable quality control standards were invited to participate in the survey. All participating SRL's were then asked to identify about 20 regional laboratories that serve hospitals that are geographically and demographically representative for their country, and secure their participation.

For six months (from September 2006 until February 2007) the regional laboratories collected the first five successive MSSA and the first five successive MRSA isolates from individual patients with invasive disease. All isolates were sent to the central SRL and characterised by *spa* typing. MRSA was confirmed by *mecA* PCR and subsequently all SRLs submitted data to a central SQL database located at the National Institute for Public Health and the Environment of the Netherlands (RIVM) for further analysis.

Along with the *spa* type and assignment as MRSA or MSSA, additional information submitted included; the EARSS laboratory identifier, the sample number, the date of isolation, origin of clinical specimen, demographic detail (such as age and gender), epidemiological context (hospital or community acquired), antibiotic resistance (oxacillin and cefoxitin) and all-cause mortality 14 days after isolation. SRLs also provided the postal address of laboratories for mapping. Some countries (Austria, Belgium, Czech Republic and Poland) submitted regional data rather than postcode if laboratories were to be aggregated for mapping. Before describing the web application, and to place the data in context, the broad statistics of data collected will be summarised (statistics were produced by Hajo Grundmann).

5.2.2.2 Data summary

Twenty-six SRLs (representing 26 countries) contributed data. During the six-month sampling period, 357 laboratories serving 450 hospitals collected 2890 successive MSSA and MRSA isolates from patients with invasive *S. aureus* infection (90.4% from blood cultures, 0.6% from cerebrospinal fluid and 9.3% from other normally sterile sites). Some data were missing for certain isolates; gender (28 isolates), age (54), date (74), epidemiological context (568), all-cause mortality (1052) and *spa* type (40). Many laboratories were not able to collect five successive MRSA isolates within the sampling period due to low MRSA incidence in the hospitals they serve. The combined collection consisted of 1923 MRSA (66.5%) and 967 MSSA (33.5%). Patients with MRSA infections were older with a median age of 69 years compared to 63 with MSSA and patients with MRSA had a higher all-cause mortality (20.8%) 14 days after the first isolation of *S. aureus* than MSSA patients (13.2%). More males (61.7%) than females had invasive *S. aureus* infections, however the proportion of MRSA compared to MSSA did not differ significantly between sexes.

A total of 660 *spa* types were reported. 565 and 155 were assigned to MSSA and MRSA respectively, of which, 505 were exclusively MSSA and 95 exclusively MRSA. MRSA was less diverse than MSSA. Only five *spa* types accounted for 50% of all MRSA isolates, whereas the same proportion of MSSA isolates comprised the 26 most frequent MSSA *spa* types.

The central database was made available for the development of the SRL-Maps application. Postal address, or region name, was converted into latitude and longitude coordinates using the geocoding facility developed and described in section 5.2.4 (http://www.spatialepidemiology.net/user_maps) and all data were entered into MS SQL server.

During the annual EARSS plenary meeting held in Athens, Greece, in November 2008, I demonstrated two versions of SRL-Maps. Functionally identical, the first is password-restricted and allows each of the country representatives to log in and view only data submitted for their country, including all the associated data listed above. The second gives a European view, again consisting of all data, to demonstrate the benefit of a geographically broader display. Following the meeting it was agreed that

the country specific view of the data would stay password restricted and accessible only by country representatives. A European view, displaying restricted epidemiological data (to protect data privacy) would be made publicly available following an initial publication. As mentioned, the functionality available within both views is identical, however the password-restricted view allows only one countries data to be seen and the public version allows viewing the entire network but with limited epidemiological data. It is this second, public version, which will now be described.

5.2.2.3 SRL-Maps Web Application

The initial page details the introduction to the project and a navigation menu is located at the top right of the screen containg the following sections:

i) Introduction.
ii) Isolate Database.
iii) *spa* type Google Map.
iv) *spa* type Google Earth.
v) Acknowledgements.

The introduction and acknowledgements are self-explanatory and shall not be discussed, but provide instructions to SRLs, and following publication, other users, on how to use the site and list participants and their contributions to the study.

5.2.2.4 Isolate database

The isolate database can be searched using a standard form based query method. As mentioned, for the public site, minimal epidemiological data are available. Upon entering a particular query (for example '*spa* type = t067' and clicking 'Submit' the results of the query are returned, listing (for each isolate of *spa* type t067) the isolate ID, country and city of origin, the MSSA or MRSA status of the isolate and the *spa* type and repeat pattern. Just above the table of results is a link allowing the results of the query to be viewed on a map. Any regional reference laboratory submitting isolate details with the variable searched for is represented by a marker, the colour of which

represents the % of the total results found in that laboratory (the key to the right can be used to turn on or off % boundaries).

5.2.2.5 *Spa* type Google Map

Clicking the '*spa* type Google Map' menu option displays the European regional laboratory view (Figure 5.12). Each regional laboratory is represented by a marker based on the latitude/longitude of the postal address provided. Those countries requiring laboratory agglomeration to regional level (Austria, Belgium, Czech Republic and Poland) are represented by a marker placed at the latitude/longitude at the geographical centre of each region.

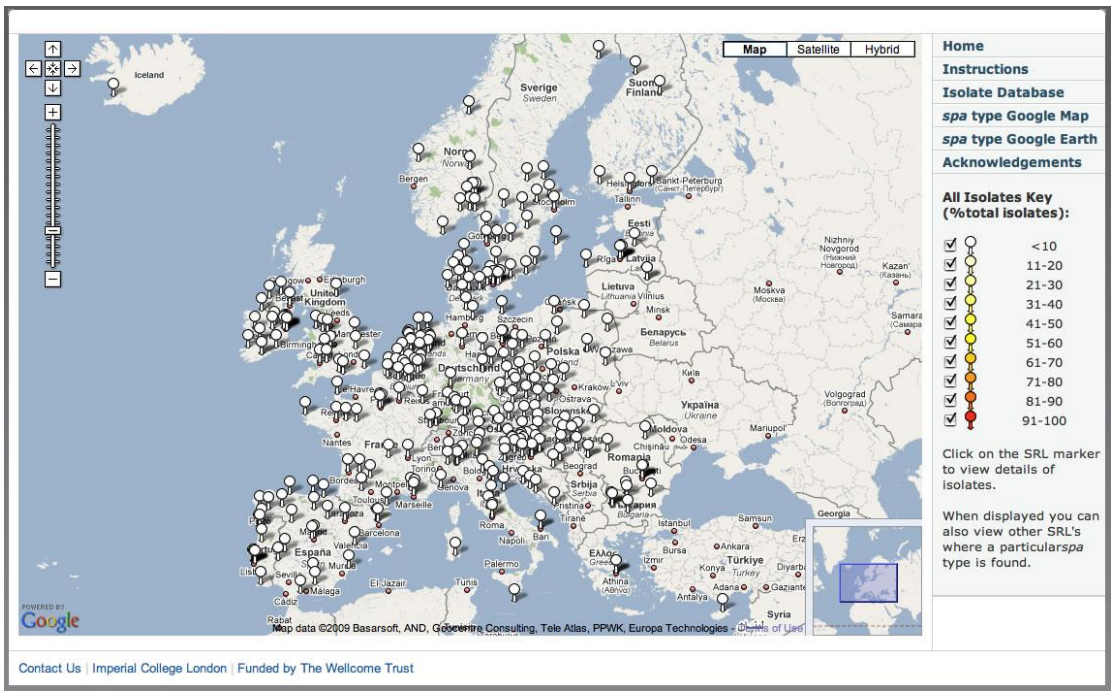
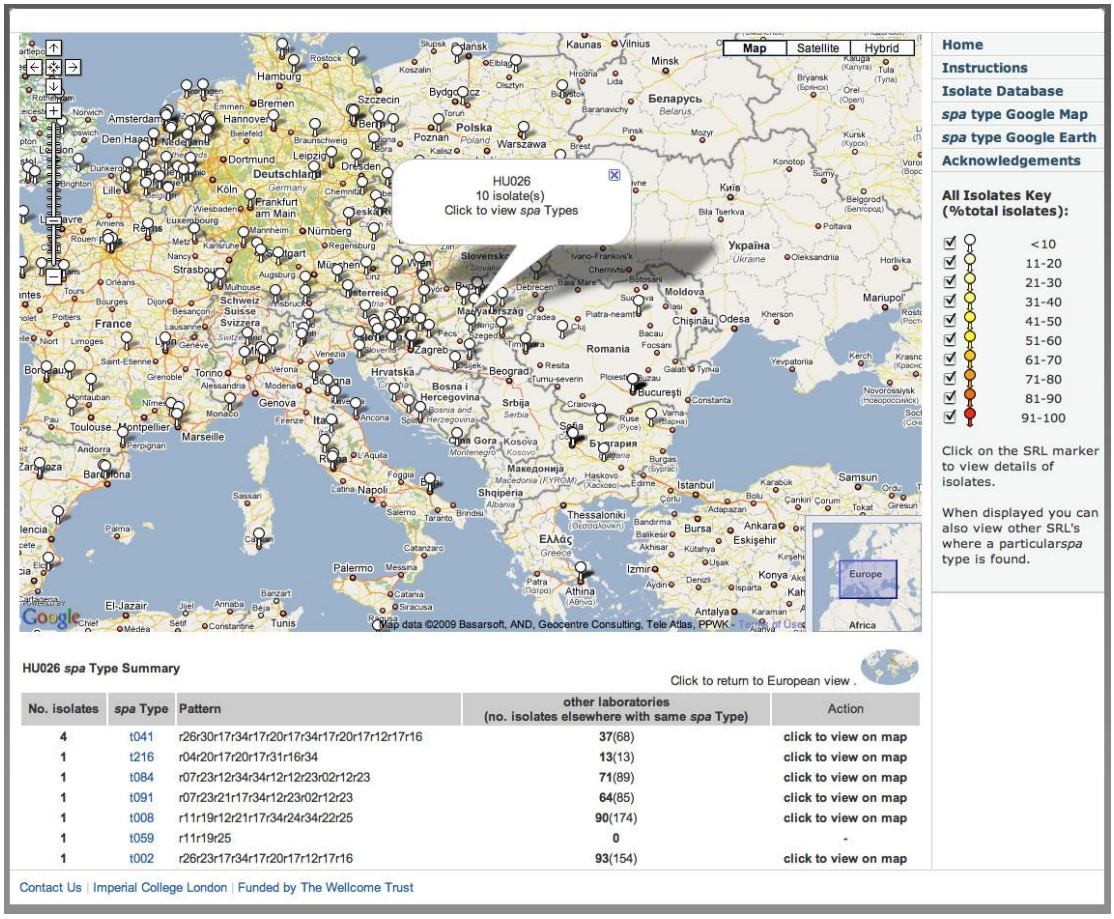


Figure 5.12 SRL-Maps: European distribution of Staphylococcal reference laboratories.

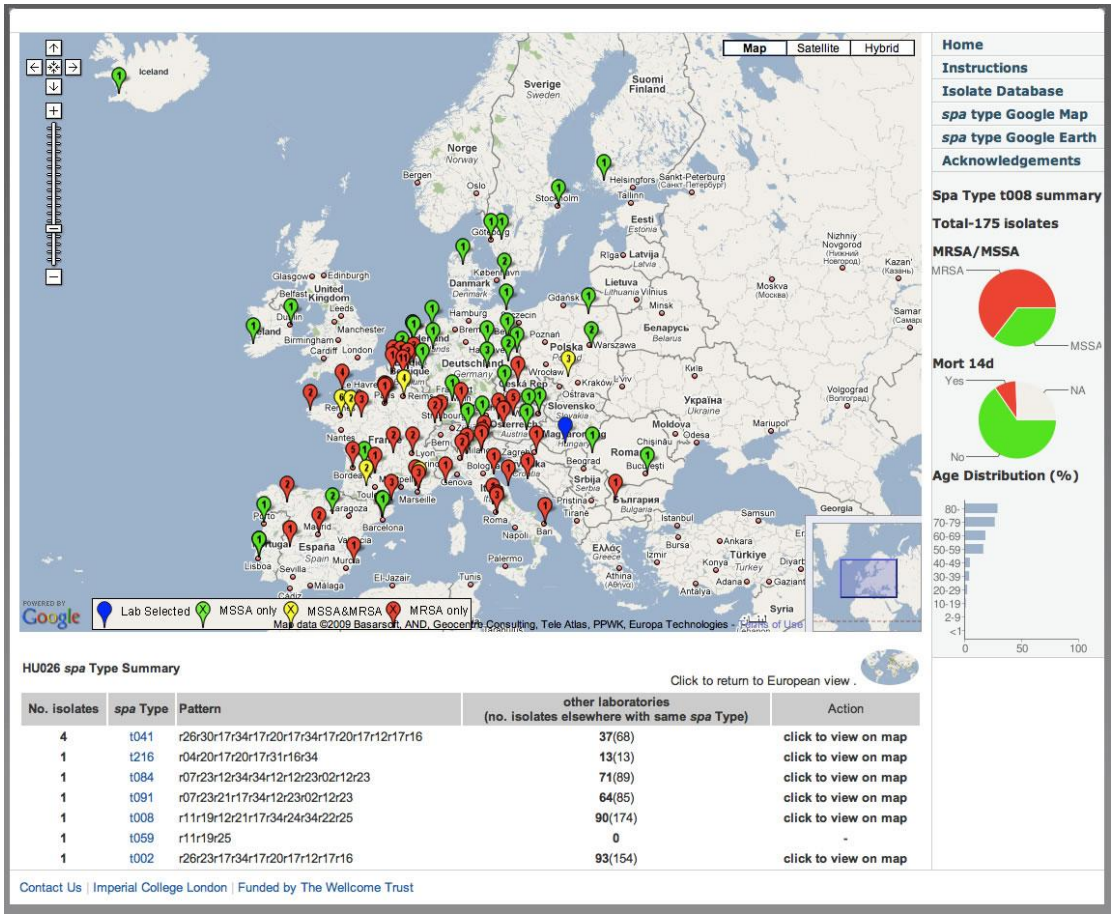
Clicking on a marker opens an Info Window indicating the number of isolates whose details have been submitted from that laboratory and a link to view their *spa* types. Clicking the link zooms the map into the laboratory location and lists the *spa* types, along with their repeat pattern and an indication of the number of other laboratories (and number of isolates) with that *spa* type. A link is also provided to view the European distribution of each *spa* type (Figure 5.13).



No. isolates	*spa* Type	Pattern	other laboratories (no. isolates elsewhere with same *spa* Type)	Action
4	t041	r26r30r17r34r17r20r17r34r17r20r17r12r17r16	**37**(68)	**click to view on map**
1	t216	r04r20r17r20r17r31r16r34	**13**(13)	**click to view on map**
1	t084	r07r23r12r34r34r12r12r23r02r12r23	**71**(89)	**click to view on map**
1	t091	r07r23r21r17r34r12r23r02r12r23	**64**(85)	**click to view on map**
1	t008	r11r19r12r21r17r34r24r34r22r25	**90**(174)	**click to view on map**
1	t059	r11r19r25	**0**	-
1	t002	r26r23r17r34r17r20r17r12r17r16	**93**(154)	**click to view on map**

Figure 5.13 Detailed view of laboratory HU026, Szekszárd, Hungary.

Figure 5.13 details the *spa* types isolated from laboratory HU026, located in the city of Szekszárd, Hungary. Ten isolates were *spa* typed and the majority of types were also found elsewhere. Clicking on the link to view isolates of *spa* type t008 (which is found in 90 other laboratories (174 isolates)) changes the view to that seen in Figure 5.14.



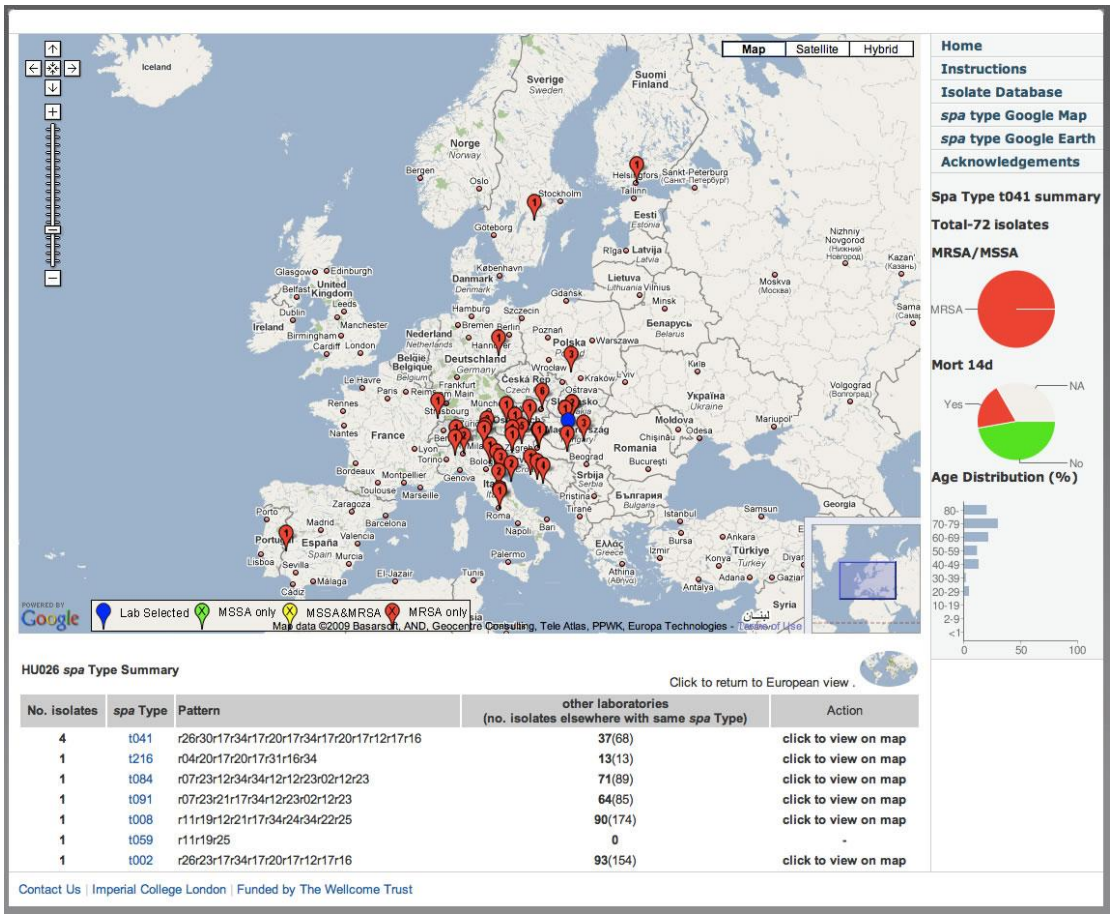
No. isolates	*spa* Type	Pattern	other laboratories (no. isolates elsewhere with same *spa* Type)	Action
4	t041	r26r30r17r34r17r20r17r34r17r20r17r12r17r16	**37**(68)	click to view on map
1	t216	r04r20r17r20r17r31r16r34	**13**(13)	click to view on map
1	t084	r07r23r12r34r34r12r12r23r02r12r23	**71**(89)	click to view on map
1	t091	r07r23r21r17r34r12r23r02r12r23	**64**(85)	click to view on map
1	t008	r11r19r12r21r17r34r24r34r22r25	**90**(174)	click to view on map
1	t059	r11r19r25	**0**	-
1	t002	r26r23r17r34r17r20r17r12r17r16	**93**(154)	click to view on map

Figure 5.14 European distribution of *spa* type t008 focussed on SRL HU026.

Laboratory HU026 is represented by a blue marker and every other laboratory submitting a t008 isolate is represented by a marker with the number of *spa* type t008 isolates indicated within. If all isolates from a laboratory are MSSA the marker is coloured green. If all isolates are MRSA the marker is red and if there is a mix of MRSA and MSSA isolates the marker is yellow; a key to the marker colour scheme is placed at the bottom of the map. To the right of the map, below the navigation menu, a graphical summary of all isolates is shown (including the total number isolated). The first pie chart represents the proportion of isolates that are MRSA and MSSA, the second indicates whether mortality was seen after 14 days ('NA' indicates data not available) and the bar chart shows the patient age distribution. It can be seen in Figere 5.14 that t008 is predominantly MRSA with low all cause mortality, generally isolated from older patients.

A further example can be viewed by clicking the link to view the distribution of *spa* type t041 (Figure 5.15). 72 isolates of *spa* type t041 were isolated (all MRSA), the majority of which form a cluster close to lab HU026, and just under half of the

patients were still alive after 14 days (the full extent is not possible to discern due to missing data).



No. isolates	*spa* Type	Pattern	other laboratories (no. isolates elsewhere with same *spa* Type)	Action
4	t041	r26r30r17r34r17r20r17r34r17r20r17r12r17r16	37(68)	click to view on map
1	t216	r04r20r17r20r17r31r16r34	13(13)	click to view on map
1	t084	r07r23r12r34r34r12r12r23r02r12r23	71(89)	click to view on map
1	t091	r07r23r21r17r34r12r23r02r12r23	64(85)	click to view on map
1	t008	r11r19r12r21r17r34r24r34r22r25	90(174)	click to view on map
1	t059	r11r19r25	0	-
1	t002	r26r23r17r34r17r20r17r12r17r16	93(154)	click to view on map

Figure 5.15 European distribution of *spa* type t041 focussed on SRL HU026.

5.2.2.6 *Spa* type Google Earth

When opening the KMZ file provided, all laboratory locations are represented as icons (Figure 5.16). Clicking on one of the icons opens an Info Window detailing the laboratory ID, the number of isolates with details from that laboratory and a link to 'Click to view *spa* types' Clicking on the link opens a web browser and a similar *spa* type summary to that viewed through the Google Map interface is displayed. As with the Google Earth view of MLST-Maps, the lack of two-way connectivity when using Google Earth precludes the subsequent viewing of the distribution of a particular *spa* type.

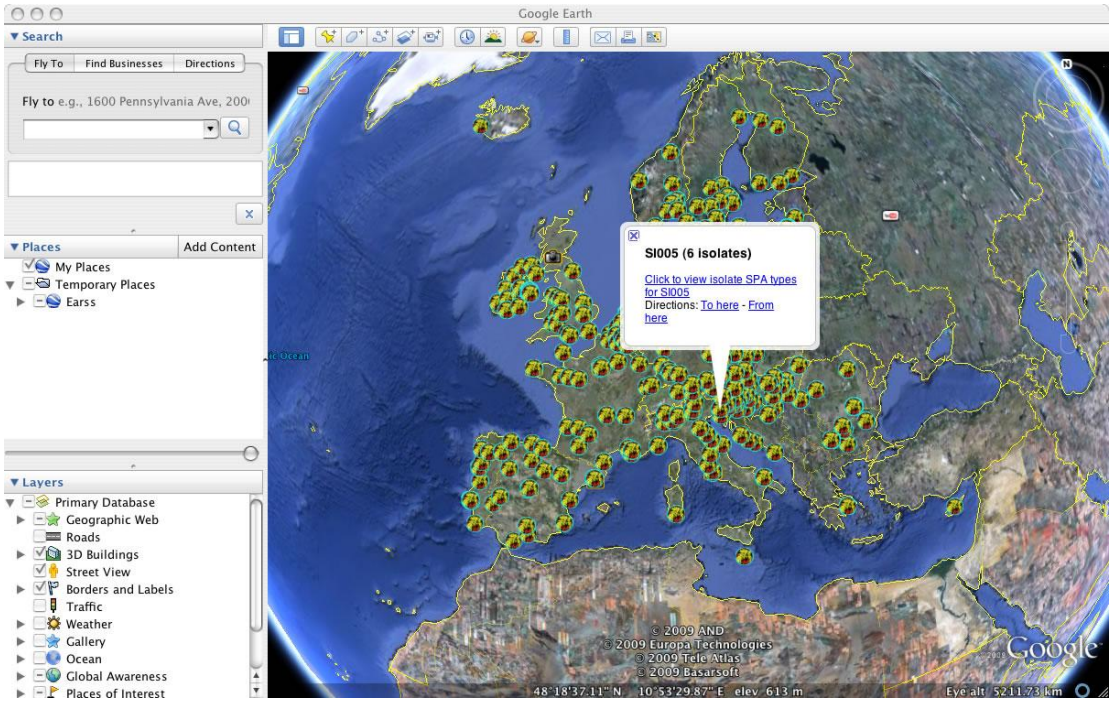


Figure 5.16 European distribution of Staphylococcal reference laboratories viewed using Google Earth.

The SRL-Maps application has provided participating laboratories across Europe with a method of viewing and interrogating the data generated for the structured survey undertaken. Feedback received has been wholly positive and has led to the decision (by all SRL's) to open the facility for public exploration following manuscript publication. Furthermore, a representative sample of isolates have been made available for further genomic sequencing by collaborators at the Sanger Institute and it is hoped that such molecular data can be incorporated within SRL-Maps to allow further investigation by the network of SRLs.

The structured survey represents a single temporal snapshot of the genetic structure of *S.aureus* isolates collected across Europe. However, it has been agreed by participating SRLs that a similar sampling strategy will be undertaken again. This presents exciting opportunities to investigate changes in the distribution of genotypes (using whichever method is chosen if not *spa*) temporally as well as spatially.

5.2.3 *Batrachochytrium dendrobatidis* Maps: *Bd*-Maps

Available: http://www.spatialepidemiology.net/Bd-Maps

The *Batrachochytrium dendrobatidis* Maps (*Bd*-Maps) application is a collaboration with Deanna H Olson of the USDA Forest Service, Corvallis, USA who leads the global *Bd* mapping initiative and Matthew Fisher, Imperial College London, a researcher in fungal disease. The web application, and its use, is described as part of the following publication, which is currently in submission.

Olson, D.H*., Aanensen, D.M*., Ronnenburg, K.L., Walker, S.F., Bielby, J., Garner, T.W.J., Weaver, G., Spratt, B.G, the *Bd* Mapping Group, and Fisher, M.C. (2009) Mapping the Global Emergence of Amphibian Chytridiomycosis.
*Joint first authors

5.2.3.1 Introduction

Batrachochytrium dendrobatidis (*Bd)* is a fungal pathogen of amphibians, which results in the disease chytridiomycosis and is implicated as a primary causative agent in recent population declines (Daszak et al., 2003). First identified by Longcore et al, (1999), death of amphibians due to chytridiomycosis has been reported in North and South America, Europe, Australasia and Africa (Wake and Vredenburg, 2008).

Infection with *Bd* occurs following exposure of an amphibian to water containing free-living *Bd* zoospores (a motile asexual spore) released from mature zoosporangia (the fungal structure producing zoospores) in the keratinized tissues of infected hosts. Infection is limited to these tissues, which are present in the outer layer of skin of adult amphibians but only in the mouthparts of early-stage larvae. Infection causes thickening of the keratinized layer of cells, excessive lethargy, irregular cell loss, hyperkeratosis and excessive skin sloughing (weakening and shedding) which can lead to death, usually at about 21 days post infection.

Bd appears to have the widest host-range of any known eukaryotic pathogen, infecting many species that are central to broad food webs (Wake and Vredenburg, 2008) and its spread is thought to occur by a number of routes; animal trafficking for commercial trade (food, scientific, bait and pet markets; Rowley et al., 2007),

fisheries management activities (Green and Dodd, 2007) and the re-introduction of infected individuals following captive-breeding programs (Walker et al., 2007). In response to the increasing role played by *Bd* in amphibian declines, the World Organisation of Animal Health (OIE) has recently listed *Bd* as a notifiable disease [W68].

To assess the impact of disease (and other factors) on wildlife (and plants) organisations such as the International Union for Conservation of Nature (IUCN) [W69], through networks of government and non-governmental member organisations, monitor the global presence and range of species. The 'IUCN Red List of Threatened Species™' [W70] provides taxonomic, conservation status and distribution information on plants and animals (that have been globally evaluated using various criteria) to determine the relative risk of extinction. Any threatened species on the Red List is categorised as Critically Endangered, Endangered or Vulnerable. While the IUCN provides valuable information on the status of a particular animal or plant species, it does not undertake pathogen surveillance, and as such can only be used as a gross measure of the impact a number of factors may be having on a particular species.

Pathogen surveillance systems for animal species of commercial interest are, in many cases, well established. However, systems aimed at the surveillance of animal pathogens that are not zoonotic are rare and focus on a limited number of species and pathogens (Kuiken et al., 2005).

Bd is a non-zoonotic emerging pathogen that infects multiple host species. To address the need for data regarding the distribution of *Bd*, in January 2007 the Partners in Amphibian and Reptile Conservation (PARC) initiated the global *Bd* mapping project [W71]. The project aimed to initiate collation of *Bd* detections and sampling locations and to produce and publish regular (annual) updates as to the global picture of incidence. Data were collated from literature searches, websites and personal communications with scientists. Regional and country data coordinators were identified and helped by collating data for their area (and emailing Excel spreadsheets to PARC). The first collated data was published on a website in January 2008 as a series of five static maps [W72] showing 1) *Bd* Surveys; countries where surveys have occurred or are ongoing. 2) *Bd* detections; map showing locations. 3) *Bd* not

detected. 4) Zoomed in North America; showing *Bd* detected and not detected, and 5) *Bd* detections with mortality; whether animals tested had died.

The first international conference regarding *Bd* and amphibian declines was held in Phoenix, Arizona, USA in Nov 2007. At this conference I demonstrated a basic version of the current *Bd*-Maps application (using location data mostly extracted from literature searches by Matthew Fisher and then geocoded; detailed in section 5.2.4). It was subsequently agreed with Deanna Olson, who initiated and runs the global mapping project at PARC that, following the initial static map release, data would be migrated to *Bd*-Maps. The following description details the current version of *Bd*-Maps that now acts as the central global mapping and surveillance tool for *Bd*.

5.2.3.2 Data summary

Data requested for the initial PARC mapping project were as follows [W71] :

Region – Country, and region of country.
Location – Latitude/Longitude if known.
Species – Amphibian species tested; life stage and sample size.
Bd detection – Positive *Bd* test result.
Method – Which used to test for *Bd* (histology or PCR).
Dead Animal – Positive *Bd* from an animal that was dead, or clearly ill and noted.
Live Animal – Animal not known to die, and with no signs of being ill.
Other – *Bd* detected in another way (e.g. from water sampling).
Bd not detected – Negative *Bd* test result.
Source – Name and email address of sender.
Date – Date of data compilation.
Notes – Any comments about the data or study.

All data (from email submission, literature and web search) were collated by Dede Olson and Kathryn Ronnenburg. Those locations for which precise latitude and longitudes were not known were geocoded using Google Earth and the centroid latitude/longitude for the area was used. For sample sites where multiple samples were recorded, data were agglomerated by species so that a record for a single site may contain multiple samples but only one species. The IUCN status (see above) for

each species was recorded in a separate column. All data were stored in an Excel spreadsheet and sent to Imperial College so that I could use as a framework for further developing *Bd*-Maps

Data included a total of 1701 sample points (i.e. unique latitude/longitude combinations) representing 6667 samples from 72 countries and 799 different amphibian species. Three columns were added to the database as follows – 'Number positive', Number Negative' and 'Number Sampled'. Many of the records sent had missing data fields and little quality control was undertaken (as stated on the PARC website [W71]). These three extra columns were used to inflict stringency on the data and shall be referred to again when discussing the web application in section 5.2.3.4.

5.2.3.2 Handling large datasets using Google Maps

In a conceptually similar fashion to MLST-Maps and SRL-Maps, within *Bd*-Maps each sample point acts as a reference for accessing the underlying data. However, the *Bd* data are more geographically diverse, contain many more records and furthermore, this number is likely to increase greatly with time. The current number of points in MLST-Maps is limited by number of countries and in SRL-Maps, the number of laboratories. However, a primary aim of developing such mapping facilities is to increase user awareness and encourage the deposition of more precise geographical data. Therefore, the applications should be designed to be scalable and capable of handling large datasets. The *Bd* dataset, although lacking in molecular data, provided the opportunity to design and implement such a method.

Google Maps is very proficient at displaying many points, however, the application speed begins to decrease (due mainly to the JavaScript engine of the web browser) substantially when plotting more than about 500 points within one window. Furthermore, similar performance issues occur when loading large KML files into Google Earth. To address this issue within *Bd*-Maps I implemented a zoom-sensitive method for displaying data. At the global view, a thematic map displays a summarised version of the data. Overlays are positioned on top of the representation of a country from where data has been submitted and this allows colour to be used as an indication of the number of positive or negative samples recorded per country. By far the greatest amount of data in the *Bd* dataset comes from the USA (due to heightened

awareness of the pathogen and subsequent sampling efforts), and the thematic map for the USA is further delimited by State boundaries. Zooming in one level on the map places a single marker at the centroid of each country (or state in the USA), allowing all of one country's (or state's) data to be viewed by clicking. Zooming further displays all individual sampling points, coloured based on criteria described below. In the USA a further zoom is required to view site data. In a similar fashion to Google Maps itself, which only using a minimal number of tiles to represent the section of the earth that is being viewed, this three-level zoom method only displays points on the maps based on zoom and view area.

To create the thematic map, a world borders dataset was used (derived from the ThematicMapping website [W73]). The borders dataset is a single KML file which includes data representing each country as a polygon, consisting of a set of latitude and longitudes which trace the countries outline; the polygon 'fill colour' is specified by a hexadecimal colour code. Utilising this static borders dataset, the 'outline' latitude and longitudes for each country were parsed into a database resulting in a single record per country, containing the country name, ISO country code and the border represented as a set of points. For the United States, a similar dataset was derived from a US State border KML file from the Cocoa Research Centre [W74].

Storing the country/state borders in such a fashion allows one to use server-side scripting to dynamically generate a global thematic map. Country/State polygon fill colours can be specified using the results of a database lookup, for example, to colour based on a scale representing number of records. Specification 2 of the Google Maps API [W55] allows the direct overlay of KML onto a Google Map and thus, this method can be used to create a dynamic thematic world view of any dataset.

Marker display at differing geographical levels can be handled by detecting, using JavaScript, the users zoom level and either loading or removing sets of markers, using AJAX. For example, imagine three zoom levels (1-3) with 1 being the lowest level and representing the entire world view. At level 1 the thematic world view is displayed. At zoom level 2, markers are positioned on each country/state centroid. When zooming to level 3, the country/state centroid markers are removed, and individual sample site markers are added, based on the section of the world that is being viewed.

In this fashion we can potentially cope with very large datasets by providing a global thematic overview with further detail displayed when requested during map navigation. Delimiting the USA into states was undertaken to cope with the large amount of data submitted from the USA however in theory, any country can be broken down into individual units using border datasets, and these borders could be governmental or any other system used to delimit an area into smaller regions.

5.2.3.3 *Bd*-Maps Web Application

The homepage displays standard navigation menus to the right and also a list of current publications (as listed in PubMed) and recent BLOG posts from the *Bd*-Maps BLOG (Figure 5.17). The centre of the page displays a snapshot of the current database in map form. This utilises the Google Visualisation API [W75]. This API provides various data visualisations dynamically and simply requires two parameters, a data series to be used and the type of visualisation. One of the visualisations provides a static map view of data used. A simple count, per country, of positive *Bd* samples was calculated and used to display within the map visualisation. This provides an up-to date, and simple, snapshot of the entire database showing from where *Bd* positive records have been submitted and also an overview of relative incidence based on a simple colour scale. The 'Home' link on the navigation menu allows a user to return to this frontpage from anywhere within the site. The 'Information' link provides details of the projects origin, information regarding *Bd* and instructions for using the site.

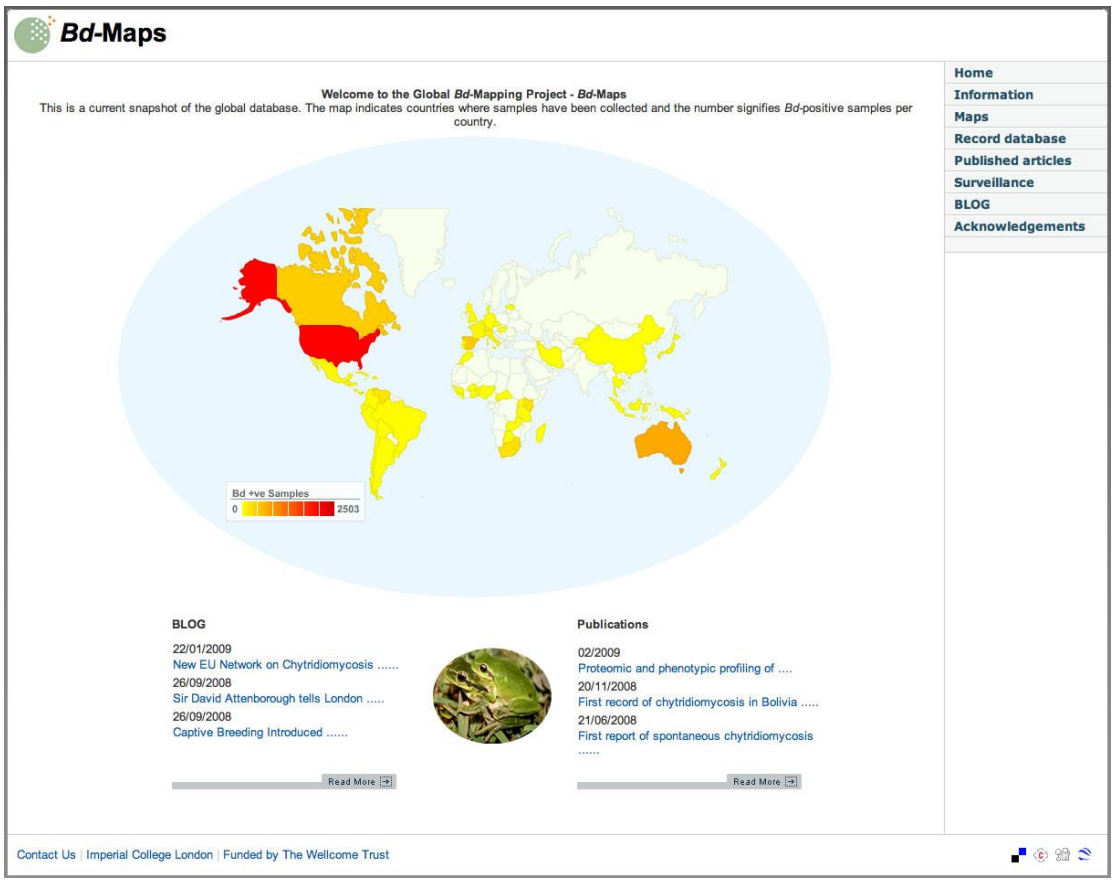


Figure 5.17 *Bd*-Maps: Homepage.

5.2.3.4 *Bd* Google Map

Clicking the 'Maps' option opens the main Google Map page as seen in Figure 5.18. Initially, the thematic map overlay as described above is displayed (and this overlay can be turned on or off using the link below the navigation menu). As mentioned, due to the lack of some data fields in the initial dataset received from PARC, three extra columns were added to the database; 'Number Positive', Number Negative' and 'Number Sampled'. These columns were completed (with help from Susan Walker) so that for each sample site it would be possible to identify whether it is to be classed as 'positive' or 'negative'. If such a system is to be potentially useful for any kind of basic surveillance purposes, we need to indicate where the disease is. However, as important, if not more so, is the identification and definition of disease-negative locations. This can then be used to highlight those sites that switch from negative to positive, or vice-versa.

The three additional database columns allow the definition of three types of site. This definition forms part of the application logic and is used when generating the maps:

1) If a value is entered into the ‘Number Positive’ column a site is designated ‘Positive’.
2) If a zero value is entered in the ‘Number Positive’ column, a positive integer in the ‘number negative’ column and also a value in the ‘Number sampled’ column, this site is designated as ‘Negative.
3) If a zero value is entered in the ‘Number Negative’ column and either the ‘Number Positive’ and/or the ‘Number Sampled’ column contains no value (missing data etc.) this site is designated as ‘Negative (Sample size unknown)’.

Any site with a value greater than zero in the Number Positive column is a positive site, but for a site with a zero positive value and a ‘Number Negative’ value greater than zero, where sample size is unknown, it cannot be said for certain that the site is negative, as it could just be that laboratory diagnosis has not been completed. The lack of data is partially due to the fact that many records were retrieved from literature searches and personal communication (collated at PARC) however, using these site definitions allows future sample reporting to be logically organised.

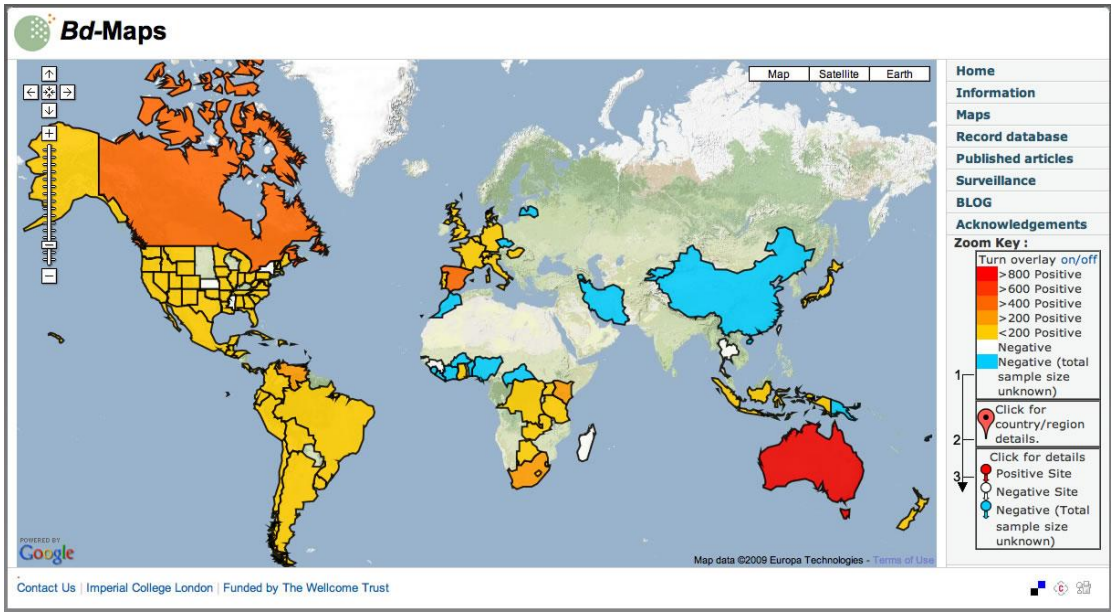


Figure 5.18 *Bd*-Maps: Global thematic overview. The outline of each country where positive samples have been recorded is shown and does not indicate equeal distribution of sample sites within a country.

This site definition is applied when indicating status for a region, or country and the thematic global map view reflects this. Any positive area (country/region) is coloured based on the actual number of positive samples recorded using a yellow-red scale indicated to the right of the map (figure 5.18). Any truly negative region/country (where sample sizes are known) is coloured white and any Negative region/country which adheres to criteria 3 above) is coloured blue. As can be seen from the map all samples tested in Madagascar are negative and sample sizes known, however many countries in West Africa have only reported negative samples, but exact sample sizes are unknown. A key to this colouring is indicated below the Navigation menu.

Zooming in one level calls a JavaScript function that detects the zoom level and creates markers for each country where one or more record is found in the database (Figure 5.19).

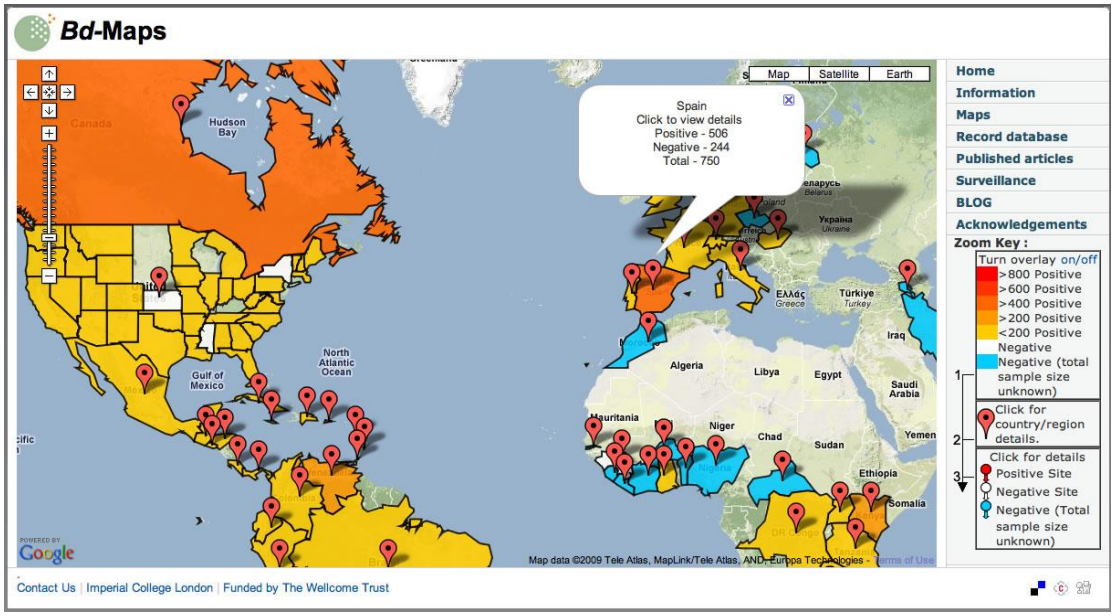


Figure 5.19 Country centroid markers.

Each marker can be clicked and the positive, negative and sample size values are summarised by country, and displayed in the resultant Info Window. A link to view details is also provided and produces output below the map (as detailed below).
As mentioned, within the USA the thematic map is split into States to cope with the large number of sample sites. Zooming one level further on the map places markers for every sample site recorded for the area of the map being viewed, except in the USA, where each State now has a marker at its centroid (Figure 5.20). Each State is

represented by a regional marker and outside of the USA (e.g. Mexico) markers represent actual sample sites. Clicking on a State marker displays an Info Window (in Figure 5.20, Nevada) and if the included link is clicked, the map zooms in one level further, which calls another JavaScript function that will now display all sample site locations and also the details of all records for that State (Figure 5.21).

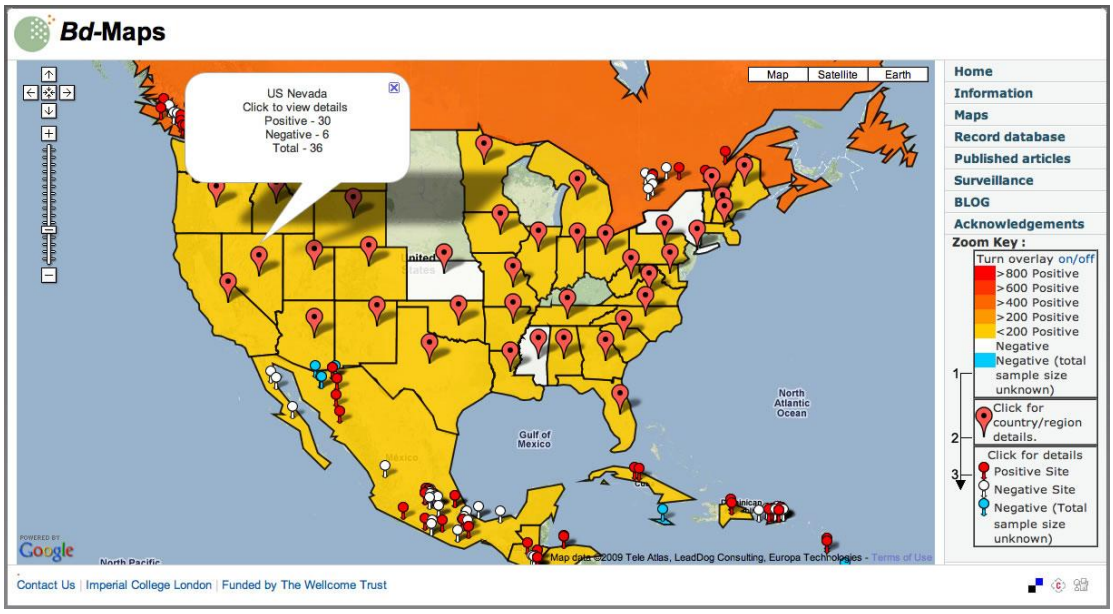
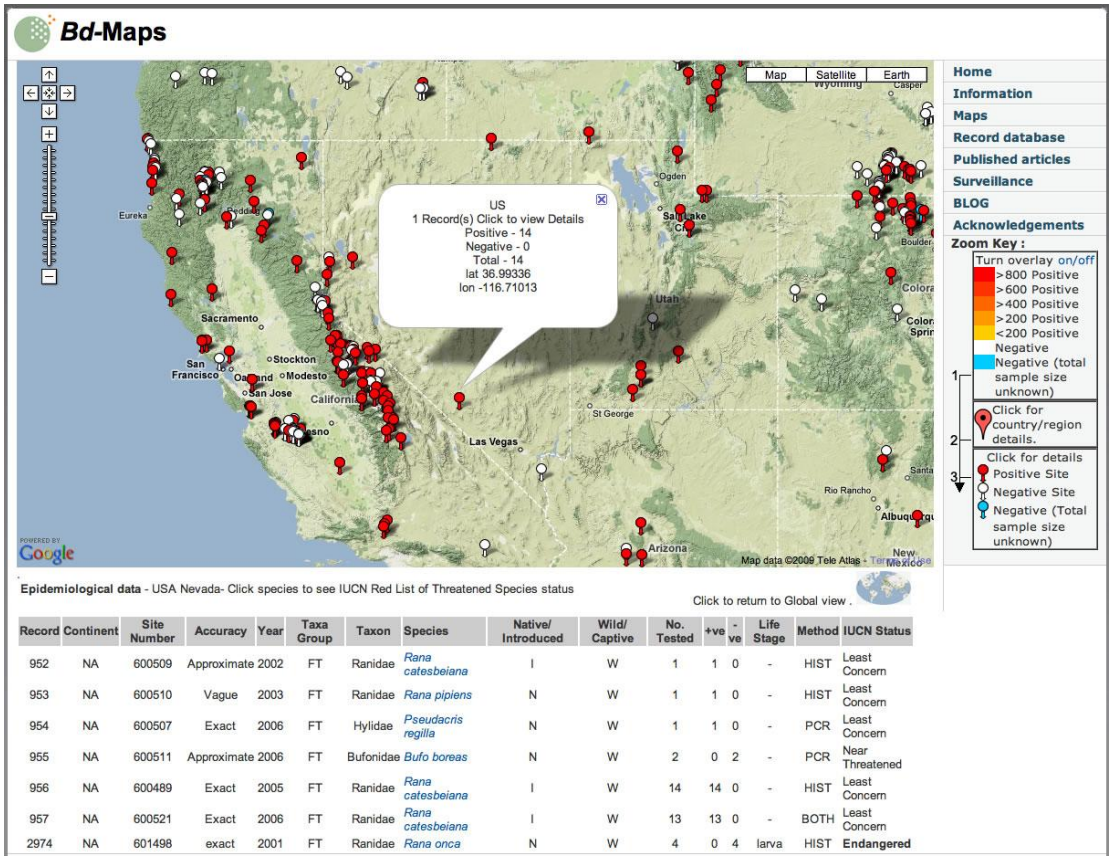


Figure 5.20 USA State (excepting Alaska and Hawaii) centroid markers.

Using the 'Turn overlay on/off' link below the Navigation Menu, the thematic overlay can be removed and allows the location of sites to be explored in further detail. Each site marker is colour coded in accordance with the positive/negative criteria defined above and the key indicates the colour meaning. Each sample site can be clicked to view the Info Window. Below the map, all records from Nevada are listed (Figure 5.21). Details include the site number, the accuracy of the location (whether the latitude /longitude is precise or represents the centroid of an area) and further details regarding the species of amphibian sampled.

The species name and IUCN status are also listed and in Figure 5.21 we can see that the final record listed (on the screenshot which has been truncated for size), the species *Rana onca*, has a Red List ranking of 'Endangered'. Each of the species names are hyperlinked, and clicking opens a new browser window with the relevant IUCN page for that species (Figure 5.22). Amphibian taxonomy undergoes frequent revision and family and species names can differ depending on which taxonomic

scheme is used. The IUCN keeps track of such changes and references the most up-to date accepted species name with reference to any synonyms that exist. *Rana onca* is now more commonly known as *Lithobates onca* and this is reflected in the IUCN page launched (Figure 5.22). It can be seen that *Rana onca* is listed as a synonym. The IUCN page contains detailed information regarding each species including range maps (i.e. known geographic distribution for healthy populations of a species).



Record	Continent	Site Number	Accuracy	Year	Taxa Group	Taxon	Species	Native/ Introduced	Wild/ Captive	No. Tested	+ve	-ve	Life Stage	Method	IUCN Status
952	NA	600509	Approximate	2002	FT	Ranidae	*Rana catesbeiana*	I	W	1	1	0	-	HIST	Least Concern
953	NA	600510	Vague	2003	FT	Ranidae	*Rana pipiens*	N	W	1	1	0	-	HIST	Least Concern
954	NA	600507	Exact	2006	FT	Hylidae	*Pseudacris regilla*	N	W	1	1	0	-	PCR	Least Concern
955	NA	600511	Approximate	2006	FT	Bufonidae	*Bufo boreas*	N	W	2	0	2	-	PCR	Near Threatened
956	NA	600489	Exact	2005	FT	Ranidae	*Rana catesbeiana*	I	W	14	14	0	-	HIST	Least Concern
957	NA	600521	Exact	2006	FT	Ranidae	*Rana catesbeiana*	I	W	13	13	0	-	BOTH	Least Concern
2974	NA	601498	exact	2001	FT	Ranidae	*Rana onca*	N	W	4	0	4	larva	HIST	**Endangered**

Figure 5.21 Zooming the map to view site sample sites in Nevada, USA.

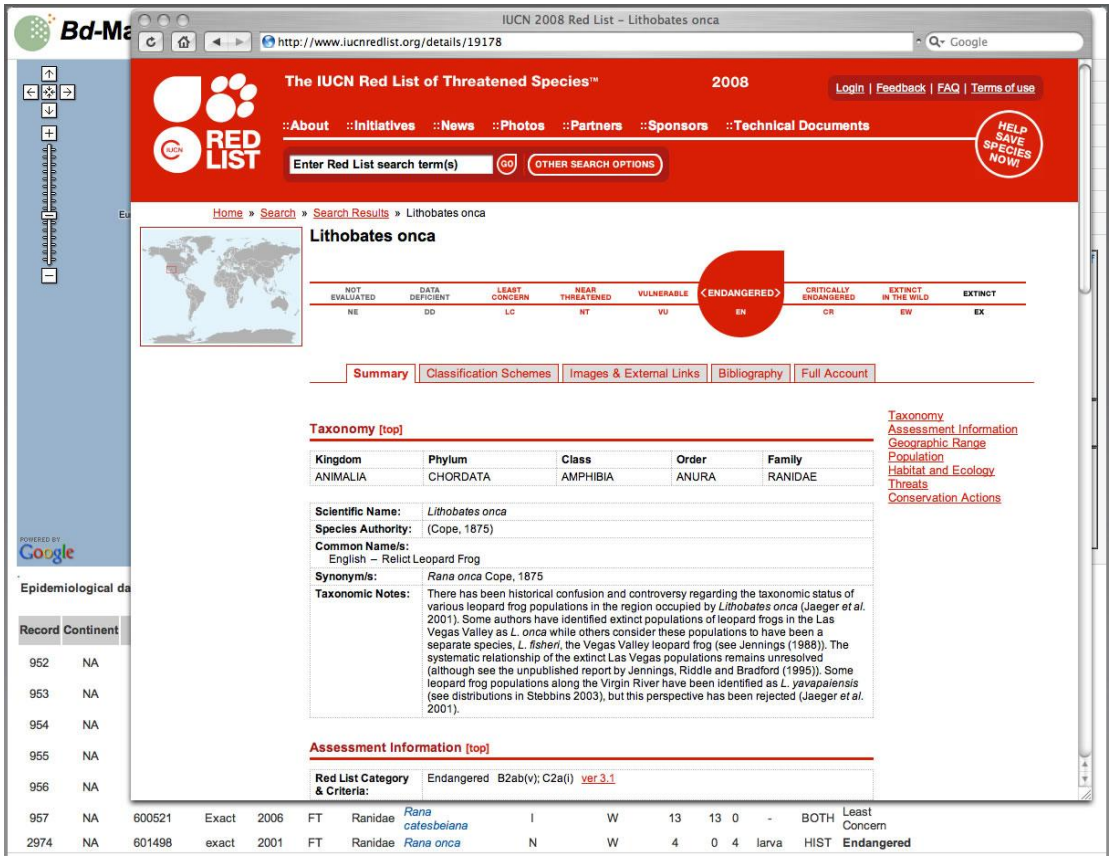


Figure 5.22 IUCN Red List page for *Rana onca* see text regarding discrepant species name.

5.2.3.5 Surveillance

In the future, as record numbers in the database increase, it will become important to have a method for viewing increasing trends within the data and, for *Bd*-maps to prove more useful than simply displaying points on a map, the surveillance pages were developed to offer a simple basic framework for further development (Figure 5.23).

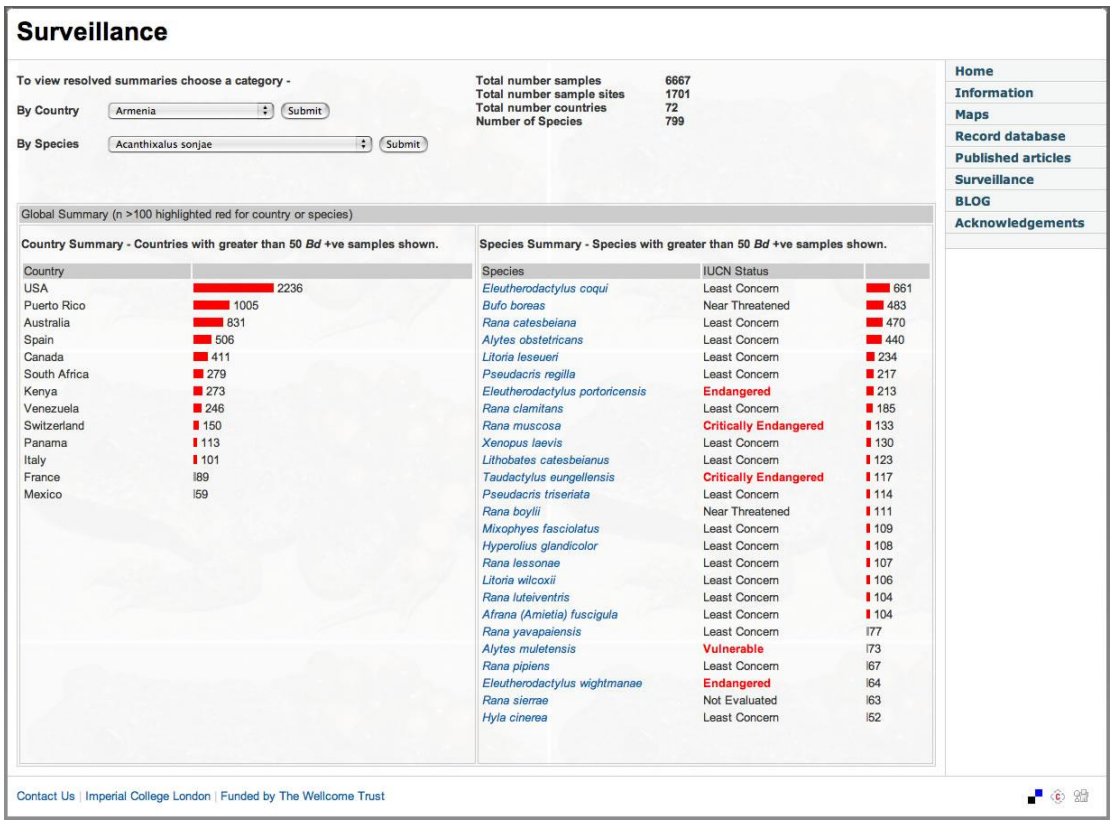
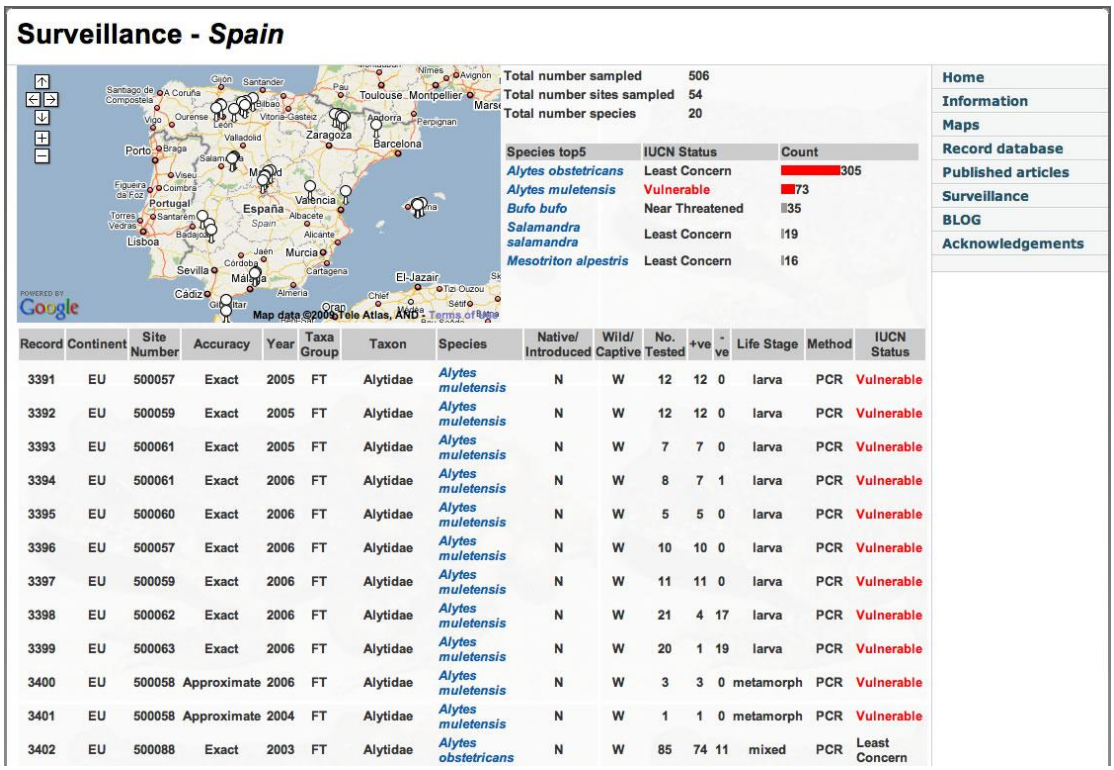


Figure 5.23 *Bd*-Maps: Surveillance summary.

The latter half of the page contains two tables. The first table lists, ranked by number of positive sample sites, the countries with the highest incidence of *Bd*. The second table lists, by amphibian, those species with the greatest number of positive records in the database along with the corresponding IUCN status (and a hyperlink to the IUCN species page as above). Both tables also contain a simple graphical summary which highlights in red those countries or species with greater than 100 positive records.

This represents a very basic summary, however it does allow the database to be monitored and warnings raised should a species or country exceed the cut-off level (100 positive samples). Above the tables is a form containing the option to view a summary of any country or species in the database. Choosing 'Spain' from the country list produces a page as in Figure 5.24



Record	Continent	Site Number	Accuracy	Year	Taxa Group	Taxon	Species	Native/ Introduced	Wild/ Captive	No. Tested	+ve	-ve	Life Stage	Method	IUCN Status
3391	EU	500057	Exact	2005	FT	Alytidae	*Alytes muletensis*	N	W	12	12	0	larva	PCR	Vulnerable
3392	EU	500059	Exact	2005	FT	Alytidae	*Alytes muletensis*	N	W	12	12	0	larva	PCR	Vulnerable
3393	EU	500061	Exact	2005	FT	Alytidae	*Alytes muletensis*	N	W	7	7	0	larva	PCR	Vulnerable
3394	EU	500061	Exact	2006	FT	Alytidae	*Alytes muletensis*	N	W	8	7	1	larva	PCR	Vulnerable
3395	EU	500060	Exact	2006	FT	Alytidae	*Alytes muletensis*	N	W	5	5	0	larva	PCR	Vulnerable
3396	EU	500057	Exact	2006	FT	Alytidae	*Alytes muletensis*	N	W	10	10	0	larva	PCR	Vulnerable
3397	EU	500059	Exact	2006	FT	Alytidae	*Alytes muletensis*	N	W	11	11	0	larva	PCR	Vulnerable
3398	EU	500062	Exact	2006	FT	Alytidae	*Alytes muletensis*	N	W	21	4	17	larva	PCR	Vulnerable
3399	EU	500063	Exact	2006	FT	Alytidae	*Alytes muletensis*	N	W	20	1	19	larva	PCR	Vulnerable
3400	EU	500058	Approximate	2006	FT	Alytidae	*Alytes muletensis*	N	W	3	3	0	metamorph	PCR	Vulnerable
3401	EU	500058	Approximate	2004	FT	Alytidae	*Alytes muletensis*	N	W	1	1	0	metamorph	PCR	Vulnerable
3402	EU	500088	Exact	2003	FT	Alytidae	*Alytes obstetricans*	N	W	85	74	11	mixed	PCR	Least Concern

Figure 5.24 Surveillance summary for Spain.

Each sample site is highlighted on the map and the record details listed below. Numerical summaries for number of sites, species and total number sampled are given to the right of the map as well as a table detailing the five species with the highest *Bd* prevalence and their corresponding IUCN status.

5.2.3.6 Other options

The underlying database is fully searchable using standard search forms by selecting the 'Record database' option from the navigation menu. Further options link to the *Bd* BLOG and an acknowledgements section detailing contributions to the project. The 'Publications' link loads a page containing all recent articles, referenced in PubMed (Figure 5.25). This page utilises an RSS feed from Hubmed [W38] mentioned in chapter 3. The RSS feed is incorporated into *Bd*-maps and provides an up-to-date list of articles that can be accessed directly. Using differential display of DIVs, the article abstracts are included and can be toggled on/off as required.

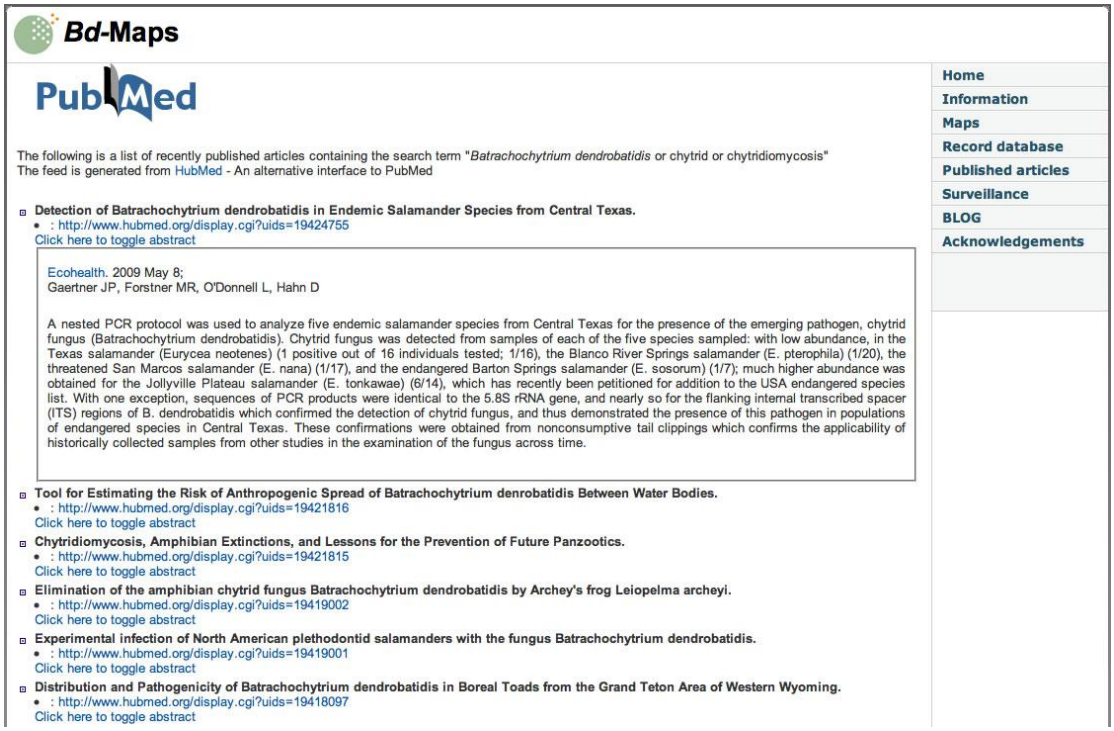


Figure 5.25 Recent *Bd* articles indexed in PubMed, derived from Hubmed.org.

Currently, *Bd*-Maps is accessed and utilised by members of the *Bd* research community and further development and usefulness will be guided by feedback on facilities widely agreed to be necessary. As mentioned, there is currently no genotypic data within the database, however, similarly to SRL-Maps, the addition of further data can be achieved easily and, within *Bd*-Maps, the geographic distribution of particular genotypes could be displayed in the same fashion as within MLST-Maps and SRL-Maps.

Since its initial release, *Bd*-Maps has been positively greeted and provides a much richer tool for the investigation of the picture of global incidence than the planned annual static map release. However, key to further usefulness and continued data submission is the development of a comprehensive curation system allowing the submission of data directly via the web rather then sending (by email) centrally for collation. Furthermore, *Bd* incidence data (with confirmation undertaken by subsequent laboratory diagnosis) are collected by multiple field workers, from many different locations. A potential area for further investigation is the use of field devices (such as mobile phones) that would allow direct submission of data to the central database from the field.

5.2.4 Geocoding and generating user maps: User-Maps

Available: http://www.spatialepidemiology.net/user_maps

5.2.4.1 Introduction

The applications detailed above (MLST-Maps, SRL-Maps and *Bd*-Maps) all present a geographical view for the interrogation of epidemiological databases. A major aim when developing web mapping tools for such databases is to promote further user submission of records with highly resolved spatial data, particularly within MLST-Maps where the majority of isolate data deposited includes only the country of origin. It is therefore imperative that current users, and interested users, have access to facilities allowing point co-ordinates for their data (latitude and longitudes) to be found. Excepting field workers who have access to GPS units (i.e. the majority of *Bd*-Maps data contributors), how does someone who has a number of geographically linked data points, plot their data on a map without access to a) GIS software and b) someone skilled in using such software?

To address this question I developed a very simple, generic geocoding and map creation tool allowing users to find latitude/longitudes for points and, once achieved, a simple method for plotting these points on a permanent map.

5.2.4.2 User-Maps Web Application

The 'user maps' facility at spatialepidemiology.net consists of four sections:

i) Map Latitude/Longitude Lookup.
ii) Batch Geocode Addresses.
iii) Simple Map Creation Tool.
iv) Advanced Map Creation Tool.

5.2.4.3 Map Latitude/Longitude lookup

Initially, the user is presented with a zoomed out view of Europe, By clicking, dragging and zooming, the user can locate the part of the world they wish to retrieve co-ordinates for and, by clicking the map, a marker appears with the latitude and

longitude indicated in an Info Window. Using click-drag-zoom, any point on the virtual representation of the earth can be geocoded in this way.

5.2.4.4 Batch geocoding

This facility allows a user to paste multiple addresses into a form for geocoding. The form consists of two text fields. The user pastes addresses into the first (search) and clicks submit. Following submission, each address is passed, via JavaScript, to the geocoding function within the Google Maps API [W19]. Geocoded results are returned directly to JavaScript and, using the DOM, the results text field is updated.
Data can be entered into the form in comma or TAB delimited format (e.g. New York, USA or New York USA). Geocoding can be undertaken at many levels of resolution: from street address or postal code (UK) or Zip (USA) (for example, "10 Downing Street, London" or "SW1 2AA, UK") to country level e.g. Brazil. For countries, cities, districts, postal codes and street addresses, the centroid of the area is returned. Results are returned as a series of columns, formatted as follows:

```
Address | Latitude | Longitude | Accuracy | No. of Addresses returned | Address or error.
```

Accuracy is an integer that corresponds to the following list:

0	Unknown location.
1	Country level accuracy.
2	Region (state, province, prefecture, etc.) level accuracy.
3	Sub-region (county, municipality, etc.) level accuracy.
4	Town (city, village) level accuracy.
5	Post code (zip code) level accuracy.
6	Street level accuracy.
7	Intersection level accuracy.
8	Address level accuracy.

As previously, the latitude and longitude represent the centroid of the location returned. Multiple locations may be returned if the address entered is ambiguous. For example, if we geocode 'London, USA', eight values are returned corresponding to the eight towns in the USA called London. It is therefore important for users to check the accuracy and include as much address information as possible when geocoding.

5.2.4.5 Simple map creation tool

The simple map creation tool allows a user to enter a list of geocoded locations, which are used to create a permanent map. An excel template is provided so that data can be formatted correctly prior to entering. Data must be entered into the simple map creation tool in TAB delimited format consisting of four columns as follows:

Name Latitude Longitude Description

Name represents the name given to the point.
Latitude and **longitude** are self-explanatory.
Description represents the data that appears in the Info Window when a point is clicked.

Following form submission, HTML representing a webpage containing a Google Map, with points plotted on the map, is written to the spatialepidemiology.net server and a link to this HTML document is provided to the user. These HTML files are kept on the server and the user can therefore quote the URL.

5.2.4.6 Advanced map creation tool

To allow slightly more flexibility in display options for User-Maps, the advanced map creation tool allows one extra column of data to be pasted in. This column contains an integer that references an icon that is used to represent the point on the map. A simple set of ten icon options are given (numbered 1-10) which are coloured on a temperature gradient allowing users to differentially display their data in a number of ways.

The geocoding and map creation tools provide a simple way for users to begin exploring their data geographically. However, the tools are currently very limited in display options. Ideally, further development would allow many more icon options, grouping of sets of icons and the ability to re-edit the map. Currently, once produced, a map cannot be amended without re-entering data as the overall aim was simply to allow users to explore possibilities. However, since launching the ‘User-Maps’ tool, a

number of requests have been made to produce more complex maps and it is therefore desirable for such further development to occur.

5.3 Discussion

The addition of web mapping tools to epidemiological databases provides users with an interactive method of exploring data within the geographical context they were collected. Each of the developments described (excepting User-Maps) provides simple methods for visualising and further interrogating the geographic structure of the data variables held within each database.

Within MLST-Maps and SRL-Maps, the focus of data interrogation is the genotype (the ST or the *spa* type) and the tools allow, based on this primary variable, the relationships of secondary variables (e.g. serotype or MRSA/MSSA status etc.) to be viewed as graphical summaries to aid interpretation of the data. Such a framework is applicable to other molecular epidemiological databases where individuals of a population are characterised by molecular methods and the aim is to relate genotypic to phenotypic variables. Such systems should be scaleable and the *Bd*-Maps application provided the opportunity to develop a method for incorporating large datasets with such mapping tools. All developments (again, excepting User-Maps) allow the easy addition of further variables (for example a genotypic method for *Bd*) and the choice of secondary variables utilised when exploring the data geographically should be guided by the questions we (and the user communities) wish to ask. A variable not currently utilised within the applications is time. Within MLST-Maps this is largely due to sampling bias and the lack of structured sampling. However, the decision by the SRLs who provided data for SRL-Maps, to repeat their structured survey provides the potential to develop a method for displaying and analysing the change in genotypic structure of the population over time. Furthermore, such methods are also applicable to the other developments described.

The main aim when initially incorporating mapping tools into MLST databases was to encourage the communities of researchers (and public health laboratories) who use mlst.net to begin submitting higher geographical resolution to the databases. The development of User-Maps was undertaken to provide simple methods for users to access the information required. The ability to geocode any location on the planet,

with or without reference to an address, provides users with the tools to achieve this aim. As mentioned in chapter 3, since the introduction of MLST-Maps, such finer geographical detail has been incorporated into the *Leptospira interrogans* MLST database (http://leptospira.mlst.net) and it is hoped that such collection strategies will be utilised within other organism-specific databases.

All the applications described utilise the principles of AJAX web development. The simple components of AJAX (described in the thesis introduction) provide a highly flexible toolkit for the addition of further functionality. Through the addition of further JavaScript functions which link to server-side scripts and/or software stored on the server, such applications are highly extensible and can adapt to the needs of the communities and the questions we (and they) wish to ask of the data. The rather grand aim, eluded to in the introduction, of providing GIS like functionality within web applications has only been achieved in part. One of the criticisms of web mapping tools is that they provide 'quasi-GIS' facilities, which are simply a descriptive framework for data display rather than any analytical functionality. This is true to some extent with respect to the applications described, however, the principles of AJAX do offer the potential for the addition of analytical methodologies only really limited by the questions being asked. For example, within both MLST-Maps and SRL-Maps, the distribution of genotypes can be viewed geographically (and secondary variables detailed) but there is no further inference of geographic structure within the population. eBURST provides a method for grouping closely related sets of STs within MLST databases but at present there is no linking of the resultant grouping to the location in geographic space of the STs analysed. Such a method could however be developed.

Currently there are a number of software packages that address the inference of structure within population genetic data, both at the sequence level and for multilocus data (e.g. ClonalFrame, Didelot and Falush, 2007; Structure, Pritchard et al., 2000 and BAPS, Corander et al., 2003): for a comprehensive review see Excoffier and Heckel (2006). Such software is used (similarly to eBURST) for inferring and investigating sub-divisions of closely related genotypes within a population. However, the methods vary greatly and few utilise spatial data when inferring sub-divisions. One package that demonstrates potential within this area is Geneland (Guillot et al., 2005). Geneland is an open-source program written in the R statistical programming

language [W80] which takes as inputs a set of multilocus genotypes and their corresponding point locations (e.g. latitude and longitude). Geneland incorporates an individual-centred method that focuses on the detection of immigrants within a population, based on the simple assumption that immigrants will exhibit different multilocus genotypes from natives. Once population sub-divisions are defined, the probability of sampled genotypes belonging to a particular subdivision can be viewed spatially as a series of heat-maps. Methods such as Geneland may prove useful in the future as increasing analysis of the biogeography of populations becomes neccessary. Rigorous testing, ideally utilising simulated data with known patterns of geographic and genotypic history, would be necessary before being confident that the technique would prove useful (in a similar fashion to the validation of the eBURST methodology undertaken by Turner et al, (2007)). However, from an application development point of view, such software could easily be incorporated using AJAX, in much the same way MEGA was utilised within eMLSA.net (detailed in chapter 4). Heat maps, indicating population membership, could be incorporated as map overlays and each time new data is added to a database, probability maps could be updated.

Another potential area, which would increase the analytical capabilities of such applications, is the addition of further data layers. A Web Mapping Service (WMS) is a standard protocol, developed by the Open Geospatial Consortium, for serving geographical data over the Internet [W81]. WMS servers provide web services, which allow developers to programmatically request particular sets of data (most often tile images) for incorporation into their applications, much in the same way that Google Maps provides a base mapping facility. A multitude of different WMS servers exist providing diverse data (e.g. average temperature over a period of time [W82]) and the incorporation of such data into web mapping applications holds great promise for the use of further analytical techniques. For example, the identification of climactic and geographic factors affecting the presence of *Bd* has led to the recognition that certain species of amphibian, at certain altitude ranges, are more at risk than others (Drew et al., 2005). Furthermore, climate change, and in particular large-scale warming, has a very strong positive correlation with species decline (Pounds et al., 2006). Using WMS servers, data pertaining to both average temperature and altitude could, theoretically, be combined with the location of sampling sites and the Surveillance pages within *Bd*-Maps could be developed to provide a richer tool. Furthermore, the inclusion of data layers (for example population density (and movement), transport

networks etc.) provides exciting potential for molecular epidemiological databases such as SRL-Maps and MLST-Maps.

Of primary concern when placing details of the location of a pathogen (and therefore the host) on a map, particularly when such data relates to human infection, is data privacy and patient confidentiality. A commonly used method to address this issue within disease mapping is to aggregate point data to higher groupings such as postcode, region or country (as is the case for MLST-Maps, through necessity rather than choice). However, concerns have been raised in the literature regarding regional differences in aggregated data. When aggregating by ZIP code in the USA, Eisen and Eisen (2007, 2008) point out that such aggregations are based on population density and that as such, the much greater population density in the north eastern United States creates finer scale dispersion than in the east. An alternative to aggregation, reviewed by Armstrong et al, (1999) is to randomly move each point location slightly to ensure privacy and to undertake analysis on this new set of locations. Concerns regarding the loss of fidelity when correlating epidemiological parameters to such randomly fuzzied data have recently been addressed by Wieland SC et al, (2008) who propose a method which adds noise to point locations but has been shown to both preserve the distribution of a disease and guarantee patient privacy.

An alternative method currently in use within the Biodiversity community is the point-radius method for georeferencing locality descriptions (Wieczorej et al., 2004). This method uses a 'best-guess' for exposure to a pathogen (e.g. an individual's residence or workplace) and describes the uncertainty in that georeference as a radius that expresses spatial uncertainty in a record. The geographical resolution within SRL-Maps, which is based on regional laboratory location, does not contain any detail regarding the potential exposure site for the individual presenting with a *S.aureus* infection at a hospital. The point-radius method may be useful in the future for recording, on a per-patient basis, more useful geographical detail. The same principles are also applicable for MLST data should geographic resolution increase.

In conclusion, web mapping for molecular epidemiological databases (and those which do not contain molecular data) can provide intuitive methods to geographically explore the data held within. Such methods are highly extensible and through the use

of AJAX, further tools can be provided which focus on a number of questions not currently addressed.

Chapter 6: General Discussion

The web development techniques detailed in the introduction to this thesis have been used to produce web applications that address different, but not exclusive, areas of microbial infectious disease microbiology. Generality was a guiding factor throughout. Chapter 2 detailed the development of an application utilised by a small group of researchers (including myself) to expand on previous knowledge and to understand the genetics of the capsular biosynthetic loci whose products lead to the production of the 90 *S.pneumoniae* serotypes known at the time of study. The methods utilised are directly applicable to other systems, consisting of sets of large chromosomal regions (or even full genomes) where sequences have a complicated evolutionary history. Chapter 3 detailed mlst.net, which by its nature is a generic system and applicable to a wide range of microbial species. Conceptually similar to mlst.net, eMLSA.net, detailed in Chapter 4 is also applicable to a number of different organisms, however the question addressed is somewhat different, that of bacterial speciation. The developments linked together within spatialepidemiology.net, the subject of Chapter 5, provide a generic framework for the geographical interrogation of epidemiological datasets with the initial aim of encouraging the submission of data, from users of databases such as mlst.net, at higher levels of spatial resolution.

The increasing sophistication of web development methodologies is apparent in the toolsets used to produce each application and the increasing use of the newer web development methods has allowed richer and more user-friendly applications to be developed which provide tools, which in a number of cases, would not have been possible without such programming principles.

AJAX and the use of APIs are currently driving innovation in HTML web development. This is evidenced by the increasingly frequent release cycle of the common web browsers. The developers of Firefox (Mozilla), Safari (Apple), Opera (Opera Software) and Chrome (Google) have placed increasing importance on developing JavaScript engines that are capable of running ever more complex JavaScript based applications efficiently. Indeed, the initial release of the Google Chrome browser was delayed significantly while the developers completely re-wrote the JavaScript engine for future-proofing. The only browser which does not have such

a frequent release cycle is Internet Explorer (Microsoft: currently the most widely used browser), however, the planned release of Internet Explorer 9 contains a significantly updated JavaScript engine and as such represents a substantial investment, by Microsoft, in the future of AJAX intensive web applications.

All of the standards utilised in AJAX, (and more generally in the set of technologies used for HTML development) are open. That is, the basic vocabulary of each language is provided by the W3C and each specification release is decided upon by a combination of differing sets of developers. Anyone can take part in the conversation leading up to the release of a new specification and as such allows provision of a truly open set of frameworks for developing web applications.

Open-source software is the cornerstone of traditional bioinformatics development, however, within web development, the paradigm is shifted somewhat. Within traditional software development, an application is produced using a particular programming language (e.g. C++ or Java) and in most cases for a particular platform (e.g. Windows, Mac OS or Linux). The software is compiled and deployed either through downloading and installing the compiled software, or from direct installation via another media (e.g. CD). Source code is subsequently opened by the developer for download and further amendment.

The increasing shift towards the internet as a platform for application development, providing functionality more often associated with traditional software, has led to the 'browser deployment model'. Briefly, whereas a GUI provides the interface to underlying software functionality in traditional software, the browser provides the GUI (which is specified using CSS etc. within the web applications presentation logic) and whereas the application logic is complied within traditional software, JavaScript provides part of the application logic within a web application with further logic coded within server-side scripts. Web applications are not compiled and as such it is easy to access the source code for most of the application logic directly from within the browser.

The majority of web browsers provide the option to 'View Source'. For example, when using Firefox, Safari or Internet Explorer this is achieved by selecting the 'File' menu option and then 'View Source' (or 'Page Source' in Firefox). The source

HTML for the page being viewed can be seen as plain text. As mentioned in the introduction, JavaScript files (and libraries) and CSS files are often linked to in the HEAD of a HTML document. These can also be retrieved using the 'view source' method. In this way, almost all components of a web application can be viewed directly by anyone. The only part of a web application that is not immediately accessible for viewing are the server-side scripts used for database and server software interaction. Server-side scripts produce HTML, which is passed to the browser, and in many cases it is quite easy to work out what each script is doing (e.g. if a database search is undertaken, the fields returned are displayed on the page). Furthermore, within AJAX development, the data passed via the *XMLHttpRequest* DOM object to JavaScript is kept to a minimum, and often the specific makeup of the data structures are transparent within the JavaScript functions. As such, this 'almost 100%' open source method of web software delivery provides the opportunity for other developers to understand the application logic within a web application and use and re-use as required. Furthermore, this will increase in importance for true open source applications as the web progresses further towards the primary platform for software delivery.

There are a number of areas in which current web applications, including those detailed within this thesis, lack the functionality found in traditional software. The first is the graphic capabilities within HTML development and the second is persistence of data and interaction with a user's computer. Graphic manipulation in HTML is almost non-existent; images are passed in common formats (such as JPEG, PNG etc.) and simply displayed within a HTML page. Google Maps is entirely powered by HTML development standards and each map view consists, in essence, of a set of different static images (tiles) Further static images, such as those used for markers, are then dynamically overlayed or removed. Any further graphical interactivity or dynamic animation requires the embedding, within HTML of other technologies. For example, Adobe Flash (such as those used to display dendrogram sub-clusters detailed in Chapter 2) or Java applications such as PhyloWidget described within eMLSA.net in Chapter 4. PhyloWidget is very much a minority development, which through the use of a comprehensive API, does allow a great deal of interactivity with a web application, and indeed its development was driven with the aim of tight integration within the web environment. However, the vast majority of such embedded applications, do not allow such a level of interactivity.

Furthermore, compiled Java applications such as ACT (Chapter 1) or eBURST (chapters 3 and 5), which use graphical methods to display, sequence alignments (and annotation) and population clusters respectively also offer a paucity of programmatic interactivity. Ideally, it would be advantageous to be able to create, amend and interact with complex images and graphics directly using JavaScript.

A further lack of functionality within web applications compared to traditional software is the lack of interaction with a user's computer. When using Excel, for example, one opens a spreadsheet from a location on the hard drive, undertakes the work required, then saves the file for re-use when next using the software. For a user to interact with a web application using their own data, an upload is required which sends data to the web server for processing. Ideally, such manipulation would take place directly within the browser (through application logic coded within JavaScript), requiring no server upload, and furthermore such manipulation should remain persistant (through an equivalent to 'saving') mirroring the traditional software workflow.

HTML 5 (originally referred to as Web Applications 1.0) is the next major revision of the HTML specification [W83]. Currently in working draft form, it contains, among other features, new markup and APIs for dynamic, scriptable rendering of bitmap images and offline storage of data. HTML5 represents a major advance in open web development and it has been suggested that it could make obsolete such embedded technologies as Flash and JavaFX [W84]. The 'canvas' element of HTML5 consists of a drawable region in HTML code which can be accessed through JavaScript and includes a full set of drawing functions similar to other 2D APIs for traditional programming languages (e.g. Java). The AppCache API within HTML5 is being developed for the provision of 'Offline Web Applications'. Currently, within a browser, the only way to store data on a users machine is using a 'Cookie': a simple text file which can contain simple data variables which can be accessed each time a user visites a particular URL. Cookies generally store very simple data such as passwords, or a users preference for viewing a site (colours etc.). AppCache provides a way of persistently storing JavaScript libraries, CSS files, images etc. that will allow all application logic currently undertaken within the browser to be stored and the web application to be used when offline. A further enhancement to the DOM

specification, ‘DOM storage’ provides data structures for the offline storage of data in a very similar (scriptable) fashion to data stored in databases.

Currently, the vast majority of the HTML 5 specification is supported in several browsers (Firefox, Opera and Chrome) with the notable exception being the most widely used browser, Internet Explorer. As such, the use of these new features was not exploited within the developments detailed within this thesis. However, there are a number of avenues worth exploring. Firstly, Canvas could be investigated as an alternative for the graphical rendering and manipulation within both eBURST and PhyloWidget. Secondly, to address issues that users may have with data privacy, and to provide the ability for users to analyse their data locally, without upload to central servers, AppCache could be utilised for the querying of MLST databases.

Whichever methods are utilised in the future, the prospect for the addition of increased functionality and tools for understanding the biology within current and subsequent web applications is great.

Bibliography

Publications

Aanensen, D.M. and Spratt, B.G. (2005) The multilocus sequence typing network: mlst.net *Nucleic Acids Res* 1(**33**):W728-33

Aanensen, D.M., Mavroidi, A., Bentley, S.D., Reeves, P.R. and Spratt, B.G. (2007) Predicted functions and linkage specificities of the products of the Streptococcus pneumoniae capsular biosynthetic loci. *J Bacteriol* 189(**21**):7856-76

Abbott, J.C., Aanensen, D.M., Rutherford, K., Butcher, S. and Spratt, B.G. (2005) WebACT – an online companion for the Artemis Comparison Tool. *Bioinformatics* 21(**18**):3665-6

Abbott, J.C., Aanensen, D.M. and Bentley, S.D. (2007) WebACT: an online genome comparison suite. *Methods Mol Biol* 395:57-74

Achtman, M. and Wagner, M. (2008) Microbial diversity and the genetic nature of microbial species. *Nat Rev Microbiol* 6(**6**):431-40

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25(**17**):3389–3402

Armstrong, M.P., Rushton, G. and Zimmerman, D.L. (1999) Geographically masking health data to preserve confidentiality. *Stat Med* 18:497-525

Arrecubieta, C., Garcia, E. and Lopez, R. (1995) Sequence and transcriptional analysis of a DNA region involved in the production of capsular polysaccharide in *Streptococcus pneumoniae* type 3. *Gene* 167:1-7

Austrian, R. (2004) *In* E.I, Tuomanen (ed), The Pneumococcus. ASM Press, Washington, D.C. ISBN:155581297X

Bennett, J.S., Jolley, K.A., Sparling, P.F., Saunders, N.J., Hart, C.A., Feavers, I.M. and Maiden, M.C. (2007) Species status of Neisseria gonorrhoeae: evolutionary and epidemiological inferences from multilocus sequence typing. *BMC Biol* 5:35

Bentley, S.D., Aanensen, D.M., Mavroidi, A., Saunders, D., Rabbinowitsch, E., Collins, M., Donohoe, K., Harris, D., Murphy, L., Quail, M.A., Samuel, G., Skovsted, I.C., Kaltoft, M.S., Barrel,l B., Reeves, P.R., Parkhill, J. and Spratt, B.G. (2006) Genetic analysis of the capsular biosynthetic locus from all 90 pneumococcal serotypes. *PLoS Genet* 2(**3**):e31.

Benson, D.A., Boguski, M., Lipman, D.J. and Ostell, J. (1994) GenBank. *Nucleic Acids Res* 22(**17**):3441-4

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostall, J. and Sayers, E.W. (2009) GenBank. *Nucleic Acids Res* 37:D26-31

Bernheimer, H.P. and Wermundsen, I.E. (1972) Homology in capsular transformation reactions in Pneumococcus. *Mol Gen Genet* 116:68-83

Bishop, C.J., Aanensen, D.M., Jordan, G.E., Kilian, M., Hanage, W.P. and Spratt, B.G. (2009) Assigning strains to bacterial species via the internet. *BMC Biology* **7**:3

Bohne-Lang, A., Lang, E., Forster, T. and von der Lieth, C.W. (2001) LINUCS: linear notation for unique description of carbohydrate sequences. *Carbohydr Res* 336(**1**):1-11

Bonofiglio, L., Garcia, E. and Mollerach, M. (2005) Biochemical characterization of the pneumococcal glucose-1-phosphate uridylytransferase (GalU) essential for capsule biosynthesis. *Curr Microbiol* 41(**4**):217-21

Bougnoux, M.E., Morand, S. and d'Enfert, C. (2002) Usefulness of multilocus sequence typing for characterization of clinical isolates of Candida albicans. *J Clin Microbiol* 40(**4**):1290-7

Bougnoux, M.E., Tavanti, A., Bouchier, C., Gow, N.A., Magnier, A., Davidson, A.D., Maiden, M.C., D'Enfert, C. and Odds, F. (2003) Collaborative consensus for optimized multilocus sequence typing of Candida albicans. *J Clin Microbiol* 41(**11**):5265-6

Bougnoux, M.E., Aanensen, D.M., Morand, S., Théraud M., Spratt, B.G. and d'Enfert, C. (2004) Multilocus sequence typing of *Candida albicans*: strategies, data exchange and applications. *Infect Genet Evol*. 4(**3**):243-52

Brueggemann, A.B., Griffiths, D.T., Meats, E., Peto, T., Crook, D.W. and Spratt, B.G. (2003) Clonal relationships between invasive and carriage Streptococcus pneumoniae and serotype- and clone-specific differences in invasive disease potential. *J Infect Dis* 187(**9**):1424-32

Cartee, R.T., Forsee, W.T. and Yother, J. (2005) Initiation and synthesis of the *Streptococcus pneumoniae* type 3 capsule on a phosphatidylglycerol membrane anchor. *J Bacteriol* 187(**13**):4470-9

Cartright, K. (2002) Pneumococcal disease in Western Europe: Burden of disease, antibiotic resistance and management. *Eur J Pediatr* 161:188-195

Carver, T., Thomson, N., Bleasby, A., Berriman, M. and Parkhill, J. (2009) DNAPlotter: circular and linear interactive genome visualisation. *Bioinformatics* 25(**1**):19-20

Carver, T.J., Rutherford, K.M., Berriman, M., Rajandream, M.A., Barrell, B.G. and Parkhill, J. (2005) ACT: the Artemis Comparison Tool. *Bioinformatics* 21(**16**):3422-3

Chan, M.S., Maiden, M.C. and Spratt, B.G. (2001) Database-driven multi locus sequence typing (MLST) of bacterial pathogens. *Bioinformatics* 17(**11**):1077-83

Chevenet, F., Brun, C, Bañuls, A.L., Jacq, B. and Christen, R. (2006) TreeDyn: towards dynamic graphics and annotations for analyses of trees. *BMC Bioinformatics* 7:439

Cieslewicz, M.J., Kasper, D.L., Wang, Y. and Wessels, M.R. (2001) Functional analysis in type 1a group B Streptococcus of a cluster of genes involved in extracellular polysaccharide production by diverse species of streptococci. *J Biol Chem* 276:139-146

Cieslewicz, M.J., Chaffin, D., Glusman, G., Kasper, D., Madan, A., Rodrigues, S., Fahey, J., Wessels, M.R. and Rubens, C.E. (2005) Structural and Genetic Diversity of Group B *Streptococcus* Capsular Polysaccharides. *Infect Immun* 73(**5**):3096-3103

Clamp, M., Cuff, J., Searle, S.M. and Barton, G.J. (2004) The Jalview Java alignment editor. *Bioinformatics* 20(**3**):426-7

Coffey, T.J., Enright, M.C., Daniels, M., Morona, J.K., Morona, R., Gryniewicz, W., Paton, J.C. and Spratt, B.G. (1998) Recombinational exchange at the capsular polysaccharide biosynthetic locus leads to frequent serotype changes among natural isolates of *Streptococcus pneumoniae Mol Microbiol* 27:73-83

Coffey, T.J., Daniels, M., Enright, M.C. and Spratt, B.G. (1999) Serotype 14 variants of the Spanish penicillin-resistant serotype 9V clone of *Streptococcus pneumoniae* arose by large recombinational replacements of the *cpsA-pbp1a* region. *Microbiol* 145:2023-2031

Cohan, F.M. (2002) What are bacterial species? *Annu Rev Microbiol* 56:457-87

Corander, J., Waldmann, P. and Sillanpää, M.J. (2003) Bayesian analysis of genetic differentiation between populations. *Genetics*, 163:367-374.

Croze, M., Dauwalder, O., Dumitrescu, O., Badiou, C., Gillet, Y., Genestier, A.L., Vandenesch, F., Etienne, J. and Lina, G. (2009) Serum antibodies against Panton-Valentine leukocidin in a normal population and during *Staphylococcus aureus* infection. *Clin Microbiol Infect* 15(**2**):144-8

Daszak, P., Cunningham, A.A. and Hyatt, A.D. (2003) Infectious disease and amphibian population declines. *Divers Distrib* 9:141-150

Didelot, X. and Falush, D. (2007) Inference of Bacterial Microevolution Using Multilocus Sequence Data. *Genetics* 175:1251-1266

Dodgson, A.R., Pujol, C., Denning, D.W., Soll, D.R. and Fox A.J. (2003) Multilocus sequence typing of *Candida glabrata* reveals geographically enriched clades. *J Clin Microbiol* **12**:5709-5717

Doolittle, W.F. (1999) Phylogenetic Classification and the Universal Tree. *Science* 284:2124-2128

Dowson, C.G., Hutchison, A., Brannigan, J.A., George, R.C., Hansman, D., Liñares, J., Tomasz, A., Smith, J.M. and Spratt, B.G. (1989) Horizontal transfer of penicillin-binding protein genes in penicillin-resistant clinical isolates of Streptococcus pneumoniae. *Proc Natl Acad Sci USA* 86(**22**):8842-6

Drew, A., Allen, E.J. and Allen, L.J. (2005) Analysis of climactic and geographic factors affecting the presence of chytridiomycosis in Australia. *Dis Aquat Organ* 68(**3**):245-50

Eaton, A.D. (2006) HubMed: a web-based biomedical literature search interface. *Nucleic Acids Res* 34:W745-7

Eisen, L. and Eisen, R.J (2008) Improving Methods for Reporting Spatial Epidemiologic Data. *Emerg Infect Dis* 14(**8**):1335-1337

Eisen, L. and Eisen, R.J. (2007) Need for Improved Methods to Collect Present Spatial Epidemiologic Data for Vectorborne Diseases. *Emerg Infect Dis* 13(**12**):1816-1820

Emmert, D.B., Stoehr, P.J., Stoesser, G. and Cameron, N. (1994) The European Bioinformatics Institute (EBI) databases. *Nucleic Acids Res* 44(**17**):3445-3449

Enright, A.J., Hunin, V. and Ouzounis, C.A. (2003) Protein families and TRIBES in genome sequence space. *Nucleic Acids Res* 15:4632-4638

Enright, M.C. and Spratt, B.G. (1998) A multilocus sequence typing scheme for Streptococcus pneumoniae: identification of clones associated with serious invasive disease. *Microbiology* 144 (**11**):3049-60

Enright, M.C., Day, N.P., Davies, C.E., Peacock, S.J. and Spratt, B.G. (2000) Multilocus sequence typing for characterization of methicillin-resistant and methicillin-susceptible clones of Staphylococcus aureus. *J Clin Microbiol* 38(**3**):1008-15

Enright, M.C., Spratt, B.G., Kalia, A., Cross, J.H. and Bessen, D.E. (2001) Multilocus sequence typing of Streptococcus pyogenes and the relationships between emm type and clone. *Infect Immun* 69(**4**):2416-27

Enright, M.C., Robinson, D.A., Randle, G., Feil, E.J., Grundmann, H. and Spratt, B.G. (2002) The evolutionary history of methicillin-resistant Staphylococcus aureus (MRSA). *Proc Natl Acad Sci USA* 99(**11**):7687-92

Excoffier, L. and Heckel, G. (2006) Computer programs for population genetics data analysis: a survival guide. *Nature Reviews Genetics* 7:745-758

Feil, E.J., Cooper, J.E., Grundmann, H., Robinson, D.A., Enright, M.C., Berendt, T., Peacock, S.J., Smith, J.M., Murphy, M., Spratt, B.G., Moore, C.E. and Day, N.P. (2003) How clonal is *Staphylococcus aureus*? *J Bacteriol* 185:3307-3316.

Feil, E.J., Li, B.C., Aanensen, D.M., Hanage, W.P. and Spratt, B.G. (2004) eBURST: inferring patterns of evolutionary descent among clusters of related bacterial genotypes from multilocus sequence typing data. *J Bacteriol* 186(**5**):1518-30

Feil, E.J., Holmes, E.C., Besses, D.E., Chan, M.S., Day, N.P., Enright, M.C., Goldstein, R., Hood, D.W., Kalia, A., Moore, C.E., Zhou, J. and Spratt, B.G. (2001) Recombination within natural populations of pathogenic bacteria: short-term

empirical estimates and long-term phylogenetic consequences. *Proc Natl Acad Sci USA* 98(**1**):182-7

Finn, A. (2004) Bacterial polysaccharide-protein conjugate vaccines. *Br Med Bull* 70:1-14

Fraser, C., Hanage, W.P. and Spratt, B.G. (2007) Recombination and the nature of bacterial speciation. *Science* 315(**5811**):476-80

Freifeld, C.C., Mandi, K.D., Reis, B.Y. and Brownstein, J.S. (2008) HealthMap: global infectious disease monitoring through automated classification and visualisation of internet media reports. *JAMIA* 15(**2**):150-7

Frénay, H.M., Bunschoten, A.E., Schouls, L.M., van Leeuwen, W.J., Vandenbroucke-Grauls, C.M., Verhoef, J. and Mooi, F.R. (1996) Molecular typing of methicillin-resistant Staphylococcus aureus on the basis of protein A gene polymorphism. Eur J Clin Microbiol Infect Dis 15(**1**):60-4

Galperin, M.Y. and Cochrane, G.R. (2009) Nucleic Acids Research annual Database Issue and the NAR online Molecular Biology Database Collection in 2009. *Nucleic Acids Res* 37:D1-4

Gevers, D., Cohan, F.M., Lawrence, J.G., Spratt, B.G., Coenye, T., Feil, E.J., Stackenbrandt, E., Van de Peer, Y., Vandamme, P., Thonpson, F.L. and Swings, J. (2005) Opinion: Re-evaluating prokaryotic species. *Nat Rev Microbiol* 3(**9**):733-9

Glucksmann, M.A., Reuber, T.L. and Walker, G.C. (1993) Genes needed for modification, polymerization, export and processing of succinoglycan by *Rhizobium meliloti*: a model for succinoglycan biosynthesis. *J Bacteriol* 182:1854-1863

Godoy, D., Randle, G., Simpson, A.J., Aanensen, D.M., Pitt, T.L., Kinoshita, R. and Spratt, B.G. (2003) Multilocus sequence typing and evolutionary relationships among the causative agents of melioidosis and glanders, Burkholderia pseudomallei and Burkholderia mallei. *J Clin Microbiol* 41(**5**):2068-79

Green, D.E. and Dodd Jr, K. (2007) Presence of amphibian chytrid fungus *Batrachochytrium dendrobatidis* and other amphibian pathogens at warm-water fish hatcheries in Southeastern North America. *Herpetol Conserv Biol* 2(**1**):25

Guillot, G., Mortier, F. and Estoup, A. (2005) Geneland: A program for landscape genetics. *Mol Ecol Notes.* 5:712-715.

Hakenbeck, R., Balmelle, N., Weber, B., Gardès, C., Keck, W. and de Saizieu, A. (2001) Mosaic genes and mosaic chromosomes: intra- and interspecies genomic variation of Streptococcus pneumoniae. *Infect Immun* 69(**4**):2477-86

Hanage, W.P., Fraser, C. and Spratt, B.G. (2006) Sequences, sequence clusters and bacterial species. *Phil Trans Roy Soc B* 361:1917-1927

Hanage, W.P., Fraser, C. and Spratt, B.G. (2005) Fuzzy species in recombinogenic bacteria. *BMC Biol* 3:6

Hardy, G.G., Caimano, M.J. and Yother, J. (2000) Capsule biosynthesis and basic metabolism in *Streptococcus pneumoniae* are linked through the cellular phosphoglucomutase. *J Bacteriol* 182(**7**):1854-1863

Heidelberger, M. (1983) Precipitating cross-reactions among pneumococcal types. *Infect Immun* 41:1234-44

Henrichsen, J. (1995) Six newly recognized types of *Streptococcus pneumoniae. J Clin Microbiol* 33:2759-62

Hohwy, J. and Kilian, M. (1995) Clonal diversity of the *Streptococcus mitis* biovar 1 population in the human oral cavity and pharynx. *Oral Microbiol Immunol* 10:19-25

Holden, M.T., Heather, Z., Paillot, R., Steward, K.F., Webb, K., Ainslie, F., Jourdan, T., Bason, N.C., Holroyd, N.E., Mungall, K., Quail, M.A., Sanders, M., Simmonds, M., Willey, D., Brooks, K., Aanensen, D.M., Spratt, B.G., Jolley, K.A., Maiden, M.C., Kehoe, M., Chanter, N., Bentley, S.D., Robinson, C., Maskell, D.J., Parkhill, J. and Waller, A.S. (2009) Genomic evidence for the evolution of *Streptococcus equi*: host restriction, increased virulence, and genetic exchange with human pathogens. *PLoS Pathog* 5(**3**):e1000346

Homan, W.L., Tribe, D., Poznanski, S., Li, M., Hogg, G., Spalburg, E., Van Embden, J.D. and Willems, R.J. (2002) Multilocus sequence typing scheme for Enterococcus faecium. *J Clin Microbiol* 40(**6**):1963-71

Hoshino, T., Fujiwara, T. and Kilian, M. (2005) Use of phylogenetic and phenotypic analyses to identify nonhemolytic streptococci isolated from bacteremic patients. *J Clin Microbiol* 43(**12**):6073-85

Huson, D.H. (1998) SplitsTree: analyzing and visualizing evolutionary data. *Bioinformatics* 14(**1**):68-73

Huson, D.H., Richter, D.C., Rausch, C., Dezulian, T., Franz, M. and Rupp, R. (2007) Dendroscope: An interactive viewer for large phylogenetic trees. *BMC Bioinformatics* 8:460

Iannelli, F., Pearce, B.J. and Pozzi, G. (1999) The Type 2 Capsule Locus of *Streptococcus pneumoniae*. *J Bact* 181(**8**):2652-2654

Jordan, G.E and Piel, W.H. (2008) PhyloWidget: web-based visualisations for the tree of life. *Bioinformatics* 24(**14**):1641-2

Kamerling, J.P. (2000) Pneumococcal polysaccharides: A chemical view. *In* A.Tomasz (ed) *Streptococcus pneumoniae*: Molecular biology and mechanisms of disease. Mary Ann Liebert, Larchmont, USA. ISBN: 0913113859

Keseler, I.M., Collado-Vides, J., Gama-Castro, S., Ingraham, J., Paley, S., Paulsen, I.T., Peralta-Gil, M. and Karp, P.D. (2005) EcoCyc: a comprehensive database resource for Escherichia coli. *Nucleic Acids Res* 33:D334-7

Kikuchi, N., Kameyama, A., Nakaya, S., Ito, H., Sato, T., Shikanai, T., Takahashi, Y. and Narimatsu, H. (2005) The carbohydrate sequence markup language (CabosML): and XML description of carbohydrate structures. *Bioinformatics* 21(**8**):1717-8

Kilian, M., Poulsen, K., Blomqvist, T., Håvarstein, L.S., Bek-Thomsen, M., Tettelin, H. and Sørensen, U.B. (2008) Evolution of *streptococcus pneumoniae* and its close commensal relatives. *PLoS ONE* 5:35

Kilian, M (2005) p833-881 *In* Topley and Wilson's Microbiology and Microbial Infections. Borriello, P., Murray, P.R., Funke, G (eds.) Hodder Arnold, London

King, S.J., Leigh, J.A., Heath, P.J., Luque, I., Tarradas, C., Dowson, C.G. and Whatmore, A.M. (2002) Development of a multilocus sequence typing scheme for the pig pathogen Streptococcus suis: identification of virulent clones and potential capsular serotype exchange. *J Clin Microbiol* 40(**10**):3671-80

Kolkman, M.A.B., Wakarchuk, W., Nuijten, P.J.M. and van der Zeijst, B.A.M. (1997) Capsular polysaccharide synthesis in *Streptococcus pneumoniae* serotype 14: molecular analysis of the complete *cps* locus and identification of genes encoding glycosyltransferases required for the biosynthesis of the tetrasaccharide subunit. *Mol Microbiol* 26(**1**):197-208

Konstantinidis, K.T. and Tiedje, J.M. (2005) Towards a genome-based taxonomy for prokaryotes. *J Bacteriol* 187(**18**):6258-64

Kraak, M.N. (2001) Settings and needs for web cartography.p3-4 *In* Kraak and Allan (eds.) Web Cartography (Geographic Information Systems Workshop). Francis and Taylor, New York. ISBN: 978-0748408696

Kuiken, T., Leighton, F.A., Fouchier, R.A., LeDuc, J.W., Peiris, J.S., Schudel, A., Stöhr, K. and Osterhaus, A.D. (2005) Public health. Pathogen surveillance in animals. *Science* 309(**5741**):1680-1

Kumar, S., Nei, M., Dudley, J. and Tamura, K. (2008) MEGA: a biologist-centric software for evolutionary analysis of DNA and protein sequences. Brief *Bioinformatics* 9(**4**):299-306

Laible, G., Spratt, B.G. and Hakenbeck, R. (1991) Interspecies recombinational events during the evolution of altered PBP 2^{x} genes in penicillin-resistant clinical isolates of *Streptococcus pneumoniae. Mol Microbiol* 5:1993-2002

Lawson Handley, L.J., Manica, A., Goudet, J. and Balloux, F. (2007) Going the distance: human population genetics in a clinal world. *Trends Genetics* 23(**9**):432-439
Lemercinier, X. and Jones, C. (2006) Full assignment of the 1H and 13C spectra and revision of the O-acetylation site of the capsular polysaccharide of Streptococcus pneumoniae Type 33F, a component of the current pneumococcal polysaccharide vaccine. *Carbohydr Res* 341(1):68-74

Llull, D., Lopez, R. and Garcia E. (2001) Genetic basis and medical relevance of capsular biosynthesis in pathogenic streptococci. *Curr Mol Med* 1:475-91

Longcore, J.E., Pessier, A.P. and Nichols, D.K. (1999) *Batrachochytrium dendrobatidis* gen. et sp. Nov., a chytrid pathogenic to amphibians. *Mycologia* 91(**2**):219-227

Maiden, M.C.J., Bygraves, J.A., Feil, E., Morelli, G., Russell, J.E., Urwin, R., Zhang, Q., Zhou, J., Zurth, K., Caugant, D.A., Feavers, I.M., Achtman, M. and Spratt, B.G. (1998). Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc Natl Acad Sci USA* vol. 95(**6**):3140-5

Margos, G., Gatewood, A.G., Aanensen, D.M., Hanincová, K., Terekhova, D., Vollmer, S.A., Cornet, M., Piesman, J., Donaghy, M., Bormane, A., Hurn, M.A., Feil, E.J., Fish, D., Casjens, S., Wormser, G.P., Schwartz, I. and Kurtenbach, K. (2008) MLST of house-keeping genes captures geographic population structure and suggests a European origin of Borrelia burgdorferi. *Proc Natl Acad Sci USA* 105(**25**):8730-5

Martens, M., Dawyndt, P., Coopman, R., Gilis, M., De Vos, P. and Willems, A. (2008) Advantages of multilocus sequence analysis for taxonomic studies: a case study using 10 house-keeping genes in the genus *Ensifer* (including former *Sinorhizobium*) *Int J Syst Evol Microbiol* 58:200-214

Mavroidi, A., Aanensen, D.M., Godoy, D., Skovsted, I.C., Kaltoft, M.S., Reeves, P.R., Bentley, S.D. and Spratt, B.G. (2007) Genetic relatedness of the Streptococcus pneumoniae capsular biosynthetic loci. *J Bacteriol* 189(**21**):7841-55

McGee, L., McDougal, L., Zhou, J., Spratt, B.G., Tenover, F.C., George, R., Hakenbeck, R., Hryniewicz, W., Lefévre, J.C., Tomasz, A. and Klugman, K.P. (2001) Nomenclature of Major Antimicrobial-Resistant Clones of *Streptococcus pneumoniae* Defined by the Pneumococcal Molecular Epidemiology Network. *J Clin Micro* 39(**7**):2565-2571

Meats, E., Feil, E.J., Stringer, S., Cody, A.J., Goldstein, R., Kroll, J.S., Popovic, T. and Spratt, B.G. (2003) Characterization of encapsulated and noncapsulated Haemophilus influenzae and determination of phylogenetic relationships by multilocus sequence typing. *J Clin Microbiol* 41(**4**):1623-36

Meyer, W., Aanensen, D.M., Boekhout, T., Cogliati, M., Diaz, M.R., Esposto, M.C., Fisher, M., Gilgado, F., Hagen, F., Kaocharoen, S., Litvintseva, A.P., Mitchell, T.G., Simwami, S.P., Trilles, L., Viviani, M.A. and Kwon-Chung, J. (2009) Consensus multi-locus sequence typing scheme for Cryptococcus neoformans and Cryptococcus gattii. *Med Mycol* 12:1-14

Morona, J.K., Morona, R. and Paton, J.C. (1997) Characterization of the locus encoding the *Streptococcus pneumoniae* type 19F capsular polysaccharide biosynthetic pathway. *Mol Microbiol* 23(**4**):751-63

Morona, J.K., Paton, J.C., Miller, D.C. and Morona, R. (2000) Tyrosine phosphorylation of CpsD negatively regulates capsular polysaccharide biosynthesis in streptococcus pneumoniae. *Mol Microbiol* 35(**6**):1431-42

Morona, J.K., Morona, R. and Paton, J.C. (2006) Attachment of capsular polysaccharide to the cell wall of *Streptococcus pneumoniae* type 2 is required for invasive disease. *Proc Natl Acad Sci USA* 103(**22**):8505-10

Munoz, R., Mollerach, M., Lopez, R. and Garcia E. (1997) Molecular organization of the genes required for the synthesis of type 1 capsular polysaccharide of *Streptococcus pneumoniae*: formation of binary encapsulated pneumococci and identification of cryptic dTDP-rhamnose biosynthesis genes. *Mol Microbiol* 25(**1**):79-92

Munoz, R.M., Moller ach, R., Lopez, R. and Garcia, E. (1999) Characterization of the Type 8 Capsular Gene Cluster of *Streptococcus pneumoniae*. *J Bact* 181(**19**):6214-6219

Neumann, A. (2008) Open Source GIS Software. pp1262 *In* S., Shekar and H, Xiong (eds.) Encyclopedia of GIS. Springer. ISBN: 978-0-387-30858-6

Nielsen, J. (1999) Designing Web Usability. New Riders Publishing, Indianapolis, Indiana USA. ISBN:1-56205-810-X

Nørskov-Lauritsen, N., Bruun, B. and Kilian, M. (2005) Multilocus sequence phylogenetic study of the genus Haemophilus with description of Haemophilus pittmaniae sp. nov. *Int J Syst Evol Microbiol* 55(**1**):449-56

Page, R.D. (1996) TreeView: an application to display phylogenetic trees on personal computers. Comput Appl Biosci 12(**4**):357-8

Paiment, A., Hocking, J. and Whitfield, C. (2002) Impact of phosphorylation of specific residues in the tyrosine autokinase, Wzc, on its activity in assembly of group 1 capsules *Escherichia coli J Bacteriol* 184(**23**):6437-47

Park, I.H., Pritchard, D.G., Cartee, R., Brandao, A., Brandileone, M.C. and Nahm, M.H. (2007) Discovery of a new capsular serotype (6C) within serogroup 6 of *Streptococcus pneumoniae*. *J Clin Microbiol* 45(**4**):1225-33

Pelosi, L., Boumedienne, M., Saksouk, N., Geiselmann, J. and Geremia, R.A. (2005) The glucosyl-1-phosphate transferase WchA (Cap8E) primes the capsular polysaccharide repeat unit biosynthesis of *Streptococcus pneumoniae* serotype 8. *Biochem Biophys Res Com* 327:857-865

Poolman, J.T. (2004) Pneumococcal vaccine development. *Expert Rev Vaccines* 3:597-604

Postic, D., Garnier, M. and Baranton, G. (2007) Multilocus sequence analysis of atypical Borrelia burgdorferi sensu lato isolates--description of Borrelia californiensis sp. nov., and genomospecies 1 and 2. *Int J Med Microbiol* 297(**4**):263-71

Pritchard, J.K., Stephens, M. and Donnelly, P. (2000) Inference of Population structure using Multilocus Genotype Data. *Genetics* 155:945-959

Pounds, J.A., Bustamante, M.R., Coloma, L.A., Consuegra, J.A., Fogden, M.P., Foster, P.N., La Marca, E., Masters, K.L., Merino-Viteri, A., Puschendorf, R., Ron,

S.R. Sanchez-Azofeifa, G.A., Still, C.J., and Young, B.E. (2006) Widespread amphibian extinctions from epidemic disease driven by global warming. *Nature* 439(**7073**):161-7

Raskin, J. (2000) The Humane Interface: New directions for designing Interactive Systems. Addison Wesley, UK. ISBN:0201379376

Reeves, P.R., Hobbs, M., Valvano, M.A., Skurnik, M., Whitfield, C., Coplin, D., Kido, N., Klena, J., Maskell, D., Raetz, R.H. and Rick, P.D. (1996) Bacterial polysaccharide synthesis and gene nomenclature. *Trends Microbiol* 4:495-503

Rice, P., Longden, I. and Bleasby, A. (2000) EMBOSS: the European Molecular Biology Open Source Software Suite. *Trends Genet* 16(**6**):276-7

Romero, P. and Karp, P. (2003) PseudoCyc, A pathway-Genome Database for *Pseudomonas aeruginosa*. *J Mol Microbiol Biotechnol* 5(**4**):230-9

Rowley, J.J., Chan, S.K., Tang, W.S., Speare, R., Skerratt, L.F., Alford, R.A., Cheung, K.S., Ho, C.Y. and Campbell, R. (2007) Survey for the amphibian chytrid Batrachochytrium dendrobatidis in Hong Kong in native amphibians and in the international amphibian trade. *Dis Aquat Org* 78(**2**):87-95

Ruiz-Garbajosa, P., Bonten, M.J., Robinson, D.A., Top, J., Nallapareddy, S.R., Torres, C., Coque, T.M., Canton, R., Baquero, F., Murray, B.E., del Campo, R. and Willems, R. (2006) Multilocus sequence typing scheme for Enterococcus faecalis reveals hospital-adapted genetic complexes in a background of high rates of recombination. *J Clin Microbiol* 44(**6**):2220-8

Saksouk, N., Pelosi, L., Colin-Morel, P., Boumedienne, M., Abdian, P.L. and Geremia, R.A. (2005). The capsular polysaccharide biosynthesis of Streptococcus pneumoniae serotype 8: functional identification of the glycosyltransferase WciS (Cap8H). *Biochem J* 389(**1**):63-72

Sawabe, T., Kita-Tsukamoto, K. and Thompson, F.L. (2007) Inferring the evolutionary history of vibrios by means of multilocus sequence analysis. *J Bacteriol* 189:7932-7936

Selander, R.K., Caugant, D.A., Ochman, H., Musser, J.M., Gilmour, M.N. and Whittam, T.S. (1986) Methods of multilocus enzyme electrophoresis for bacterial population genetics and systematics. *Appl Environ Microbiol* 51(**5**):873-84

Selander, R.K., Musser, J.M., Caugant, D.A., Gilmour, M.N. and Whittam, T.S. (1987) Population genetics of pathogenic bacteria. *Microb Pathog* 3(**1**):1-7

Snel, B., Huynen, M.A. and Dutilh, B.H. (2005) Genome trees and the nature of genome evolution. *Annu Rev Microbiol* 59:191-209

Snel, B., Bork, P. and Huynen, M.A. (1999) Genome phylogeny based on gene content. *Nat Genet* 21(**1**):108-10

Sonnhammer, E.L., Eddy, S.R. and Durbin, R. (1997) Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins* 28(**3**):405-20

Sørenson, U.B., Henrichsen, J., Chen, H.C. and Szu, S.C. (1990) Covalent linkage between the capsular polysaccharide and the cell wall peptidoglycan of *Streptococcus pneumoniae* revealed by immunochemical methods. *Microb Pathol* 8:325-334

Spratt, B.G., Feil, E.J. and Smith, N.H. (2001) Population and evolutionary dynamics of pathogenic bacteria. *Bioessays* 22(**12**):1115-22

Spratt, B.G., Hanage, W.P. and Brueggemann, A. (2004) Evolutionary and population biology of *Streptococcus pneumoniae*. ASM Press, Washington (D.C.)

Spratt, B.G., Hanage, W.P., Li, B., Aanensen, D.M. and Feil, E.J. (2004) Displaying the relatedness among isolates of bacterial species - the eBURST approach. FEMS *Microbiol Lett* 241(**2**):129-34

Stackebrandt, E., Frederiksen, W., Garrity, G.M., Grimont, P.A., Kämpfer, P., Maiden, M.C., Nesme, X., Rosselló-Mora., Swings, J., Trüper, H.G., Vauterin, L., Ward, A.C. and Whitman, W.B. (2002) Report of the ad hoc committee for the re-evaluation of the species definition in bacteriology. *Int J Syst Evol Microbiol* 52(**3**):1043-7

Tacconelli, E., Pop-Vicas, A.E. and D'Agata, E.M. (2006) Increased mortality among elderly patients with meticillin-resistant *Staphylococcus aureus* bacteraemia. *J Hosp Infect* 64:251-256.

Tang, C., Reyes, J.F., Luciani, F., Francis, A.R. and Tanaka M.M. (2008) spolTools: online utilities for analyzing spoligotypes of the *Mycobacterium tuberculosis* complex. *Bioinformatics* 24(**20**):2414-5.

Tateno, Y., Fukami-Kobayashi, K., Miyazaki, S., Sugawara, H. and Gojobori, T. (1998) DNA Data Bank of Japan at work on genome sequence data. *Nucleic Acids Res* 26(**1**):16-20

Taylor, C.M. and Roberts, I.S. (2005) Capsular polysaccharides and their role in virulence. *Contrib Microbiol* 12:55-66

Tettelin, H., Nelson, K.E., Paulsen, I.T., Eisen, J.A., Read, T.D., Peterson, S., Heidelberg, J.,DeBoy, R.T., Haft, D.H., Dodson, R.J., Durkin, A.S., Gwinn, M., Kolonay, J.F., Nelson, W.C., Peterson, J.D., Umayam, L.A., White, O., Salzberg, S.L., Lewis, M.R., Radune, D., Holtzapple, E., Khouri, H., Wolf, A.M., Utterback, T.R., Hansen, C.L., McDonald, L.A., Feldblyum, T.V., Angiuoli, S., Dickinson, T., Hickey, E.K., Holt, I.E., Loftus, B.J., Yang, F., Smith, H.O., Venter, J.C, Dougherty, B.A., Morrison, D.A., Hollingshead, S.K. and Fraser, C.M. (2001) Complete genome sequence of a virulent isolate of *Streptococcus pneumoniae*. *Science* 293(**5529**):498-506

Thaipadungpanit, J., Wuthiekanum, V., Chierakul, W., Smythe, L.D., Petkanchanpong, W., Limpaiboon, R., Apiwatanaporn, A., Slack, A.T., Suputtamongkol, Y., White, N.J., Feil, E.J., Day, N.P. and Peacock, S.J. (2007) A Dominant Clone of Leptospira interrogans Associated with an Outbreak of Human Leptospirosis in Thailand. *PLoS Negl Trop Dis* 1(**1**):e56

Thomas, J.C., Vargas, M.R., Miragaia, M., Peacock, S.J., Archer, G.L. and Enright, M.C. (2007) Improved multilocus sequence typing scheme for Staphylococcus epidermidis. *J Clin Microbiol* 45(**2**):616-9

Thomson, J.D., Higgins, D.G. and Gibson, T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple alignment through sequence weighting, position specific gap penalties and weight matrix choice. *Nucleic Acids Res* 22:4673-4680

Thompson, J.D., Gibson, T.J. and Higgins, D.G. (2002) Multiple sequence alignment using ClustalW and ClustalX. *In Current protocols in bioinformatics* Andreas D Baxevanis [ed] vol. Chapter 2 pp. Unit 2.3 John Wiley & Sons, UK

Thorisson, G.A., Smith, A.V., Krishnan, L. and Stein, L.D. (2005) The International HapMap Project Web Site. *Genome Research* 15(**11**):1592-1593

Turner, K.M., Hanage, W.P., Fraser, C., Connor, T.R. and Spratt, B.G. (2007) Assessing the reliability of eBURST using simulated populations with known ancestry. *BMC Microbiol* 7:30

Vandamme, P., Pot, B., Gillis, M., de Vos, P., Kersters, K. and Swings, J. (1996) Polyphasic taxonomy, a consensus approach to bacterial systematics. *Microbiol Rev* 60(**2**):407-38

Wake, D.B. and Vredenburg, V.T. (2008) Colloquium paper: are we in the midst of the sixth mass extinction? A view from the world of amphibians. *Proc Natl Acad Sci USA* 105(**1**):11466-73

Walker, S.F., Bosch, J., James, T.Y., Litvintseva, A.P., Oliver Vallas, J.A., Piña, S., Garcia, G., Rosa, G.A., Cunningham, A.A., Hole, S., Griffiths, R. and Fisher, M.C. (2007) Invasive pathogens threaten species recovery programs. *Curr Biol* 18(**18**):R853-4

Wang, L., Liu, D. and Reeves, P.R. (1996) C-terminal half of *Salmonella enterica* WbaP (RfbP) is the galactosyl-1-phosphate transferase domain catalyzing the first step of *O*-antigen biosynthesis. *J Bacteriol* 178:2598-2604

Wayne, L.G., Brenner, D.J., Colwell, R.R., Grimont, P.A.D., Kandler, O., Krichevsky, M.I., Moore, L.H., Moore, W.E.C., Murray, R.G.E., Stackenbrandt, E., Starr, M.P. and Trüper, H.G. (1987) Report of the ad hoc committee on reconciliation of approaches to bacterial systematics. *Int J System Bact* 37:463-464

Wertheim, H.F., Melles, D.C., Vos, M.C., van Leeuwen, W., van Belkum, A., Verbrugh, H.A. and Nouwen, J.L. (2005) The role of nasal carriage in *Staphylococcus aureus* infections. *Lancet Infect Dis* 5:751-762.

Whiley R.A., Hall, L.M., Hardie, J.M. and Beighton, D. (1999) A study of small-colony, beta-haemolytic, Lancefield group C streptococci within the anginosus group: description of Streptococcus constellatus subsp. pharyngis subsp. nov., associated with the human throat and pharyngitis. *Int J Syst Bacteriol* 49(**4**):1443-9

Wieland, S.C., Cassa, A.A., Mandl, K.D. and Berger, B. (2008) Revealing the spatial distribution of a disease while preserving privacy. *Proc Natl Acad Sci USA* 105(**46**):17608-17613

WHO (1999) Pneumococcal vaccines, World Health Organisation position paper. *Wkly Epidemiol Rec* 74(**23**):177-83

Wieczorek, J., Guo, Q. and Hijmans R. (2004) The point-radius method for georeferencing locality descriptions and calculating associated uncertainty. *Int J Geo Information Sci*. 12:746-67

Yother, J. (2004) Capsules, pp30-48. *In* E.I, Tuomanen (ed), The Pneumococcus. ASM Press, Washington, D.C. ISBN:155581297X

Zmasek, C.M., and Eddy, S.R. (2001) ATV: display and manipulation of annotated phylogenetic trees. *Bioinformatics* 17(**4**):383-4

Web references

1. http://www.insdc.org/ The International Nucleotide Sequence Database Collaboration.
2. http://www.w3.org/ The World Wide Web Consortium (W3C).
3. http://www.w3.org/Style/CSS/ W3C specification for Cascading Style Sheets.
4. http://www.w3.org/Protocols/ W3C Hypertext Transfer Protocol.
5. http://www.ebi.ac.uk/Tools/clustalw2/ ClustalW server at European Bioinformatics Institute.
6. http://www.iis.net/ Microsoft Internet Information Server (IIS).
7. http://httpd.apache.org/ The open source Apache HTTP Server.
8. http://www.ebi.ac.uk/Tools/blast2 Wahington University Basic Local Alignment Search Tool Version 2.0 (WU-BLAST 2.0); European Bioinformatics Institute.
9. http://www.w3.org/DOM/ W3C specifications for the Document Object Model.
10. http://www.w3.org/XML/ W3C Extensible Markup Language (XML) Working Group.
11. http://www.ecma-international.org/ European association for standardising information and communication systems – In particular, the general purpose, cross platform, vendor-neutral language ECMAScript.
12. http://tinyurl.com/7tcjz - Tim O'Reilly keynote speech (10/12/2006) What is Web 2.0? *O'Reilly Network*.
13. http://office.microsoft.com/excel Excel: A desktop spreadsheet application.
14. http://www.w3.org/TR/XMLHttpRequest/ W3C XMLHttpRequest Object
15. http://adaptivepath.com/ideas/essays/archives/000385.php Jesse James Garrett (18/02/2005) Ajax:A New Approach to Web Applications.
16. http://www.google.com/webhp?complete=1&hl=en Google Suggest. One of the first AJAX applications.
17. http://www.microsoft.com/Windows Microsoft Windows Operating System.
18. http://developer.apple.com/cocoa Apple's Mac OS X Operating System.
19. http://code.google.com/apis/maps/ Google Maps API.
20. http://aws.amazon.com Amazon Web Services homepage.

21. http://developer.yahoo.com/search/boss Yahoo Search BOSS (Build your own Search Service) API.
22. http://www.programmableweb.com Programmable Web – A comprehensive list of Web Mashups, APIs and details of the Web as a platform.
23. http://www.housingmaps.com HousingMaps – one of the first examples of a web Mashup.
24. http://www.craigslist.org CraigsList – Providing local classifieds and forums for jobs, housing, for sale, personals, local community and events in the USA.
25. http://ispecies.org iSpecies – A simple example using APIs of the potential of a webpage for every biological species as envisioned by Edward O Wilson.
26. http://en.wikipedia.org/wiki/API Wikipedia Search API
27. http://www.ncbi.nlm.nih.gov/entrez/query/static/esoap_help.html The NCBI Entrez Utilities Web service.
28. http://code.google.com/apis/ajaxsearch/ Google AJAX Search APIs.
29. http://www.healthmap.org HealthMap: global infectious disease monitoring.
30. http://www.genomesonline.org Comprehensive listing of microbial genome projects, including those published and those ongoing.
31. http://www.mmb.usyd.edu.au/BPGD/default.htm The Bacterial Polysaccharide Gene Database – University of Sydney, Australia.
32. http://www.glyco.ac.ru/bcsdb/ The Bacterial Carbohydrate Structure Database, N.D. Zelinsky Institute of Organic Chemistry, Moscow, Russia.
33. http://www.mlst.net MLST databases held at Imperial College London, London, UK
34. http://www.pubMLST.org MLST databases held at Oxford University, Oxford, UK
35. http://www.mlst.ucc.ie MLST databases held at University of Cork, Cork, Republic of Ireland,
36. http://www.pasteur.fr/recherche/genopole/PF8/mlst MLST databases held at The Pasteur Institute, Paris, France
37. http://sara.molbiol.ox.ac.uk/userweb/mchan/stars Sequence Typing Analysis and Retrieval System (STARS) Written by Man-Suen Chan and Nicki Ventress (2001).
38. http://www.hubmed.org An alternative interface to the PubMed medical literature database. Written by Alf Eaton.
39. http://www.pubmed.org The PubMed medical literature database.
40. http://www.phylowidget.org A Processing based tree viewer. Written by Greg Jordan (2008). Also see Jordan GE and Piel WH 2008
41. http://mail.google.com Gmail – Google's free online AJAX email facility
42. http://mail.yahoo.com Yahoo! Mail – free email service.
43. http://www.hotmail.com Microsofts free email service – Hotmail.
44. http://java.sun.com/javase/technologies/desktop/javawebstart/index.jsp Java Web Start Technology.
45. http://processing.org/ Processing: An open source visual programming language.
46. http://www.hpa.org.uk/web/HPAweb&HPAwebStandard/HPAweb_C/1195733825428 Norovirus Molecular Epidemiology Database. Health Protection Agency, London, UK
47. http://hiv.lanl.gov The Los Alomos HIV Sequence Database.
48. http://www.emi.unsw.edu.au/spolTools/ Spoligotyping molecular epidemiological database for the *Mycobacterium tuberculosis* complex

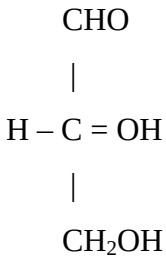
49. http://www.hpa-bioinfodatabases.org.uk/hepatitis_open/main.php The International Repository for Hepatitis B Virus Strain Data.
50. http://maps.google.com Google Maps
51. http://earth.google.com Google Earth
52. http://maps.yahoo.com Yahoo! Maps
53. http://www.microsoft.com/maps Microsoft Bing Maps.
54. http://www.openstreetmap.org/ Open Street Map.
55. http://code.google.com/apis/maps/ Google Maps API
56. https://www1.nga.mil/ National Geospatial-Intelligence Agency
57. http://www.winzip.com WinZip compression software.
58. http://code.google.com/apis/kml/ The Keyhole Markup Language (KML) specification.
59. http://www.prototypejs.org/ The Prototype JavaScript Framework
60. http://script.aculo.us/ The Script.aculo.us JavaScript Library
61. http://www.mapstraction.com The Mapstraction Javascript library provides a common API for various other JavaScript mapping APIs
62. http://www.w3.org/TR/soap/ The Simple Object Access Protocol specification provided by the W3C
63. http://pubmlst.org/api/ The pubMLST SOAP API.
64. http://pubmlst.org/neisseria The *Neisseria* spp. MLST database
65. http://www.rivm.nl/earss/ The European Antimicrobial Resistance Surveillance System Database
66. http://www.ridom.de/spa-server/ The *spa* typing Server.
67. http://www3.ridom.de/staphtype/ The *spa* StaphType software.
68. http://www.oie.int/ The World Organisation for Animal Health.
69. http://www.iucn.org/ The International Union for Conservation of Nature (IUCN).
70. http://www.iucnredlist.org/ The IUCN Red List of Threatened Species.
71. http://www.parcplace.org/bdmap.html The Global *Bd* Mapping project.
72. http://www.parcplace.org/bdmap2008update.html First release of maps from the Global *Bd* Mapping project.
73. http://www.thematicmapping.org/ The Thematicmapping web site.
74. http://www.sgrillo.net/ The Cocoa Research Centre.
75. http://code.google.com/apis/visualisation/ Google Visualisation API
76. http://www.lgcstandards-atcc.org/ The American Type Culture Collection (ATCC)- Laroratory of the Government Chemist (LGC) collaboration. LGCC is a nonprofit resource dedicated to the collection, preservation and distribution of authentic cultures of living microorganisms, viruses, DNA probes, plants and human and animal cells.
77. http://www.esri.com/software/arcgis/ ArcGIS A Complete Integrated System for geospatial data.
78. http://grass.itc.it/ GRASS, a free, open source Geographic Information System.
79. http://www.hiv.lanl.gov/components/sequence/HIV/geo/geo.comp The Los Alomos HIV Sequence database geographic search Interface.
80. http://www.r-project.org/ The R Project for Statistical Computing.
81. http://www.opengeospatial.org/ The Open Geospatial Consortium.
82. http://wms-sites.com/ A Collection of Web Mapping Service Sites.
83. http://www.w3.org/TR/html5/ W3C Wirking Draft for HTML 5.
84. http://tinyurl.com/koprh5 - Paul Krill from ComputerWorld on HTML 5
85. http://www.mapstraction.com/features.php Mapstraction feature coverage matrix.

Appendix 1: Polysaccharide Biochemistry

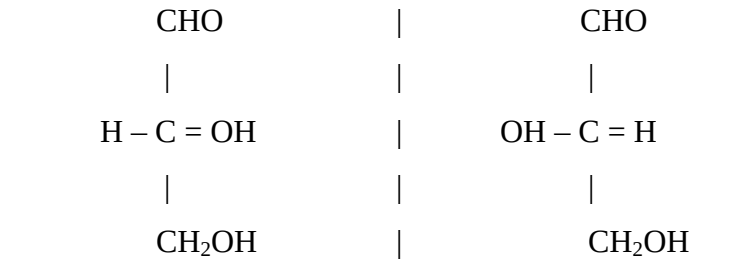
(Adapted from Essentials of Glycobiology, ED A.Varki CSHL Press NY.)

Monosaccharides are composed of an unbranched chain of carbon atoms linked together by single bonds and can occur in either linear or ring forms. All simple monosaccharides adhere to the general empirical formula $(CH_2O)_n$ where $_n$ is an integer between 3 and 9, and globally all monosaccharides can be grouped into one of two very general classes depending on functional group components. Aldoses contain a functional aldehyde group (CH=O) (eg glucose, mannose and galactose) and ketoses contain a functional ketone group (C=O) (eg fructose and ribulose).

One widely used abbreviated structural representation of individual monosaccharides is the 'Fischer projection formula', which can be used to describe the simplest aldose, Glyceraldehyde as follows –



Carbon atoms can by asymmetric (chiral) based on the phenomenon of non-superimposable mirror images of a molecule, with each different form of the molecule being termed an optical isomer. As an example the following Fischer projections show the two forms, termed enantiomers, of Glyceraldehyde



a) D-Glyceraldehyde b) L-Glyceraldehyde

The central C molecule in Glyceraldehyde contains four different groups, and the resulting non-superimposable enantiomers can be termed the D-form (Latin D-

Dexter) if the OH group is present on the right hand side of the chiral C atom as in a) and the L-form (Latin L – Laevus) if the OH group is found on the left.

With the exception of dihydroxy acetone, all monosaccharides have at least one asymmetric chiral center, the total number being equal to the number of internal (CHOH) groups.

Numbering of the carbon atoms in a monosaccharide follows a convention with Carbon 1 (C1) generally being a potential or actual aldehyde group within aldoses and the ketone group in ketoses is always C2.

Monosaccharides can exist in either linear or ring forms and in biological systems the ring form predominates due to the fact that they have lower energies than the linear form. Ring forms can be found in either five or six-sided ring structures due to the formation of intramolecular bonds, and 5-sided ring structures and termed furanoses due to their relation to the heterocyclic compound furan and 6-sided rings are termed pyranoses due to their relation to pyran.

In order for a ring structure to form, the linear structure has to bend resulting in a ring structure with a new OH group at the C1 position. This linking leads to the formation of another chiral centre at this position, and the OH goup can emerge at either side of the ring depending on the way in which the O atom of the C=O group is pointing just before ring closure. This therefore creates another two epimers, while the numbering of the C atoms is retained.

The two possible forms of the ring based on the position of the OH group are termed anomers and can be designated as either α, if the OH group is axial or β if the OH group is equatorial. Ring structures are generally depicted using the Haworth projection formula.

Fundamental to carbohydrate chemistry is the notion that monosaccharides can be linked to form disaccharides, (consisting of two monosaccharides), trisaccharides (consisting of three monosaccharides) and so on to form linear or branching chains.

Common protocol states that chains containing between 2 and 10 monosaccharides are termed oligosaccharides and polymers containing more than 10-20 monosaccharides are termed polysaccharides, even though the boundaries of the term are somewhat arbitrary, a polysaccharide is generally referred to as a molecule of unspecified length and repetitive composition. When describing polysaccharides, the names of monosaccharides are frequently abbreviated and the most commonly used are three-letter abbreviations for simple monosaccharides (e.g., Gal, Glc, Man, Xyl, Rib) and unless specified, the isomeric D- configuration is implied in these abbreviated names.

Appendix 2: Relevant publications

The following manuscripts are presented in their published form:

Bentley, S.D, Aanensen, D.M., Mavroidi, A., Saunders, D., Rabbinowitsch, E., Collins, M., Donohoe, K., Harris, D., Murphy, L., Quail, M.A., Samuel, G., Skovsted, I.C., Kaltoft, M.S., Barrell, B., Reeves, P.R., Parkhill, J. and Spratt, B.G. (2006) Genetic analysis of the capsular biosynthetic locus from all 90 pneumococcal serotypes. *PLoS Genet* 2(**3**):e31

Aanensen, D.M*., Mavroidi, A*., Bentley, S.D, Reeves, P.R. and Spratt, B.G. (2007) Predicted functions and linkage specificities of the products of the Streptococcus pneumoniae capsular biosynthetic loci. *J Bacteriol* 89(**21**):7856-76

Mavroidi, A*., Aanensen, D.M*., Godoy, D., Skovsted, I.C., Kaltoft, M.S., Reeves, P.R., Bentley, S.D. and Spratt, B.G. (2007) Genetic relatedness of the Streptococcus pneumoniae capsular biosynthetic loci. *J Bacteriol* 189:(**21**):7841-55

Aanensen, D.M. and Spratt, B.G. (2005) The multilocus sequence typing network: mlst.net *Nucleic Acids Res* 33:W729-33

Bishop, C.J*., Aanensen, D.M*., Jordan, G.E., Kilian, M., Hanage, W.P. and Spratt, B.G. (2009) Assigning strains to bacterial species via the internet. *BMC Biol* 7:3

Aanensen, D.M., Huntley, D.M., Feil, E.J., al-Own, F. and Spratt, B.G. (2009) EpiCollect: Linking Smartphones to Web Applications for Epidemiology, Ecology and Community Data Collection *PLoS ONE* 4(**9**):e6968

*Joint first authors

Genetic Analysis of the Capsular Biosynthetic Locus from All 90 Pneumococcal Serotypes

Stephen D. Bentley[1*], David M. Aanensen[2], Angeliki Mavroidi[2], David Saunders[1], Ester Rabbinowitsch[1], Matthew Collins[1], Kathy Donohoe[3], David Harris[1], Lee Murphy[1], Michael A. Quail[1], Gabby Samuel[3], Ian C. Skovsted[4], Margit Staum Kaltoft[4], Bart Barrell[1], Peter R. Reeves[3], Julian Parkhill[1], Brian G. Spratt[2]

1 Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, United Kingdom, **2** Department of Infectious Disease Epidemiology, Imperial College, London, United Kingdom, **3** School of Molecular and Microbial Biosciences, University of Sydney, Sydney, Australia, **4** Staten Serum Institut, Copenhagen, Denmark

Several major invasive bacterial pathogens are encapsulated. Expression of a polysaccharide capsule is essential for survival in the blood, and thus for virulence, but also is a target for host antibodies and the basis for effective vaccines. Encapsulated species typically exhibit antigenic variation and express one of a number of immunochemically distinct capsular polysaccharides that define serotypes. We provide the sequences of the capsular biosynthetic genes of all 90 serotypes of *Streptococcus pneumoniae* and relate these to the known polysaccharide structures and patterns of immunological reactivity of typing sera, thereby providing the most complete understanding of the genetics and origins of bacterial polysaccharide diversity, laying the foundations for molecular serotyping. This is the first time, to our knowledge, that a complete repertoire of capsular biosynthetic genes has been available, enabling a holistic analysis of a bacterial polysaccharide biosynthesis system. Remarkably, the total size of alternative coding DNA at this one locus exceeds 1.8 Mbp, almost equivalent to the entire *S. pneumoniae* chromosomal complement.

Citation: Bentley SD, Aanensen DM, Mavroidi A, Saunders D, Rabbinowitsch E, et al. (2006) Genetic analysis of the capsular biosynthetic locus from all 90 pneumococcal serotypes. PLoS Genet 2(3): e31.

Introduction

Streptococcus pneumoniae (the pneumococcus) is a major cause of morbidity and mortality worldwide, causing diseases that range in severity from meningitis, septicaemia, and pneumonia to sinusitis and acute otitis media [1,2]. Factor (typing) sera are used to divide pneumococci into serotypes and serogroups, which include immunologically related serotypes. These sera have been developed by a process of multiple cross-absorptions, which render them specific for the immunochemical differences between the pneumococcal capsular polysaccharides (CPSs) [3]. At present, 90 individual serotypes are recognised by their patterns of reactivity with the factor sera [4], and serotypes vary in the extent to which they are carried in the nasopharynx and the degree to which they are recovered from different disease states [5,6]. Expression of a capsule is important for survival in the blood and is strongly associated with the ability of pneumococci to cause invasive disease. The capsule is surface exposed, and antibodies against CPS provide protection against pneumococcal disease. Consequently, polyvalent polysaccharide vaccines have been developed in which CPS from the serotypes most commonly associated with invasive disease in children are linked to a protein carrier, and a seven-valent conjugated polysaccharide vaccine has been introduced and shown to be highly effective [7,8]. A 23-valent polysaccharide vaccine is also available for use in adults [9].

With the exception of types 3 and 37, which are synthesised by the synthase pathway [10–14], pneumococcal CPSs are generally synthesised by the Wzx/Wzy-dependent pathway (Figure 1). The genes for the latter pathway are located at the same chromosomal locus *(cps)*, between *dexB* and *aliA* [15–17]. CPSs are synthesised by transfer of an initial monosaccharide phosphate from a nucleotide diphosphate sugar to a membrane-associated lipid carrier, followed by the sequential transfer of further monosaccharides to produce the lipid-linked repeat unit. This is transferred to the outer face of the cytoplasmic membrane by the repeat-unit transporter or flippase, polymerised to form the mature CPS, and then attached to the peptidoglycan [18]. The *cps* locus therefore typically encodes the enzymes to build the repeat unit, including an initial glycosyl phosphate transferase, and additional transferases responsible for the formation of the linkages, and to allow for the addition of sugars (or other moieties), or to otherwise modify the repeat unit, as well as a repeat-unit flippase and polymerase [15].

The substantial diversity of pneumococcal CPSs is believed to have arisen as a consequence of selection for antigenic diversity imposed by the human immune system [6]. The evolutionary timescales and the genetic events by which novel serogroups and serotypes arise are unclear. Comparisons of the available *cps* loci indicate a variety of genetic mechanisms and show that the central genes responsible for the synthesis and polymerisation of the repeat unit are highly variable and often non-homologous between serotypes. These genes have a

Editor: Claire M. Fraser, The Institute for Genomic Research, United States of America

Received February 15, 2005; **Accepted** January 25, 2006; **Published** March 10, 2006

DOI: 10.1371/journal.pgen.0020031

Copyright: © 2006 Bentley et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Abbreviations: CPS, capsular polysaccharide; Galp, galactopyranose; GalpNAc, N-acetylgalactosamine pyranose; Glcp, glucopyranose; GlcpNAc, N-acetylglucosamine pyranose; HG[number], homology group[number]; IS, insertion sequence; KDQNAc, 4-keto-N-acetyl-D-quinovosamine

* To whom correspondence should be addressed. E-mail: sdb@sanger.ac.uk

Synopsis

Several major bacterial pathogens are coated by a polysaccharide capsule that is important for virulence. Each strain of *Streptococcus pneumoniae* (the pneumococcus) produces one of 90 different capsular polysaccharides, which are distinguished by using a set of antisera that recognise the chemical differences in the capsules. The capsule is important for virulence, but is immunogenic, and the large number of different capsular serotypes is believed to have been selected as a mechanism to evade the human immune response. Antibodies against capsular polysaccharide can protect against pneumococcal disease, and a highly effective protein-conjugated polysaccharide vaccine that protects children against the seven serotypes most commonly associated with serious disease has recently been introduced in the United States. Bentley and colleagues have determined the DNA sequence of the capsular biosynthesis genes for all 90 serotypes of *S. pneumoniae* and, by correlating the gene content with the serological profiles and with the known capsular polysaccharide structures, provide what they believe to be the most detailed and complete understanding of capsular biosynthesis and capsular gene evolution in any bacterial pathogen.

low percentage G+C content, and new serotypes may frequently have been generated by the introduction of novel *cps* genes into pneumococci by lateral gene transfer from other species. A much better understanding of the complex mechanisms by which antigenic diversity arises could be obtained by using the sequences of the complete set of pneumococcal *cps* loci. We therefore obtained sequences of the *cps* locus for all 90 serotypes and used these data, together with the available polysaccharide structures and the patterns of serological reactions with typing sera, to explore the genetics of capsular diversity in this major pathogen. Here we present highlights of our analysis to date, and a more exhaustive analysis will be reported elsewhere.

Results

General Features of the *dexB–aliA* Locus from 90 Serotypes

PCR products were generated from genomic DNA using primers specific for the *dexB* and *aliA* genes and ranged in size from 10,337 bp (serotype 3) to 30,298 bp (serotype 38) with an average of 20,714 bp. The synthase gene *(wchE)* of serotype 3 is located within the *cps* locus, but the type 37 *cps* locus, which was very similar to that of serotype 33F, is defective and serotype is determined by the type 37 synthase gene *(tts)* located elsewhere on the chromosome [10]. Annotation and analysis of the *cps* sequences revealed the generality of several previously observed characteristics. Genes for the generation of CPSs are always orientated in the same direction as the *dexB* and *aliA* genes (Figures 2 and S1). The regulatory and processing genes *wzg, wzh, wzd,* and *wze* (also known as *cpsABCD*) are conserved with high sequence identity in all cases and are almost always in this gene order at the 5′ end of the *cps* locus. In most *cps* clusters, the fifth gene encodes the initial glucose phosphate transferase, WchA (also known as CpsE), responsible for linkage of an activated glucose phosphate to the lipid carrier (see below). The polysaccharide polymerase *(wzy)* and flippase *(wzx)* genes are always present downstream together with a varying set of genes for glycosyl transferases, acetyl transferases, nucleotide diphosphate sugar biosynthesis, and modifying enzymes.

In every case, there is a region of low percentage G+C content within the *cps* locus. The first four genes and the non-housekeeping sugar biosynthesis genes have typical percentage G+C content for *S. pneumoniae,* while the "serotype-specific" genes, particularly *wzy* and *wzx,* tend to have more AT-rich sequences. In the regions between the *cps* genes and the flanking *dexB* and *aliA* genes, there is almost always evidence of mobile genetic elements. This is largely manifested as intact or disrupted genes for insertion-sequence (IS) transposases [19,20], although in four cases we identified group-II introns [21] (serotypes 19F, 25F, 25A, and 38).

We could assign a functional designation to the products of all but 26 of the 1,999 predicted coding sequences in the 90 *cps* regions, with most of the remainder showing weak similarities to products of genes in bacterial polysaccharide gene clusters. Unsurprisingly, many coding sequences fall into the broad functional categories of glycosyl transferase (351), acetyl transferase (74), and sugar phosphate transferase (71). To make more specific assignments within such categories, we used the TribeMCL program to assemble all the annotated proteins into homology groups (HGs). With from two to 90 members in each, 91% of the proteins assembled into 175 HGs, with the remainder forming 74 single-member HGs (Table S1). The products of *wzg, wzh, wzd,* and *wze* each fall into a single HG covering every serotype. Ignoring IS element transposases, the next largest HG comprises 65 WchA initial transferases (HG5). At the other extreme, the serotype-specific gene products are diverse, with 87 HGs for non-initial sugar transferases and 40 and 13 groups of Wzy repeat-unit polymerases and Wzx flippases, respectively.

Biosynthesis of Precursors for Sugars and Other CPS Components

Of the 18 sugars and related compounds found in *S. pneumoniae* capsules, seven are available from housekeeping metabolic pathways and nine from known dedicated pathways encoded within the *cps* cluster (Figure S2). This includes 4-keto-*N*-acetyl-D-quinovosamine (UDP-KDQNAc), which is the intermediate in the two step reaction catalysed by FnlA [22]. We found a perfect correlation between the presence of a non-housekeeping sugar in the CPS and the presence of the appropriate biosynthetic genes in the *cps* locus. Two of the three remaining compounds are the sugar alcohol phosphates arabinitol-1-P and mannitol-6-P. The precursors for these have not been identified, but nucleotide-diphosphate-linked precursors can be easily derived from D-xylulose-5-phosphate or D-fructose-6-phosphate, respectively, by two-step pathways parallel to that for CDP-ribitol formation from ribulose-5-phosphate [23]. D-xylulose-5-phosphate and D-fructose-6-phosphate are central to major pathways, and there are appropriate genes for their conversion in the associated *cps* loci. The precursor for ribofuranose has also not been identified, but a proposed pathway for its biosynthesis by the product of a gene within CPS 19F *(cps19R)* [24] is supported by our observation that an orthologous gene (renamed *rbsF*) is present for all CPS that contain ribofuranose.

Choline-1-phosphate, glycerol-1-phosphate, and glycerol-2-phosphate are also found in some of the structures. CDP-choline is known to be produced by *S. pneumoniae* as a precursor for teichoic acid biosynthesis [25]. For glycerol-1-phosphate, we find an intact *gct* gene for CDP-glycerol synthesis [26] in the *cps* where expected, and there are four genes associated with presence of glycerol-2-phosphate, three

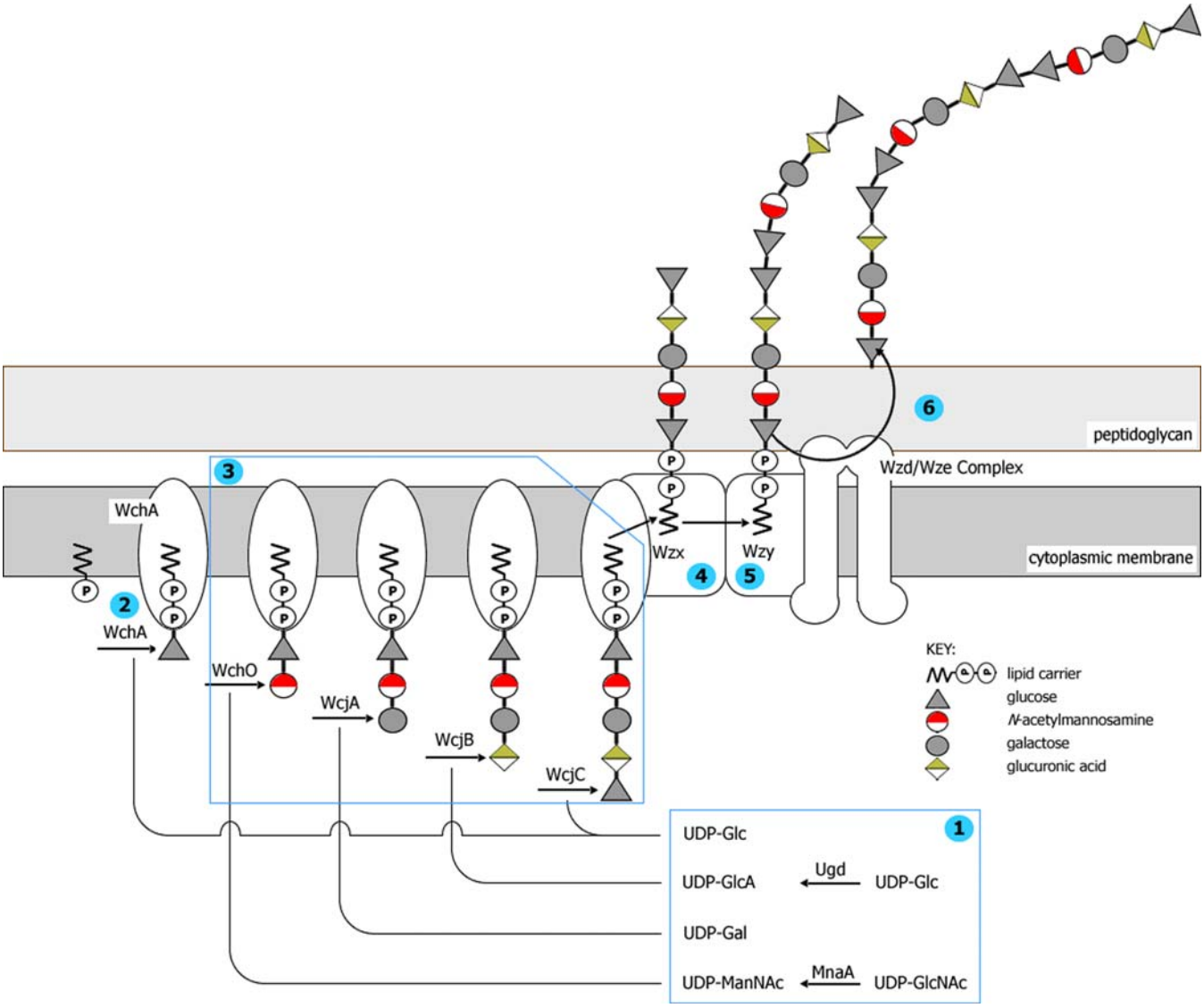
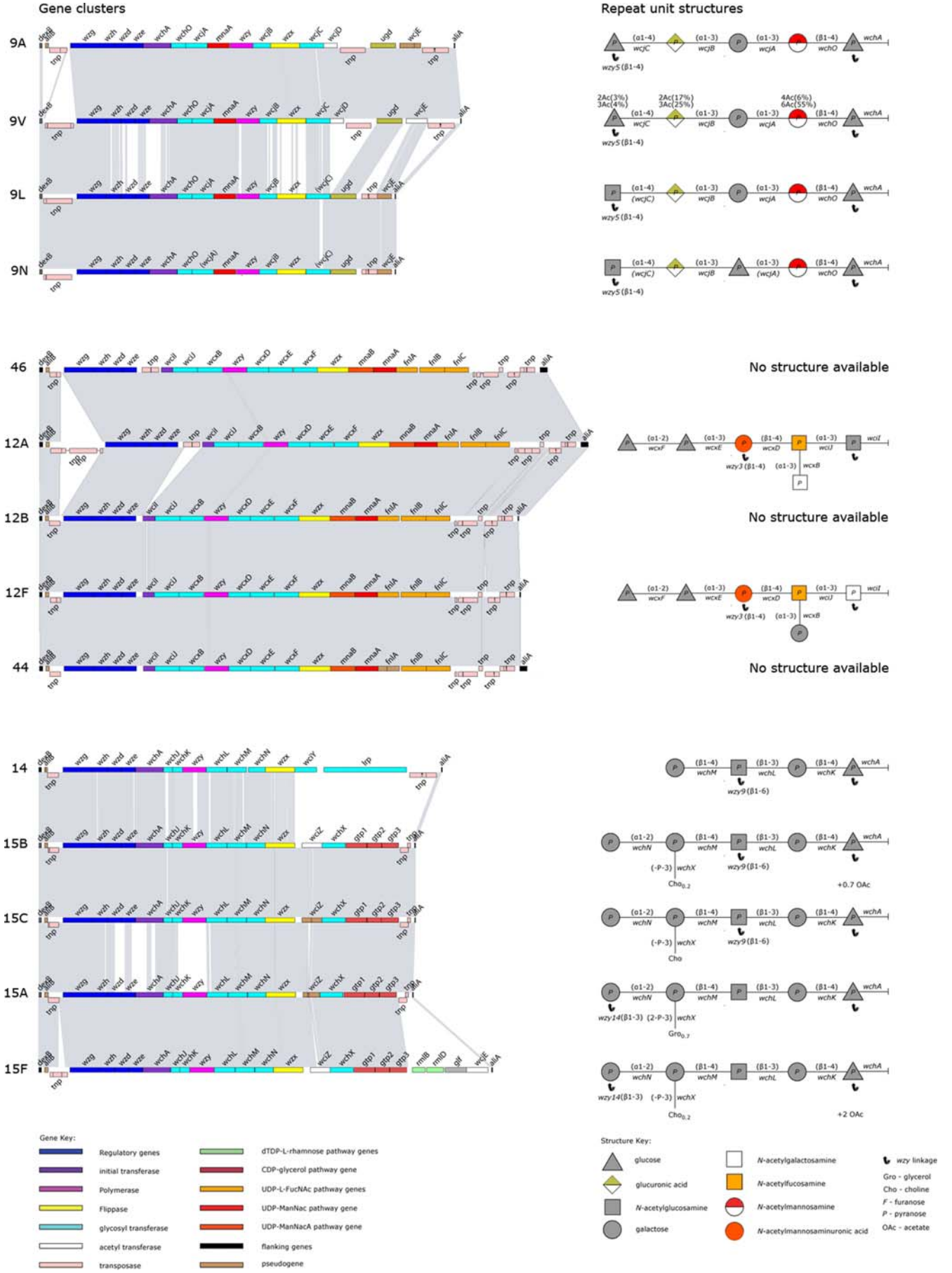


Figure 1. Representation of the Wzx/Wzy-Dependent Pathway for Biosynthesis of CPS 9A

Pictured is a hypothetical model for capsule biosynthesis in *S. pneumoniae* based on a mixture of experimental evidence and speculation. For a recent review, see Yother [15].
(1) Non-housekeeping nucleotide sugar biosynthesis.
(2) The initial transferase (WchA in this case) links the initial sugar as a sugar phosphate (Glc-P) to a membrane-associated lipid carrier (widely assumed to be undecaprenyl phosphate).
(3) Glycosyl transferases sequentially link further sugars to generate repeat unit.
(4) Wzx flippase transports the repeat unit across the cytoplasmic membrane.
(5) Wzy polymerase links individual repeat units to form lipid-linked CPS.
(6) Wzd/Wze complex translocates mature CPS to the cell surface and may be responsible for the attachment to peptidoglycan. The complex of WchA, Wzy, Wzx, Wzd, and Wze shown in the membrane is based on that in Figure 2 of Whitfield and Paiment [47] for the related *Escherichia coli* Type 1 capsule.
Found at DOI: 10.1371/journal.pgen.0020031.g001

Figure 2. Capsule Biosynthesis Genes and Repeat-Unit Polysaccharide Structures

Shown are the *cps* gene clusters for cases discussed in the text, together with the polysaccharide structure of the encoded repeat unit where known [31] (the full set is shown in Figure S1). Genes are represented on the forward and reverse strands by boxes coloured according to the gene key, with gene designations indicated above each box. Grey blocks indicate regions of sequence similarity between gene clusters. Repeat-unit structures are displayed with the linkage to undecaprenyl pyrophosphate at the right-hand side (not necessarily the case for the published structures [31]), so residue numbers are counted from right to left. Monosaccharides are represented as shapes coloured according to the structure key. Housekeeping sugars are coloured grey. Non-housekeeping sugar colours correspond to the associated sugar biosynthesis gene colours. Glycerol, choline, and acetate are indicated as text. Also shown are the nature of linkages with the associated gene, and the linkages between repeat units created by the Wzy polymerase. Gene designations are in parentheses where their substrate specificity is unclear.
Found at DOI: 10.1371/journal.pgen.0020031.g002


Gene clusters
Repeat unit structures
9A
9V
9L
9N
46
12A
12B
12F
44
14
15B
15C
15A
15F
No structure available
No structure available
No structure available
Gene Key:
Regulatory genes
initial transferase
Polymerase
Flippase
glycosyl transferase
acetyl transferase
transposase
dTDP-L-rhamnose pathway genes
CDP-glycerol pathway gene
UDP-L-FucNAc pathway genes
UDP-ManNac pathway gene
UDP-ManNacA pathway gene
flanking genes
pseudogene
Structure Key:
glucose
glucuronic acid
N-acetylglucosamine
galactose
N-acetylgalactosamine
N-acetylfucosamine
N-acetylmannosamine
N-acetylmannosaminuronic acid
wzy linkage
Gro - glycerol
Cho - choline
F - furanose
P - pyranose
OAc - acetate

of which are thought to encode a CDP-2-glycerol pathway [27], while *wchX* encodes the glycerol phosphotransferase.

The situation is illustrated in Figure 1 for *cps*9A, which has pathway genes for *N*-acetylmannosamine pyranose and glucuronic acid, but not for glucopyranose (Glc*p*) or galactopyranose (Gal*p*) as these are available in *S. pneumoniae* from central metabolism.

Initial Transferases and Polymerisation

Initial transferase WchA adds glucose-1-phosphate to undecaprenol phosphate [28] to create Und-PP-Glc (Figure 1), and we assume it performs that function in all 65 serotypes where it is present. For the known structures, there is a perfect correlation between the presence/absence of *wchA* and the presence/absence of glucose in the repeating unit. Where *wchA* is absent, the products of the fifth *cps* gene fall into three HGs (WciI, WcjG, and WcjH) all with the same Pfam [29] domain and similar hydrophobicity profiles to the carboxy-terminal region of WchA. We suggest that they function as the initial sugar transferases, as it is known that for the *Salmonella enterica wchA* homologue, *wbaP,* the 3′ end of the gene is sufficient for transferase activity [30]. By correlation with CPS constituents, we predict the transferred initial sugars as *N*-acetylgalactosamine pyranose (Gal*p*NAc) or *N*-acetylglucosamine pyranose (Glc*p*NAc) for WciI, Gal*p* or galactofuranose for WcjG and Gal*p* for WcjH. Serotype 1 is an exception as no gene product with similarity to an initial sugar transferase has yet been identified.

The initial sugar of the repeat unit is also the donor sugar in the polymerisation of the repeat units (Figure 1), and the specificity of the Wzy polymerase determines the other component of this linkage, which in the case of CPS 9A is a beta (1–4) linkage to the terminal glucose of the next repeat unit. For the known structures [31], identification of the initial sugar allowed us to determine the polymerase linkage as both donor and acceptor sugar, and the linkages were defined once the initial sugar had been identified (see Figures 2 and S1). Where there is ambiguity due to two residues of the initial sugar in the repeat unit, the polymerase linkage can be provisionally identified by considering the linkage catalysed by other members of the same Wzy HG. The predictions for initial sugars, and subsequent repeat-unit polymerisation linkage, correlate well with the polymerase HGs (Table S2). There are 32 polymerase HGs associated with WchA, five with WciI, four with WcjG and one with WcjH. These associations are mostly exclusive, with only five polymerase HGs associated with two initial transferases. In such cases, the linkages involve the same acceptor sugar anomerism (α or β isomer) and the same or a closely related donor sugar. This adds strong support to the inferences drawn for the specificity of the initial transferases.

Relating *cps* Genes with CPS Structure and Serological Profile

The availability of all of the annotated *cps* sequences allowed us to look for correlations between genes, known CPS structures, and serology (gene clusters, CPS structures, and antigenic formulae are summarised in Figure S1 and Table S3). In this way, we can attempt both to infer gene function and, by comparing related *cps* loci, to account for differences in CPS structure and serology. Variations between *cps* loci range from two base substitutions for 18B and 18C to wholesale differences in gene complement. Within this range, the variations likely to have a phenotypic effect include gene inactivation due to single base substitutions generating a premature stop codon, single base insertion/deletions leading to translational frameshifts, change of sequence leading to change of enzyme specificity, recombination or IS element insertion leading to gene truncation, and insertion/deletion/replacement of single and multiple genes. Within serogroups, the genetic differences were often subtle but were also sometimes surprisingly prominent. Comparisons also revealed some strong commonality between the *cps* of different serogroups and serotypes. Illustrative examples that demonstrate how structure, genetics, and serology were combined to analyse the *cps* loci are shown in Figure 2 and are discussed below.

Serogroup 9

Previously described CPS structures [31] for all four serotypes of serogroup 9 show only subtle differences and provide an example of multiple serotypes arising by divergence from a single *cps* locus. Their *cps* genes fall into two pairs, with 9A highly similar to 9V [32], and 9L highly similar to 9N, but with the two pairs differing significantly in sequence (Figure 2), suggesting an initial divergence to form two ancestral serotypes; this split correlates with a difference at residue 5 of the repeat unit, where 9L and 9N CPSs have Glc*p*NAc, whereas 9A and 9V have Glc*p*. Factor sera 9d reacts with 9A and 9V but not with 9L and 9N, suggesting that it is interacting with Glc*p* but not with Glc*p*NAc. Both are housekeeping sugars, and their differential incorporation is likely to be due to divergent forms of glycosyl transferase WcjC. Subsequently, one of these ancestral serotypes diverged to form 9L and 9N, the latter becoming unique in the group in having Glc*p* rather than Gal*p* as residue 3 in the repeat unit. Their *dexB–aliA* loci have the same gene complement, and within the *cps* genes there are only 79 nucleotide differences. The highest number of amino acid substitutions (13) is within glycosyl transferase WcjA; ten are unique to 9N and presumably result in its altered specificity for Glc*p* rather than for Gal*p*.

The other ancestral serotype gave rise to 9V and 9A, which differ from each other only in their CPS acetylation; the former CPS has an *O*-acetylation pattern unique in the serogroup. This is likely due to the *O*-acetyl transferase–encoding *wcjE* gene, which is intact and apparently functional in 9V, disrupted by a frameshift mutation in 9A (deletion of guanine, nucleotide 726), and truncated in 9L and 9N by the insertion of an IS element. Interestingly, factor sera 9g reacts only with serotype 9V and may recognise an acetyl-based epitope determined by *wcjE*.

Serogroup 9 *cps* loci also differ by the insertion, in 9A and 9V relative to 9L and 9N, of an *O*-acetyl transferase gene *(wcjD)* and an adjacent IS element. This correlates with recent nuclear magnetic resonance data (I. C. Skovsted, unpublished data), indicating that 9A CPS is partially acetylated.

Serotypes 44 and 46 Are Related to Serogroup 12

The *cps* gene clusters of serogroup 12 and serotypes 44 and 46 are almost identical, differing only in IS transposase genes, and provide an example of common ancestry that is not apparent from serology. Structures have been determined for serotypes 12F and 12A only, although the individual constituents for serotype-46 CPS are known and all are present in 12F and 12A [31]. Although no factor serum cross-reacts with all five serotypes, serological reactions do indicate antigenic commonalities [4]; 44 cross-reacts with factor sera 12b and 12d, while 46 cross-reacts with 12c. Given the *cps* similarities, the significant differences between 12F and 12A CPS are perhaps surprising; 12A has a Gal*p*NAc and 12F has a

Gal*p* side branch, and the first main-chain residue is Gal*p*NAc in 12F and Glc*p*NAc in 12A. The nucleotide differences are concentrated within two glycosyl transferase genes *(wciI* and *wcxB),* and we predict that the initial transferases, WciI–12A and WciI–12F, with 38 amino acid differences, link Glc*p*NAc and Gal*p*NAc, respectively, to the lipid carrier, while WcxB–12A and WcxB–12F, with 17 amino acid differences, account for the side-branch difference.

Serotype 14 Is Closely Related to Serogroup 15

Serotype 14 shares no significant serological cross-reaction with serogroup 15, or with any other serotype, but the *cps* loci of these two serotypes are clearly related. All CPS structures for serotype 14 and serogroup 15 are known [31,33,34], and comparisons of structures and genes allow inferences about one to be made from the other. The four serogroup 15 pentasaccharide repeat units are identical, but polymerisation forms a linear polymer in 15A and 15F and a branched structure in 15B and 15C that correlates with the presence of *wzy* genes of different HGs (see Table S2). Serotypes 15B and 15C differ in the presence or absence of *O*-acetylation [35] and, as previously described [36], the difference is due to a variable-length TA tandem repeat region at the 5′ end of *wciZ*—in frame in 15B and out of frame in 15C strains. This gene is in frame in 15F (acetylated), but extensively degraded, rather than simply out of frame, in 15A (not acetylated).

Genes for synthesis of glycerol-2-phosphate *(gtp1, gtp2,* and *gtp3)* are present in all serogroup 15 *cps* loci, but glycerol was reported to be present only in 15A, being replaced by choline-P in 15F, 15B, and 15C, with either residue being present on only a proportion of the repeat units [31]. In all cases, the transferase is presumed to be encoded by *wchX*, with the molecular basis of the structural polymorphism being contentious. However, recent nuclear magnetic resonance analysis indicates that 15B contains glycerol and not choline, suggesting that the same may also be true for 15F and 15C [34]. The 3′ end of 15F *cps* has four extra genes—*rmlB, rmlD, glf,* and a putative acetyl transferase gene *wcjE*—but they appear to have no effect on the structure as there is no rhamnose, galactofuranose, or extra acetylation in 15F CPS. Indeed, *rmlA* and *rmlC* would also be required for rhamnose biosynthesis. These four genes show synteny with the 3′ end of *cps* in several serotypes, particularly serotype 31, and their arrangement in 15F may indicate a recombination event.

The serotype 14 [28,37,38] and basic 15 *cps* gene clusters clearly share common ancestry and differ only at the 3′ end, where the glycerol-2-phosphate–related genes in 15 are replaced in 14 by a gene *(lrp)* encoding a large (1,359 amino acid) repetitive protein, which correlates well with CPS structures [36]. The type-14 repeat unit most resembles the branched form of 15B and 15C, with the lack of *O*-acetylation due to the absence of *wciZ*. The lack of α-D-galactose is probably due to degradation of the relevant transferase gene, *wchN*. The large repetitive protein encoded by serotype-14 *cps* has a hydrophobic C-terminal region, suggesting that it may be anchored to the cell surface. This leads us to speculate that Lrp may serve as a dominant antigen that overwhelms the serological similarities to serogroup 15 that should be evident from their very similar repeat units.

Discussion

Several bacterial pathogens exist as a large number of antigenic variants because of differences in the polysaccharides presented at the cell surface. However, the sequencing and analysis of the *cps* loci of pneumococci described here are believed to provide the only such case where the whole gene repertoire is available, allowing genetics, chemistry, and immunology to be combined to predict the role of *cps* genes. This combined approach has allowed the confident prediction of most gene functions, but it has also highlighted the limitations where subtle sequence changes may alter enzyme substrate specificity. Analysis of the *cps* loci indicates that a number of different mechanisms have generated antigenic diversity in CPSs. Some of these involve the divergence of a single serotype into two related serotypes by the accumulation of point mutations (e.g., serogroup 6 [39]), or the insertion or deletion of a single gene, resulting in slightly different CPS structures (e.g., serogroup 18). In other cases, the *cps* loci of some serotypes within a serogroup seem to be virtually unrelated and probably reflect the sharing of a dominant epitope that led to them being placed within the same serogroup (e.g., serogroups 7, 17, 33, and 35). Similarly, some serotypes placed in different serogroups show more relatedness among their *cps* loci than those within the same serogroup (e.g., types 7B and 7C are more closely related to type 40 than to 7A and 7F). This is perhaps not surprising as serogroups were defined by common epitopes in the absence of any knowledge of the CPS structures or the *cps* sequences that code for their synthesis. Shared immunodominant epitopes will lead to inclusion in the same serogroup even if there are major differences in other parts of the structure and hence in the *cps*.

A striking feature of the *cps* loci is the presence of many highly divergent forms of each of the key enzyme classes. Thus, there are 40 HGs for polysaccharide polymerases, 13 groups of flippases, and a great diversity of transferases. The presence of multiple non-homologous or highly divergent forms of these enzymes, together with the low percentage G+C content of the region in which these are encoded, supports the view that these genes have been imported into pneumococci (or their ancestors) on multiple occasions from different and unknown sources. The plethora of transferases in the pneumococcal *cps* loci provides an opportunity to continually generate new serotypes by gene shuffling, but there are no clear examples of serotypes arising as mosaics of two existing *cps* loci. One barrier to the frequent appearance of new serotypes by recombination is a lack of homology between the serotype-specific regions of *cps* loci of different serogroups. The appearance of new serotypes may also be limited by a need to change multiple *cps* genes; rare genetic events that create mosaics between existing *cps* loci probably typically fail to produce a capsule since new repeat units resulting from the capture of novel transferases are unlikely to be recognised as substrates by the resident repeat-unit polymerase. The *cps* sequences, and their associated polysaccharide structures and serological profiles, constitute an extensive dataset that, through further detailed analysis, will allow a clearer understanding of capsule biochemistry, genetics, and evolution and will precipitate advances in molecular serotyping of pneumococci [40,41].

Materials and Methods

Strain selection, serotyping, and genomic DNA isolation. Representative strains of the 90 *S. pneumoniae* serotypes were selected from among the lyophilised strains in the strain collection of the World Health Organization Collaborating Centre for Reference and

Research on Pneumococci, Statens Serum Institut (Copenhagen, Denmark) (Table S4). The strains were serotyped and cultured, and genomic DNA was extracted by standard methods [3,4,42].

PCR and DNA sequencing. PCR reactions were performed using the Expand Long Template PCR System (Roche, Basel, Switzerland), which contains proof-reading thermostable polymerases. Initial reactions used primers CPS1 (TTGCCAATGAAGAGCAAGACTTGA CAGTAG) and CPS2 (CAATAATGTCACGCCCGCAAGGGCAAGT) [26]. Where these failed to produce an adequate product, further reactions were attempted using alternative *dexB*-specific primers (CPS1A [CGACCGTCGCTTCCTAGTTGTGGCTAAC] or PCPS3f [CACACAGAAAGCATCCCATGG]) and *aliA*-specific primers (CPS1B [GTCTTGAGCTTTGACTGCCGCGTATTCT] or PCPS3r [GAGACA GACCTGATAACCTCAACTATTTG]). The *cps* cluster for our serotype-5 strain was amplified using a primer based on the EMBL file (AY336008) specific for the *wzg* gene (CPS05F [CGTTCACAGAAAGT GAAGCG]) in combination with PCPS3r. PCR products spanning the *cps* locus were used directly to construct small-insert libraries [43], with 1- to 2-kb inserts in pUC18. Clones from each library were sequenced from each end using Big-Dye terminator chemistry (Applied Biosystems, Foster City, California, United States) on ABI3730 sequencing machines, to give an average of 8- to 10-fold coverage of each product. These reads were assembled with Phrap (CodonCode, Dedham, Massachusetts, United States), and any gaps or regions of poor coverage were re-sequenced using primer-directed sequencing directly from the original PCR product using Big-Dye primer chemistry (Applied Biosystems). This sequencing procedure should prevent any PCR errors from being represented in the final consensus sequence.

Annotation and bioinformatic methods. Gene prediction and annotation were performed as previously described [44]. Predicted proteins were clustered into homology groups using TribeMCL (Centre for Mathematics and Computer Science and EMBL-EBI) [45] with a cut-off of $1e^{-50}$. The genes within the *cps* loci that encoded proteins within the same homology group were assigned the same name, the exceptions being the polymerases and flippases where we used the prior gene nomenclature, *wzy* and *wzx,* even though in both cases there were multiple homology groups. Alignment of gene clusters was performed using the Artemis Comparison Tool (Sanger Institute, Hinxton, United Kingdom). Nucleotide differences were identified using the EMBOSS program Diffseq (MRC Rosalind Franklin Centre for Genomics Research, Hinxton, United Kingdom) [46].

Supporting Information

Figure S1. Capsule Biosynthesis Genes and Repeat-Unit Polysaccharide Structure for All 90 Serotypes

Found at DOI: 10.1371/journal.pgen.0020031.sg001 (9.9 MB TIF).

Figure S2. Biosynthesis Pathways for Non-Housekeeping Sugars

Found at DOI: 10.1371/journal.pgen.0020031.sg002 (50 KB PPT).

Table S1. Homology Groups including Numbers of Members and Product Description

Proteins in different homology groups are so divergent that they are highly unlikely to have diverged from a common streptococcal ancestor.

Found at DOI: 10.1371/journal.pgen.0020031.st001 (306 KB DOC).

Table S2. Associations between Initial Transferases and Wzy Polymerase Groups

Proposed Wzy groupings represent a sequential numbering of homology groups and are represented on structural diagrams.

Found at DOI: 10.1371/journal.pgen.0020031.st002 (104 KB DOC).

Table S3. Type Designations and Antigenic Formulae for the 90 Serotypes of *S. pneumoniae*

The antigenic formulae represent arbitrary designations of cross-reactions as seen by the capsular reaction.

Found at DOI: 10.1371/journal.pgen.0020031.st003 (72 KB DOC).

Table S4. Type and Strain Designations for the 90 Strains of *S. pneumoniae* Analysed

Found at DOI: 10.1371/journal.pgen.0020031.st004 (68 KB DOC).

Accession Numbers

The EMBL Nucleotide Sequence Database (http://www.ebi.ac.uk/embl), GenBank (http://www.ncbi.nlm.nih.gov/Genbank), and DNA Data Bank of Japan (http://www.ddbj.nig.ac.jp/Welcome-e-html) accession numbers for the sequences reported in this paper for the capsular biosynthetic genes of the 90 serotypes of *S. pneumoniae* are CR931632–CR931722. The EMBL Nucleotide Sequence Database (http://www.ebi.ac.uk/embl) accession number for the *wzg* gene is AY336008. The Pfam domain (http://www.sanger.ac.uk/cgi-bin/Pfam) for WchA, WciI, WcjG, and WcjH is PF02397.

Acknowledgments

We thank Renato Morona and James Baddiley for useful discussions and Fanrong Kong for highlighting anomalies in some of our initial sequences. We also acknowledge the use of core facilities at the Wellcome Trust Sanger Institute.

Author contributions. BB, JP, and BGS conceived and designed the experiments. DS, ER, MC, DH, LM, and MAQ performed the experiments. SDB, DMA, AM, KD, GS, ICS, MSK, PRR, JP, and BGS analyzed the data. SDB, DMA, AM, PRR, and BGS wrote the paper.

Funding. This work was supported by a grant from the World Health Organization, and we acknowledge the support and encouragement of Thomas Cherian. BGS acknowledges receipt of a Wellcome Trust Principal Research Fellowship.

Competing interests. The authors have declared that no competing interests exist. ■

References

1. Austrian R (1999) The pneumococcus at the millennium: Not down, not out. J Infect Dis 179 (Suppl 2): S338–S341.
2. Cartwright K (2002) Pneumococcal disease in Western Europe: Burden of disease, antibiotic resistance and management. Eur J Pediatr 161: 188–195.
3. Lund E, Henrichsen J (1978) Laboratory diagnosis, serology and epidemiology of *Streptococcus pneumoniae*. In: Bergan T, Norris J, editors. Methods in microbiology. London: Academic Press. pp. 241–262.
4. Henrichsen J (1995) Six newly recognized types of *Streptococcus pneumoniae*. J Clin Microbiol 33: 2759–2762.
5. Hausdorff WP, Bryant J, Paradiso PR, Siber GR (2000) Which pneumococcal serogroups cause the most invasive disease: Implications for conjugate vaccine formulation and use, part I. Clin Infect Dis 30: 100–121.
6. Spratt BG, Hanage WP, Bruegemann AB (2004) Evolutionary and population biology of *Streptococcus pneumoniae*. In: Tuomanen EI, editor. The pneumococcus. Washington (D. C.): ASM Press. pp. 119–135.
7. Black S, Shinefield H, Fireman B, Lewis E, Ray P, et al. (2000) Efficacy, safety and immunogenicity of heptavalent pneumococcal conjugate vaccine in children. Northern California Kaiser Permanente Vaccine Study Center Group. Pediatr Infect Dis J 19: 187–195.
8. Reinert RR (2004) Pneumococcal conjugate vaccines—A European perspective. Int J Med Microbiol 294: 277–294.
9. Poolman JT (2004) Pneumococcal vaccine development. Expert Rev Vaccines 3: 597–604.
10. Llull D, Munoz R, Lopez R, Garcia E (1999) A single gene *(tts)* located outside the cap locus directs the formation of *Streptococcus pneumoniae* type 37 capsular polysaccharide. Type 37 pneumococci are natural, genetically binary strains. J Exp Med 190: 241–251.
11. Cartee RT, Forsee WT, Jensen JW, Yother J (2001) Expression of the *Streptococcus pneumoniae* type 3 synthase in *Escherichia coli*. Assembly of type 3 polysaccharide on a lipid primer. J Biol Chem 276: 48831–48839.
12. Dillard JP, Vandersea MW, Yother J (1995) Characterization of the cassette containing genes for type 3 capsular polysaccharide biosynthesis in *Streptococcus pneumoniae*. J Exp Med 181: 973–983.
13. Arrecubieta C, Lopez R, Garcia E (1994) Molecular characterization of cap3A, a gene from the operon required for the synthesis of the capsule of *Streptococcus pneumoniae* type 3: Sequencing of mutations responsible for the unencapsulated phenotype and localization of the capsular cluster on the pneumococcal chromosome. J Bacteriol 176: 6375–6383.
14. Waite RD, Penfold DW, Struthers JK, Dowson CG (2003) Spontaneous sequence duplications within capsule genes *cap8E* and *tts* control phase variation in *Streptococcus pneumoniae* serotypes 8 and 37. Microbiology 149: 497–504.
15. Yother J (2004) Capsule. In: Tuomanen EI, editor. The pneumococcus. Washington (D. C.): ASM Press. pp. 30–48.
16. Kolkman MA, van der Zeijst BA, Nuijten PJ (1998) Diversity of capsular polysaccharide synthesis gene clusters in *Streptococcus pneumoniae*. J Biochem (Tokyo) 123: 937–945.
17. Garcia E, Llull D, Munoz R, Mollerach M, Lopez R (2000) Current trends in capsular polysaccharide biosynthesis of *Streptococcus pneumoniae*. Res Microbiol 151: 429–435.
18. Sorensen UB, Henrichsen J, Chen HC, Szu SC (1990) Covalent linkage

between the capsular polysaccharide and the cell wall peptidoglycan of *Streptococcus pneumoniae* revealed by immunochemical methods. Microb Pathog 8: 325–334.
19. Iannelli F, Pearce BJ, Pozzi G (1999) The type 2 capsule locus of *Streptococcus pneumoniae*. J Bacteriol 180: 1381–1388.
20. Sanchez-Beato AR, Garcia E, Lopez R, Garcia JL (1997) Identification and characterization of IS1381, a new insertion sequence in *Streptococcus pneumoniae*. J Bacteriol 179: 2459–2463.
21. Lambowitz AM, Zimmerly S (2004) Mobile group II introns. Annu Rev Genet 38: 1–35.
22. Mulrooney EF, Poon KKH, McNally DJ, Brisson JR, Lam JS (2005) Biosynthesis of UDP-*N*-acetyl-L-fucosamine, a precursor to the biosynthesis of lipopolysaccharide in *Pseudomonas aeruginosa* serotype O11. J Biol Chem 280: 19535–19542.
23. Pereira MP, Brown ED (2004) Bifunctional catalysis by CDP-ribitol synthase: Convergent recruitment of reductase and cytidylyltransferase activities in *Haemophilus influenzae* and *Staphylococcus aureus*. Biochemistry 43: 11802–11812.
24. Morona JK, Morona R, Paton JC (1999) Comparative genetics of capsular polysaccharide biosynthesis in *Streptococcus pneumoniae* types belonging to serogroup 19. J Bacteriol 181: 5355–5364.
25. Bean B, Tomasz A (1977) Choline metabolism in pneumococci. J Bacteriol 130: 571–574.
26. Jiang SM, Wang L, Reeves PR (2001) Molecular characterization of *Streptococcus pneumoniae* type 4, 6B, 8, and 18C capsular polysaccharide gene clusters. Infect Immun 69: 1244–1255.
27. Morona JK, Miller DC, Coffey TJ, Vindurampulle CJ, Spratt BG, et al. (1999) Molecular and genetic characterization of the capsule biosynthesis locus of *Streptococcus pneumoniae* type 23F. Microbiology 145: 781–789.
28. Kolkman MA, Wakarchuk W, Nuijten PJ, van der Zeijst BA (1997) Capsular polysaccharide synthesis in *Streptococcus pneumoniae* serotype 14: Molecular analysis of the complete *cps* locus and identification of genes encoding glycosyltransferases required for the biosynthesis of the tetrasaccharide subunit. Mol Microbiol 26: 197–208.
29. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, et al. (2004) The Pfam protein families database. Nucleic Acids Res 32: D138–D141.
30. Wang L, Liu D, Reeves PR (1996) C-terminal half of *Salmonella enterica* WbaP (RfbP) is the galactosyl-1-phosphate transferase domain catalyzing the first step of *O*-antigen synthesis. J Bacteriol 178: 2598–2604.
31. Kamerling JP (2000) Pneumococcal polysaccharides: A chemical view. In: Tomasz A, editor. Streptococcus pneumoniae: Molecular biology and mechanisms of disease. Larchmont (New York): Mary Ann Liebert. pp. 81–114.
32. van Selm S, Kolkman MA, van der Zeijst BA, Zwaagstra KA, Gaastra W, et al. (2002) Organization and characterization of the capsule biosynthesis locus of *Streptococcus pneumoniae* serotype 9V. Microbiology 148: 1747–1755.
33. Jansson PE, Lindberg B, Lindquist U, Ljungberg J (1987) Structural studies of the capsular polysaccharide from *Streptococcus pneumoniae* types 15B and 15C. Carbohydr Res 162: 111–116.
34. Jones C, Lemercinier X (2005) Full NMR assignment and revised structure for the capsular polysaccharide from *Streptococcus pneumoniae* type 15B. Carbohydr Res 340: 403–409.
35. Venkateswaran PS, Stanton N, Austrian R (1983) Type variation of strains of *Streptococcus pneumoniae* in capsular serogroup 15. J Infect Dis 147: 1041–1054.
36. van Selm S, van Cann LM, Kolkman MA, van der Zeijst BA, van Putten JP (2003) Genetic basis for the structural difference between *Streptococcus pneumoniae* serotype 15B and 15C capsular polysaccharides. Infect Immun 71: 6192–6198.
37. Kolkman MA, Morrison DA, van der Zeijst BA, Nuijten PJ (1996) The capsule polysaccharide synthesis locus of *Streptococcus pneumoniae* serotype 14: Identification of the glycosyl transferase gene *cps14E*. J Bacteriol 178: 3736–3741.
38. Kolkman MA, van der Zeijst BA, Nuijten PJ (1997) Functional analysis of glycosyltransferases encoded by the capsular polysaccharide biosynthesis locus of *Streptococcus pneumoniae* serotype 14. J Biol Chem 272: 19502–19508.
39. Mavroidi A, Godoy D, Aanensen DM, Robinson DA, Hollingshead SK, et al. (2004) Evolutionary genetics of the capsular locus of serogroup 6 pneumococci. J Bacteriol 186: 8181–8192.
40. Brito DA, Ramirez M, de Lencastre H (2003) Serotyping *Streptococcus pneumoniae* by multiplex PCR. J Clin Microbiol 41: 2378–2384.
41. Lawrence ER, Griffiths DB, Martin SA, George RC, Hall LM (2003) Evaluation of semiautomated multiplex PCR assay for determination of *Streptococcus pneumoniae* serotypes and serogroups. J Clin Microbiol 41: 601–607.
42. Sorensen UB (1993) Typing of pneumococci by using 12 pooled antisera. J Clin Microbiol 31: 2097–2100.
43. McMurray AA, Sulston JE, Quail MA (1998) Short-insert libraries as a method of problem solving in genome sequencing. Genome Res 8: 562–566.
44. Holden MT, Titball RW, Peacock SJ, Cerdeno-Tarraga AM, Atkins T, et al. (2004) Genomic plasticity of the causative agent of melioidosis, *Burkholderia pseudomallei*. Proc Natl Acad Sci U S A 101: 14240–14245.
45. Enright AJ, Van Dongen S, Ouzounis CA (2002) An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res 30: 1575–1584.
46. Rice P, Longden I, Bleasby A (2000) EMBOSS: The European Molecular Biology Open Software Suite. Trends Genet 16: 276–277.
47. Whitfield C, Paiment A (2003) Biosynthesis and assembly of Group 1 capsular polysaccharides in *Escherichia coli* and related extracellular polysaccharides in other bacteria. Carbohydr Res 338: 2491–2502.

JOURNAL OF BACTERIOLOGY, Nov. 2007, p. 7856–7876 Vol. 189, No. 21
0021-9193/07/$08.00+0 doi:10.1128/JB.00837-07
Copyright © 2007, American Society for Microbiology. All Rights Reserved.

Predicted Functions and Linkage Specificities of the Products of the *Streptococcus pneumoniae* Capsular Biosynthetic Loci▽†

David M. Aanensen,[1]‡ Angeliki Mavroidi,[1]‡ Stephen D. Bentley,[2]
Peter R. Reeves,[3] and Brian G. Spratt[1]*

Department of Infectious Disease Epidemiology, Imperial College London, London, United Kingdom[1]; Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, United Kingdom[2]; and School of Molecular and Microbial Biosciences, University of Sydney, Sydney, Australia[3]

Received 30 May 2007/Accepted 21 August 2007

The sequences of the capsular biosynthetic (*cps*) loci of 90 serotypes of *Streptococcus pneumoniae* have recently been determined. Bioinformatic procedures were used to predict the general functions of 1,973 of the 1,999 gene products and to identify proteins within the same homology group, Pfam family, and CAZy glycosyltransferase family. Correlating *cps* gene content with the 54 known capsular polysaccharide (CPS) structures provided tentative assignments of the specific functions of the different homology groups of each functional class (regulatory proteins, enzymes for synthesis of CPS constituents, polymerases, flippases, initial sugar transferases, glycosyltransferases [GTs], phosphotransferases, acetyltransferases, and pyruvyltransferases). Assignment of the glycosidic linkages catalyzed by the 342 GTs (92 homology groups) is problematic, but tentative assignments could be made by using this large set of *cps* loci and CPS structures to correlate the presence of particular GTs with specific glycosidic linkages, by correlating inverting or retaining linkages in CPS repeat units with the inverting or retaining mechanisms of the GTs predicted from their CAZy family membership, and by comparing the CPS structures of serotypes that have very similar *cps* gene contents. These large-scale comparisons between structure and gene content assigned the linkages catalyzed by 72% of the GTs, and all linkages were assigned in 32 of the serotypes with known repeat unit structures. Clear examples where very similar initial sugar transferases or glycosyltransferases catalyze different linkages in different serotypes were also identified. These assignments should provide a stimulus for biochemical studies to evaluate the reactions that are proposed.

Streptococcus pneumoniae capsular polysaccharide (CPS) is produced by almost all isolates recovered from cases of invasive disease and is a major virulence factor (84). It is also a major immunogen, and antibodies directed against the CPS provide protective immunity. Consequently, CPS is the basis of all current pneumococcal vaccines. The serotyping scheme for *S. pneumoniae* is based on the immunochemical differences between CPS expressed by different strains, which are recognized by their reactivity with a set of typing sera. Currently, 91 different *S. pneumoniae* serotypes are recognized (31, 64), some of which show sufficient immunological similarities to be considered related and are placed within a serogroup.

The synthesis and export of bacterial polysaccharides are mediated by three pathways known as the Wzy-dependent pathway, the synthase-dependent pathway, and the ABC transporter-dependent pathway, which have been reviewed recently (77, 86). The Wzy-dependent pathway is most common in *S. pneumoniae* capsular biosynthesis, but the synthase-dependent pathway is also found (12, 34, 49, 90). The ABC transporter-dependent pathway has not been described for *S. pneumoniae* capsules.

The two known synthase-dependent pathways are present in serotypes 3 and 37 (7, 12, 48). The synthase-dependent pathway is the simplest, with a single integral inner membrane protein generally thought to be responsible for both the sequential addition of the sugars and the concurrent extrusion of the nascent polymer across the cell membrane. The repeat units with synthase-dependent pathways are generally single sugars, or two sugars may alternate, or one may form the main chain and the other a side branch on each main chain residue.

The Wzy-dependent pathway has been identified in 88 *S. pneumoniae* serotypes, and the *cps* loci are located at the same site on the chromosome, known as the CPS synthesis (*cps*) locus, between the *dexB* and *aliA* genes (49). The Wzy-dependent pathway involves the synthesis of the polysaccharide repeat unit on an undecaprenyl lipid carrier (UndPP) and then translocation of this oligosaccharide across the membrane by the Wzx flippase so that the oligosaccharide component, still attached to UndPP, is on the outer face of the membrane, where it is polymerized by Wzy to produce a UndPP-attached polymer (Fig. 1) (90).

The sequences of the *cps* loci from 90 *S. pneumoniae* serotypes have been presented recently, and the products of the genes present in these *cps* loci have been classified into 249 groups (homology groups [HGs]) (12). The great diversity of these products reflects the chemical diversity of the *S. pneumoniae* capsules, and we have presented some examples of predictions of functions by correlating gene content with known CPS structure and serology (12). Recently, we have used similarities in the sequence-weighted gene contents of the *S. pneumoniae cps* loci to identify closely related *cps* loci and have related these to their similar CPS repeat units (51). Here an analysis of the predicted functions of

* Corresponding author. Mailing address: Department of Infectious Disease Epidemiology, Imperial College London, Room G22, Old Medical School Building, St. Mary's Hospital, Norfolk Place, London W2 1PG, United Kingdom. Phone: 44 (0)20 7594 3398. Fax: 44 (0)20 7402 3927. E-mail: b.spratt@imperial.ac.uk.

† Supplemental material for this article may be found at http://jb.asm.org/.

‡ D.M.A. and A.M. contributed equally to this work.

▽ Published ahead of print on 31 August 2007.

Downloaded from jb.asm.org at Imperial College London on July 10, 2009

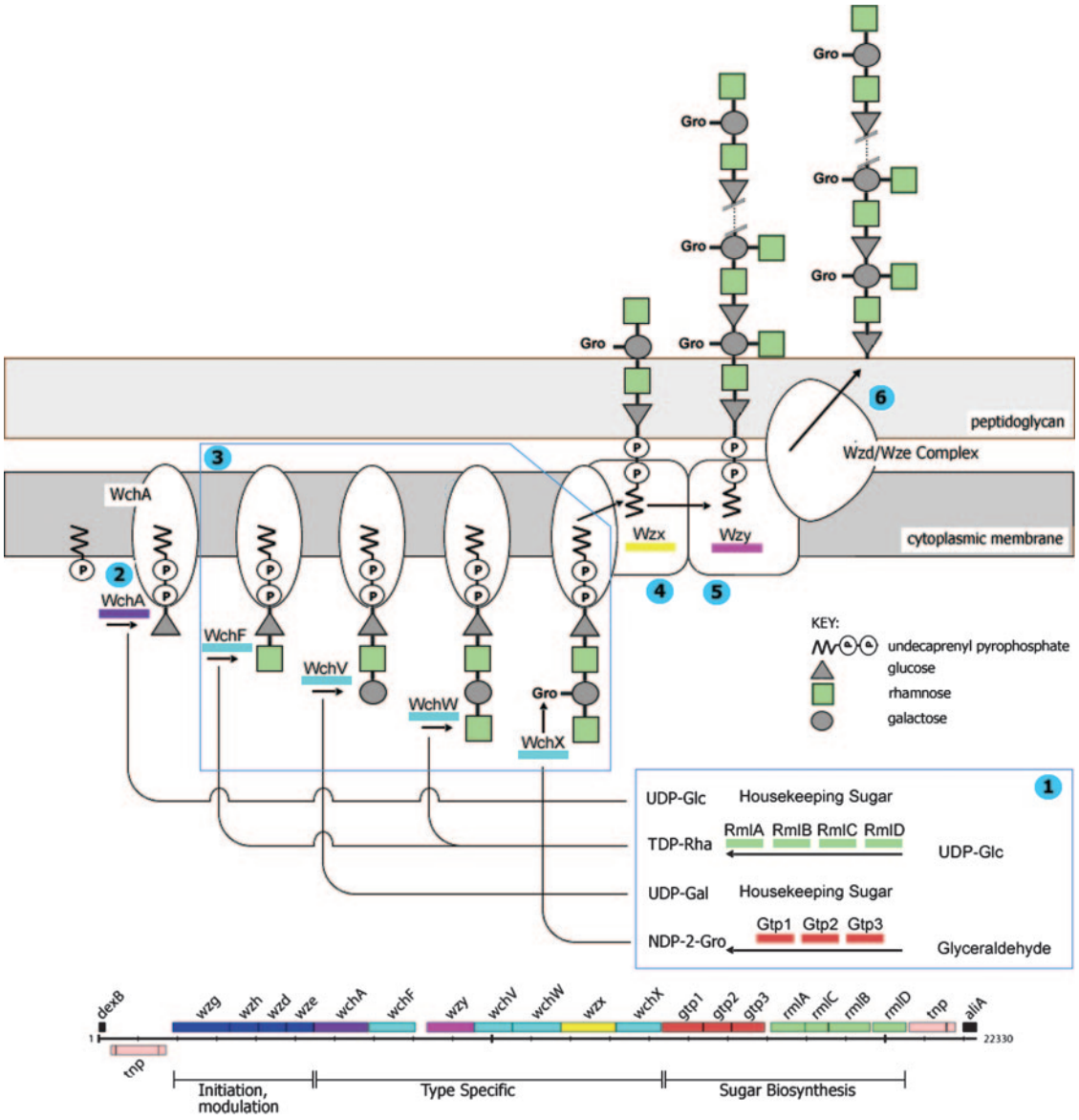


FIG. 1. Schematic representation of the biosynthesis of CPS by the Wzy-dependent pathway. The biosynthesis of the CPS of serotype 23F is represented. UDP-linked components of the repeat CPS unit are synthesized by genes encoded within the *cps* locus or are available from central metabolism (1). Repeat unit biosynthesis is initiated by the transfer of glucose phosphate to the lipid carrier by the IT WchA (2), followed by sequential addition of the other components of the repeat unit, catalyzed by the GTs (3), and the lipid-linked repeat unit is transferred across the membrane by the Wzx flippase (4) and polymerized by Wzy to result in lipid-linked CPS (5). Finally, the lipid-linked CPS is linked to the cell wall by a poorly understood process involving the Wzd/Wze complex, with release of the undecaprenyl phosphate carrier (6). The genes in the representation of the *cps* locus and their gene products are color coded; a key is given for the repeat unit constituents (modified from reference 12).

almost all of the 249 HGs encoded by the *cps* genes of the 90 serotypes is presented, along with the predicted specificity of the transferases that synthesize the CPS repeat units.

MATERIALS AND METHODS

Functional assignment of *cps* gene products. The sequences of the *S. pneumoniae cps* loci from all 90 serotypes (GenBank accession numbers CR931632 to CR931722) and the schematic representations of the available CPS structures were those of Bentley et al. (12), incorporating the revised CPS structures of serotypes 15B, 17F, and 33F (36, 37, 45). Recently, an immunological variant within serogroup 6 that has a different CPS structure (serotype 6C) has been reported, but as the *cps* sequence is not available, we do not consider it further (64). The predicted *cps* gene products have been placed into HGs, i.e., nonoverlapping groups of proteins showing highly significant amino acid sequence similarity, by using TRIBE-MCL with a TBLASTX cutoff level of $1e^{-50}$ (12). Genes encoding members of different HGs were assigned different gene names, excepting the polymerase (*wzy*) and flippase (*wzx*) genes. The polymerases and flippases fell into multiple HGs but, to maintain prior nomenclature, were each given the same gene name and were sequentially numbered to denote the members of the different HGs (12). New names for *cps* genes were assigned in accordance with the Bacterial Polysaccharide Gene Database (68). Amino acid sequence alignments were performed by using CLUSTAL_X (76). BlastN and TBLASTX comparisons of the *cps* loci were visualized by using ACT version 6 and WebACT, an online version of the Artemis comparison tool (1, 17).

To predict a functional assignment, the *cps* gene products were classified into

Downloaded from jb.asm.org at Imperial College London on July 10, 2009

families based on hidden Markov model profiles by using the Pfam database (25; http://www.sanger.ac.uk/Software/Pfam/). If there was no Pfam hit for an HG, the GenBank database (3; http://www.ncbi.nlm.nih.gov/) and the InterPro database (5; http://www.ebi.ac.uk/interpro/) were used for searches of homologs with known function in other organisms. The MetaCyc database (18; http://metacyc.org/) was used for the identification of genes involved in biosynthetic pathways. The genomic sequences of *S. pneumoniae* R6 (32; GenBank accession number AE007317) and *S. pneumoniae* TIGR4 (75; GenBank accession number AE005672) were used for the identification of genes involved in the biosynthesis of housekeeping components of the repeat units. The families of glycosyltransferases (GTs) encoded by the *S. pneumoniae cps* loci were retrieved from the carbohydrate-active enzymes (CAZy) database (20; http://afmb.cnrs-mrs.fr/CAZY/). There were 26 *cps* gene products that did not possess any known Pfam domain and were annotated as "conserved hypothetical proteins" if they showed weak similarity to products involved in bacterial polysaccharide biosynthesis or else as "hypothetical proteins" (12). Of these 26 *cps* gene products, 12 were present in serotypes with known CPS structures and their function could be predicted; 8 were present in serotypes with unknown CPS structures, and the rest were pseudogenes.

Assignment of reactions catalyzed by transferases. The approaches used to identify and assign functions to the sugar/polyalcohol phosphate transferases and pyruvyltransferases and for identification of polymerase linkages and initial sugar transferase specificities are described in Results. A number of approaches were used to tentatively assign the linkages catalyzed by the GTs, as follows.

(i) GT-linkage correlation across serotypes. Initially, all known repeat unit structures were broken down into the triplet linkages (donor sugar-linkage-acceptor sugar) catalyzed by the GTs, and correlations between the presence of a particular GT gene with the triplet linkages present in the repeat units were made. A perfect correlation between the presence of a GT gene in the *cps* loci and a single triplet linkage in the CPS of the corresponding structures, that was not found in structures where the gene was absent, corresponded to a strong assignment. This procedure was started with the GT gene (and thus the HG) that was most often present within the *cps* loci of the serotypes whose CPS structures are known. Having assigned the linkage catalyzed by the most common GT, we repeated the process with the GT gene that is next most common in the *cps* loci of serotypes with known CPS structures, ignoring all incidences of GT linkages that have already been assigned by linkage correlation. This iterative process was continued to assign the linkages catalyzed by further GTs, eliminating at each cycle the linkages already assigned, until assignments became ambiguous.

(ii) Gene content-linkage comparison between similar serotypes. Comparisons between gene content in very similar *cps* loci and their CPS structures can also be used to assign linkages. Should two *cps* loci differ only in a single GT gene and the CPS structures of those serotypes differ only in a single linkage, this can be used to assign enzyme specificity.

(iii) CAZy GT family membership and linkage status. GTs were assigned membership, where possible, to either inverting or retaining GT families based on the CAZy database (20). Each linkage in the repeat unit is catalyzed by a GT that either retains or inverts the stereochemistry of the donor sugar. Correlating the number of inverting or retaining GTs encoded in the *cps* loci with the number of inverting or retaining linkages in the CPS structures can in some cases be used to assign specificity. For example, the *cps* locus of a serotype may encode a single inverting GT, as predicted by CAZy family membership, and the CPS structure contains a single linkage that requires an inverting GT. We can therefore assign the specificity of the inverting GT to the synthesis of the inverting linkage.

(iv) Combination of the methods described above. In practice, the methods described above were often used in combination to assign GT specificities. For example, the *cps* locus of a serotype may possess three GT genes. One of these may have already been assigned by GT-linkage correlation and another by CAZy membership and linkage status. The third GT can then be assigned to the remaining linkage.

RESULTS AND DISCUSSION

***S. pneumoniae* capsules synthesized by the synthase-dependent pathway.** The type 37 synthase (Tts) directs the synthesis of serotype 37 CPS with a glucose main chain and an additional glucose side branch in each repeat unit (47, 48). The synthase gene of serotype 37 is not located within the *cps* locus, but a defective type 33F *cps* locus is present between *dexB* and *aliA* (12, 48). In the serotype 37 strain sequenced by Bentley et al. (12), the defective type 33F *cps* locus has frameshift mutations in the *wciC* and *wciG* genes and a stop codon in *glf*, compared to the *cps* locus of a serotype 33F isolate (GenBank accession no. CR931702).

The type 3 synthase (WchE; also known as Cap3B) directs the synthesis of serotype 3 CPS with glucose and glucuronic acid alternating (7, 82). WchE transfers glucose (Glc) from UDP-Glc to phosphatidylglycerol to form glucosephoshatidylglycerol, which serves as a primer for the addition of Glc and glucuronic acid (GlcA) (15, 26), and UDP-sugar substrate concentrations modulate the polysaccharide chain length (80). The type 3 synthase gene is within a defective *cps* locus along with three genes (*ugd*, *galU*, and *pgm*, also known as *cap3A*, *cap3C*, and *cap3D*, respectively) involved in the synthesis of sugar precursors (6, 12).

***S. pneumoniae* capsules synthesized by the Wzy-dependent pathway.** Excluding putative transposase genes, there are 1,502 *cps* gene products encoded by the *cps* loci of the 88 *S. pneumoniae* serotypes whose capsules are synthesized by the Wzy-dependent pathway (Table 1). As mentioned previously, the *cps* loci of the 88 pneumococcal serotypes that use the Wzy pathway include the polymerase (*wzy*) and flippase (*wzx*) genes, and although their gene products fall into a number of different HGs, they can be readily identified by homology searches and by their characteristic hydrophobicity patterns, providing a useful predictor of the presence of the Wzy-dependent pathway (12).

There are also four conserved genes (*wzg*, *wzh*, *wzd*, and *wze*) at the 5′ end of all *S. pneumoniae cps* loci that use the Wzy pathway (see the next section) (12, 34, 49, 90). The *cps* loci also include genes whose products are involved in the biosynthesis of nonhousekeeping components (*cps*-specific biosynthetic pathway genes; see below), initiation of capsule biosynthesis (initial sugar transferase genes), and transfer of sugars or other moieties and their assembly in the repeat unit (GT, acetyltransferase, sugar phosphate transferase, and pyruvyltransferase genes) (Fig. 1) (12).

Processing, regulation, and export of *S. pneumoniae* capsules. The highly conserved *wzg*, *wzh*, *wzd*, and *wze* genes (also known as *cpsA*, *cpsB*, *cpsC*, and *cpsD*, respectively) at the 5′ end of the *cps* locus are known to be involved in the processing, regulation, and export of CPS (56). These four genes constitute a regulatory system for CPS production via tyrosine phosphorylation of Wze (CpsD) (10, 56, 59), and recent findings suggest that this mechanism is important for attachment of the CPS to the cell wall (58).

Biosynthetic pathways for precursors of *S. pneumoniae* capsules. The precursors for the known components of the repeat units are all thought to be transferred from nucleotide diphospho (NDP) derivatives, with the exception of O-acetyl and pyruvate moieties, which are transferred from acetyl coenzyme A and phosphoenolpyruvate, respectively (Fig. 2 and 3). We consider first those CPS components for which NDP derivatives are generally present in *S. pneumoniae* because they are used for other pathways (housekeeping components); as expected, genes for these pathways are not present in the *cps* loci. Thirty-six of the pneumococcal CPS structures are unknown, and there could be additional housekeeping components besides those mentioned below. Genes should be present within *cps* loci for the biosynthesis, from available intermediates, of activated forms of the sugars or other CPS components that are not normally present in *S. pneumoniae* (nonhousekeeping or

Downloaded from jb.asm.org at Imperial College London on July 10, 2009

TABLE 1. Gene products, classified into functional groups, and their Pfam families, encoded by the 88 *cps* loci of *S. pneumoniae* that use the Wzy-dependent pathway of CPS biosynthesis[a]

Functional group	No. of genes (pseudogenes)	Pfam family (or families)[b]
Processing, regulation, export		
Wzg	88 (0)	(PF02916-PF03816)
Wzh	88 (0)	PF02811
Wzd	88 (0)	PF02706
Wze	88 (0)	
Polymerization and export		
Wzy	90 (2)	PF04932
Wzx	88 (0)	PF01943, PF03023
Biosynthetic pathways		
RmlA	40 (0)	PF00483
RmlC	40 (1)	PF00908 (PF01073-PF01370-PF02719-PF04321)
RmlB	41 (0)	(PF01073-PF01370-PF04321)
RmlD	41 (1)	(PF01073-PF01370-PF02719)
Glf	58 (23)	PF03275 (PF00984-PF01210-PF03720-PF03721)
Ugd	17 (0)	(PF00984-PF03720)
Gla	4 (0)	(PF01073-PF01370-PF02719-PF04321)
MnaA	15 (0)	PF02350
MnaB	5 (0)	(PF00984-PF03720-PF03721)
FnlA	8 (1)	(PF01073-PF01370-PF02719-PF04321-PF08485)
FnlB	8 (0)	PF01370
FnlC	8 (0)	PF02350
Gct	12 (2)	PF01467
Abp1	6 (1)	PF01128 (PF01073-PF01370-PF02719-PF04321)
Abp2	6 (0)	(PF01073-PF01370)
Mnp1	3 (0)	(PF00483-PF01128)
Mnp2	3 (0)	(PF01073-PF01370)
Gtp1	9 (0)	PF01761
Gtp2	9 (0)	PF00483
Gtp3	9 (0)	PF00702
RbsF	7 (0)	PF00702
Transferases		
ITs	88 (1)	PF02397
GTs	342 (8)	PF00534, PF00535, PF04488, PF03808, PF02485, PF05704, PF01501 (PF00535-PF05704)
Acetyltransferases	78 (12)	PF00132 PF01757
Sugar phosphate transferases	69 (3)	PF04991, PF04464, PF01066
Pyruvyltransferases	2 (0)	PF04230
Others		
Group II intron proteins	4 (0)	(PF00078-PF08388)
Long repetitive protein	1 (0)	(PF00746-PF05738)
GT enhancers	13 (0)	
Hypothetical proteins	26 (7)	
Total	1,502 (62)	

[a] Predicted *S. pneumoniae cps* gene products for 88 serotypes whose CPS is synthesized by the Wzy-dependent pathway. Products of transposase genes ($n = 224$) are not included.

[b] Pfam domains found in the *cps* gene products are shown. Different Pfam domains are found in different HGs of a functional group, and some contain more than one Pfam domain, as indicated by parentheses.

CPS-specific components). If some of the eight *cps* gene products without assigned functions (see Materials and Methods) are involved in biosynthetic pathways, there could also be additional nonhousekeeping components.

(i) Housekeeping components. There are no gene products present in the *S. pneumoniae cps* loci for the biosynthesis of seven components in the known CPS structures (glucopyranose, *N*-acetylglucosamine, galactopyranose, *N*-acetylgalactosamine, 2-acetamido-4-amino-2,4,6-trideoxy-D-galactopyranose, ribitol-phosphate and phosphorylcholine), and their precursors are presumably available for CPS synthesis from *S. pneumoniae* general metabolism. The first four are known to have UDP derivatives as precursors, whereas CDP-D-ribitol-5P and CDP-choline are the precursors for ribitol-phosphate and phosphorylcholine, respectively.

The genes expected for the synthesis of six of these components are present in the *S. pneumoniae* genome, and the biosynthetic pathways are well documented (68, 70), as shown in Fig. 2. There are also genes for a putative pathway for the other component, UDP-2-acetamido-4-amino-2,4,6-trideoxy-

Downloaded from jb.asm.org at Imperial College London on July 10, 2009

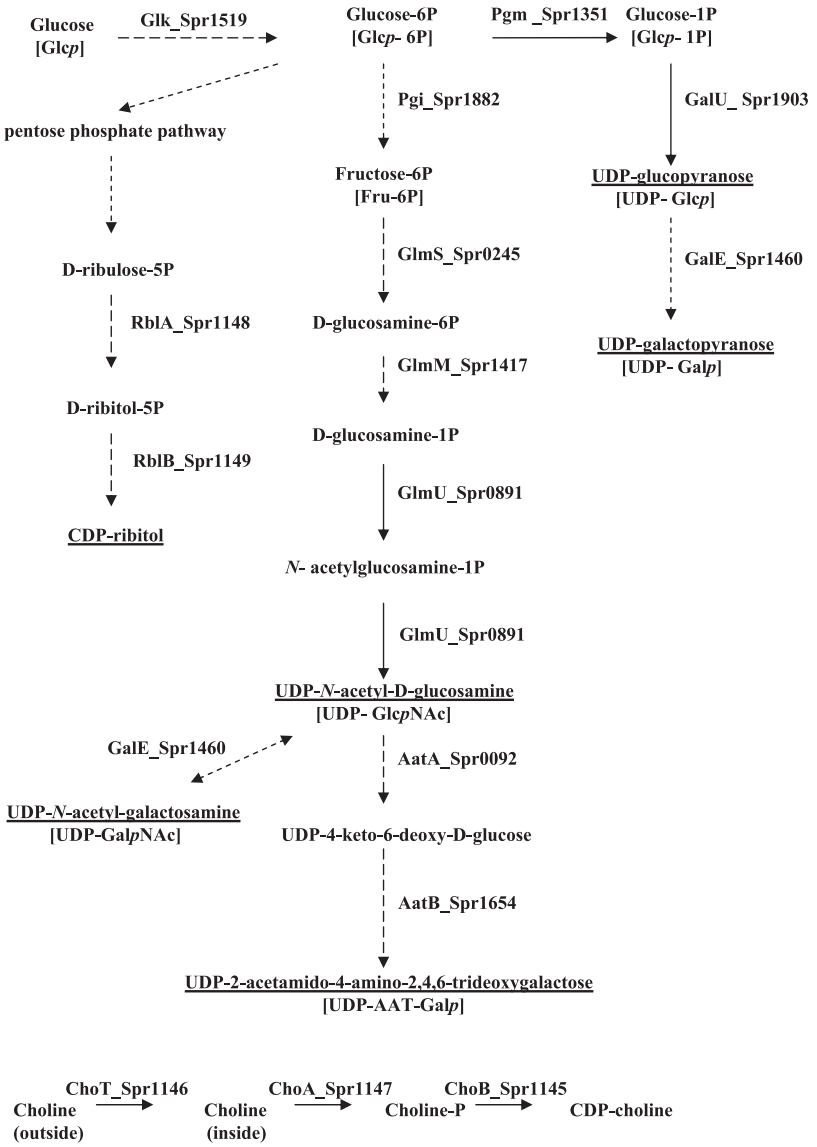


FIG. 2. Biosynthetic pathways of housekeeping components of *S. pneumoniae* CPS. Putative pathways are denoted by a dotted line, and the constituents of the repeat units are underlined. Glk, glucokinase; Pgi, glucose-6-phosphate isomerase; Pgm, phosphoglucomutase; GlmS, L-glutamine-D-fructose-6-phoshate amidotransferase; GlmM, phosphoglucosamine mutase; GlmU, UDP-*N*-acetyl-glucosamine pyrophosphorylase; GalU, UTP-glucose-1P uridylyltransferase; GalE, UDP-*N*-acetylglucosamine-4-epimerase/UDP-glucose-4-epimerase; RblA, ribitol-5-phosphate dehydrogenase; RblB, D-ribitol-5-phosphate cytidylyltransferase; AatA, UDP-GlcNAc 4,6-dehydratase; AatB, UDP-4-*keto*-6-deoxy-D-glucose 4-transaminase; ChoT (also known as LicB), choline transporter; ChoA (also known as LicA), choline kinase; ChoB (also known as LicC), cholinephosphate cytidylyltransferase. Spr numbers refer to the annotation of the *S. pneumoniae* R6 genome sequence.

D-galactose (UDP-AAT-Gal*p*), as discussed below. Glucopyranose, *N*-acetylgalactosamine, AAT-Gal*p*, ribitol-phosphate, and phosphorylcholine are components of *S. pneumoniae* teichoic and lipoteichoic acids, often called C-polysaccharide and F-polysaccharide, respectively (9, 38, 40).

Glucopyranose (Glc*p*) and galactopyranose (Gal*p*). UDP-Glc*p* is the precursor of Glc*p*, synthesized from Glc*p*-6P (Fig. 2) in two steps catalyzed by the cellular phosphoglucomutase Pgm and UTP-Glc-1P uridylyltransferase GalU (14, 54). Although these two genes are present on the pneumococcal chromosome (14, 30), they also are present within the type 3 *cps* locus but can be deleted, presumably due to this redundancy, and type 3 CPS synthesis depends mostly on the chromosomally encoded Pgm activity (29).

Exogenous galactose can be used for the synthesis of UDP-Gal*p* via the enzymes galactokinase (GalK) and Gal-1P uridylyltransferase (GalT) (27, 88). UDP-Glc*p* can also be converted to UDP-Gal*p* by UDP-Glc-4-epimerase (GalE) (28). Putative carbohydrate-specific ABC transporter genes and the *galE*, *galT*, and *galK* genes are present in the *S. pneumoniae* R6 genome (GenBank accession no. AE007317). Therefore, *S. pneumoniae* should be able to use exogenous galactose, although this has not been shown experimentally.

***N*-Acetylglucosamine (Glc*p*NAc).** Fructose-6P (Fru-6P) is housekeeping in *S. pneumoniae* as part of the glycolytic pathway (Fig. 2) and is a substrate for the synthesis of *N*-acetylglucosamine-1-phosphate (GlcNAc-1P); the first two steps of

Downloaded from jb.asm.org at Imperial College London on July 10, 2009

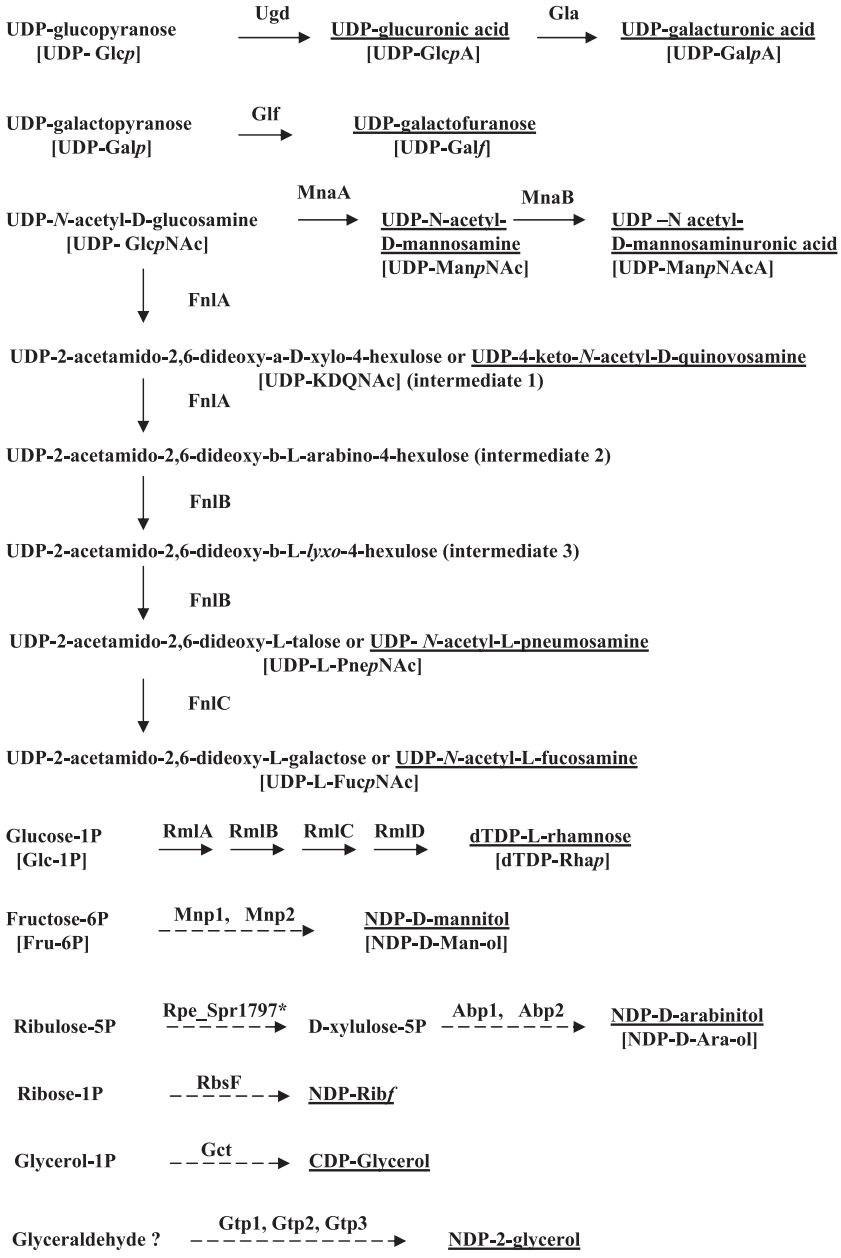


FIG. 3. Biosynthetic pathways of nonhousekeeping (CPS-specific) components of *S. pneumoniae* CPS. Putative pathways are denoted by a dotted line, and the constituents of the repeat units are underlined. All gene products are encoded within the *cps* loci, excepting ribulose phosphate 3-epimerase (Rpe), which is chromosomally encoded and marked by an asterisk. Ugd, UDP-glucose 6-dehydrogenase; Gla, UDP-galacturonate 4-epimerase; Glf, UDP-galactopyranose mutase; MnaA, UDP-*N*-acetylglucosamine-2-epimerase; MnaB, UDP-*N*-acetylmannosamine dehydrogenase; FnlA, steps 1 and 2 of UDP-FucNAc synthesis; FnlB, steps 3 and 4 of UDP-FucNAc synthesis; FnlC, step 5 of UDP-FucNAc synthesis; RmlA, glucose-1-phosphate thymidylyltransferase; RmlB, dTDP-D-glucose 4,6-dehydratase; RmlC, dTDP-4-keto-6-deoxy-D-glucose 3,5-epimerase; RmlD, dTDP-4-keto-L-rhamnose reductase; Mnp1, putative nucleotidyltransferase (NDP-mannitol pathway); Mnp2, putative reductase (NDP-mannitol pathway); RbsF, putative epimerase/dehydratase (NDP-ribose biosynthesis); Gct, CDP-glycerol biosynthetic protein; Gtp1-3, NDP-2-glycerol pathway; Abp1, putative nucleotidyltransferase (NDP-arabinitol pathway); Abp2, putative reductase (NDP-arabinitol pathway).

the pathway are catalyzed by L-glutamine-D-Fru-6P amidotransferase (GlmS) and phosphoglucosamine mutase (GlmM). The two genes appear to be present in the *S. pneumoniae* genome (spr0245 and spr1417, respectively; 32). The resultant Glc*p*N-1P is converted to UDP-Glc*p*NAc in a two-step reaction by the bifunctional enzyme GlmU, which possesses a pyrophosphorylase and an acetyltransferase domain (74). The latter sugar is also the substrate for the synthesis of some CPS nonhousekeeping components (Fig. 3).

***N*-Acetylgalactosamine (Gal*p*NAc).** A specific UDP-Glc*p*NAc-4-epimerase (Gne) is required for the conversion of UDP-Glc*p*NAc to UDP-Gal*p*NAc in *Escherichia coli* and *Yersinia enterocolitica* (11, 83). For a long time, it was assumed that GalE, a UDP-Glc-4-epimerase, was also a UDP-Glc*p*NAc-4-epimerase.

Downloaded from jb.asm.org at Imperial College London on July 10, 2009

It is now recognized that there are three classes of GalE-like proteins (33), i.e., epimerases that preferentially catalyze the conversion between UDP-Glc*p* and UDP-Gal*p* (group 1), epimerases that do not show a preference for either UDP-Glc*p*/UDP-Gal*p* or UDP-Glc*p*NAc/UDP-Gal*p*NAc (group 2), and epimerases that preferentially convert between UDP-Glc*p*NAc and UDP-Gal*p*NAc (group 3).

Several *S. pneumoniae* capsules contain Gal*p*NAc, but none of their *cps* loci have a *galE*/*gne* homolog. Nevertheless, Spr1460/NP_346051 (GenBank accession numbers AE007317 and AE005672), encoded within the *S. pneumoniae* genome, is closely related to the *Y. enterocolitica* AAC60777 and *Bacillus subtilis* NP_391765 enzymes, which are known to epimerize both UDP-Glc/UDP-Gal and UDP-GlcNAc/UDP-GalNAc. We conclude that Spr1460 is the product of a group 2 *galE* gene responsible for the synthesis of GalNAc in teichoic acid and CPS and Gal in CPS in the absence of exogenous galactose.

2-Acetamido-4-amino-2,4,6-trideoxy-D-galactose (AAT-Gal*p*). The uncommon sugar AAT-Gal*p* is a component of the CPS of *S. pneumoniae* serotype 1, of *S. pneumoniae* lipoteichoic and teichoic acids (38), and of *Streptococcus mitis* biovar I teichoic acid (13). It is also found in the *Bacteroides fragilis* capsules, the *Shigella sonnei* form I antigen (where it is called 4n-D-FucNAc) and *Plesiomonas shigelloides* serotype O17 antigen (21, 72, 87). To the best of our knowledge, these are the only occurrences in bacteria. As the latter three species lack teichoic acids, one might expect the AAT-Gal*p* pathway genes to be present in the polysaccharide biosynthetic loci and there are indeed appropriate genes that correlate with the presence of AAT-Gal*p*, and this made it possible to propose a biosynthetic pathway.

In *S. sonnei*, the proposed pathway for UDP-AAT-Gal*p* (87) involves a dehydratase (WbgZ) that converts UDP-Glc*p*NAc to 4-keto-6-deoxy-D-Glc*p*NAc, and then an amino group is added at carbon 4 by an aminotransferase (WbgX) to give UDP-AAT-Gal*p*. Finally, as AAT-Gal*p* is assumed to be the initial sugar of the repeat unit, UndPP-AAT-Gal*p* is produced by WbgY. In the case of *S. sonnei*, all three genes are present in the O-antigen gene cluster, whereas in the case of *B. fragilis*, the gene encoding the WbgZ homolog is located elsewhere on the chromosome (21, 87). Presumably, it is involved in other biosynthetic pathways and this would not be surprising as NDP-4-keto-6-deoxy sugars are often branch points in NDP-sugar biosynthesis (70).

In pneumococci, the AAT-Gal*p* biosynthetic genes should be chromosomally encoded due to the presence of AAT-Gal*p* in teichoic acid, and in the serotype 1 *cps* locus there are no gene products for UDP-AAT-Gal*p* synthesis. The products of spr0092 and spr1654 in the *S. pneumoniae* R6 genome (32) are equivalent to WbgZ and WbgX of *S. sonnei*, respectively, which we are renaming *aatA* and *aatB* to reflect their probable role in the synthesis of UDP-AAT-Gal*p* for both teichoic acids and the serotype 1 capsule (Fig. 2).

Phosphatidylcholine (Cho-P). The CPS structures of serogroup 15 (excepting 15A), serogroup 32, and serotype 27 have been reported to include Cho-P (38), although this is now in some doubt for serogroup 15 (36; see below). Cho-P is also a component of teichoic and lipoteichoic acids in *S. pneumoniae* (38, 40). Choline is required for growth, but *S. pneumoniae* is unable to synthesize choline, obtaining it from its human host. This requirement for choline means that it is always available for CPS synthesis if needed. Genes for the uptake of choline and synthesis of the precursor of Cho-P (*lic* genes) have been identified in several bacteria (73) and studied in *Haemophilus influenzae* (85). They were first recognized to be in *S. pneumoniae* by Zhang et al. (91) and are in the genome as *licB*, *licA*, and *licC*, which we have renamed *choT* (choline transporter), *choA* (choline phosphorylation), and *choB* (conversion to CDP-choline), respectively, to reflect their role in choline transport and CDP-choline synthesis (Fig. 2).

Ribitol-phosphate (Rib-ol-P). CDP-D-ribitol-5P is the precursor for the addition of D-ribitol-5P, which is also a constituent of teichoic acids and hence available for pneumococcal CPS synthesis (40). In *H. influenzae*, CDP-ribitol synthesis is catalyzed by the bifunctional enzyme Bcs1 in a two-step reaction (93). In *B. subtilis*, two separate genes, *tarJ* and *tarI*, encode a reductase that converts ribose-5P to ribitol-5P and a cytidylyltransferase that converts ribitol-5P to the precursor CDP-ribitol, respectively (44). The products of spr1148 and spr1149 in the *S. pneumoniae* R6 genome show a high level of sequence similarity to TarJ and TarI of *B. subtilis*, and as these pneumococcal homologs should have the same function, we name them *rblA* and *rblB*, respectively, to reflect their role in the synthesis of the ribitol-5P precursor (Fig. 2). Note that the single report of ribitol-1P in types 11B and 11F (38) appears to be a case of a different name for ribitol-5P.

(ii) Nonhousekeeping (CPS-specific) components. There are 14 components in the known *S. pneumoniae* CPS biochemical structures that are not available from general metabolism and for which the biosynthetic genes are found in the pneumococcal *cps* loci (Fig. 3). Many of the genes discussed below are also found in polysaccharide biosynthetic loci of other species. There are well-documented pathways from a number of bacterial species for the synthesis of the NDP-linked precursors of rhamnose, galactofuranose, glucuronic acid, galacturonic acid, *N*-acetylmannosamine, *N*-acetylmannosaminuronic acid, *N*-acetylfucosamine, *N*-acetyl-L-pneumosamine, and glycerol-1-phosphate (70), and in pneumococci the relevant genes are always present intact in the *cps* locus when the sugar is present in the structure (12). The pathways for the remaining components are less well documented, but the allocation of *cps* genes to these pathways seems secure from the correlation between the presence of a particular gene(s) and these components in the CPS.

Rhamnose (L-Rhap)—*rmlA*, *rmlB*, *rmlC*, and *rmlD*. Rhamnose is present in 23 CPS structures, and all of their *cps* loci contain the four *rml* genes for synthesis of the precursor dTDP-rhamnose (34) (Fig. 3). The gene order (*rmlA*, *-C*, *-B*, and *-D*) is always the same in *S. pneumoniae*, but there are incomplete sets of *rml* genes in a few other serotypes, and where the structure is known they lack rhamnose (12). For example, the *cps* loci of types 1 (63) and 24B have frameshifts within the *rmlD* and *rmlC* genes, respectively, whereas the *cps* locus of type 15F has only a partial *rmlB* and an *rmlD* gene present. A truncated UDP-galactofuranose biosynthetic gene (*glf*) and/or remnants of transposase genes are usually present downstream of the *rml* genes.

Galactofuranose (D-Gal*f*)—*glf*. The precursor of galactofuranose is UDP-galactofuranose, which is synthesized by Glf from UDP-galactopyranose (34) (Fig. 3). The presence of an

Downloaded from jb.asm.org at Imperial College London on July 10, 2009

intact *glf* gene correlates with Gal*f* in CPS; however, type 15F is an exception as it has an intact *glf* gene but there is no Gal*f* in the CPS. There are cases where two copies of the *glf* gene are present in the *cps* locus, the second copy being truncated in some cases.

Glucuronic acid (D-Glc*p*A) and galacturonic acid (D-Gal*p*A)—*ugd* and *gla*. UDP-Glc*p*A is the precursor of Glc*p*A and is synthesized by Ugd from UDP-Glc*p* (Fig. 3) and then is converted by Gla to UDP-Gal*p*A, the precursor of Gal*p*A (12, 34, 61, 63). There are 18 *cps* loci that have *ugd*, including the type 3 *cps* locus. Where known, the CPS structures (or CPS constituents) of these serotypes contain Glc*p*A, with the exception of serotype 1; the latter *cps* locus has both *ugd* and *gla*, and thus serotype 1 contains Gal*p*A rather than Glc*p*A.

UDP-*N*-acetyl-D-mannosamine (D-Man*p*NAc) and UDP-*N*-acetyl-D-mannosaminuronic acid (D-Man*p*NAcA)—*mnaA* and *mnaB*. The *mnaA* gene for the synthesis of D-Man*p*NAc (34, 70) (Fig. 3) is present in 15 *cps* loci, whereas *mnaB*, for the conversion of D-Man*p*NAc to D-Man*p*NAcA, is present in only 5 of these (serogroup 12 and types 44 and 46). The available CPS structures are consistent with that expected from the presence of *mnaA* alone or *mnaA* and *mnaB*; thus, type 4 and serogroups 9 and 19 contain D-Man*p*NAc, whereas types 12A and 12F contain D-Man*p*NAcA.

***N*-Acetyl-L-fucosamine (L-Fuc*p*NAc)—*fnlA*, *fnlB*, and *fnlC*.** The precursor of L-Fuc*p*NAc is UDP-L-Fuc*p*NAc, which is synthesized in a five-step pathway (60) (Fig. 3) by the products of three genes (FnlA, -B, and -C), the first two (FnlA and FnlB) being bifunctional. The three genes are present in eight *cps* loci, and L-Fuc*p*NAc is present in the CPS where the structures are known (serotypes 4, 5, 12F, 12A, and 45).

***N*-Acetyl-L-pneumosamine (L-Pne*p*NAc) and 2-acetamido-2,6-dideoxy-D-xylo-hexos-4-ulose (4-keto-*N*-acetyl-D-quinovosamine, KDQNAc).** The type 5 CPS contains L-Pne*p*NAc and KDQNAc, in addition to L-Fuc*p*NAc, and the *fnlA*, *fnlB*, and *fnlC* genes are present in the *cps* locus. The precursor of L-Pne*p*NAc is UDP-Pne*p*NAc, a UDP-Fuc*p*NAc intermediate synthesized by FnlB, and UDP-KDQNAc, the likely precursor of KDQNAc, is the first intermediate before undergoing further modification by FnlA (60) (Fig. 3). Both L-Pne*p*NAc and KDQNAc are uniquely present in type 5, and the need to divert some of the intermediate of the first FnlA reaction to a KDQNAc transferase reaction may explain the much greater sequence divergence between the *fnlA* gene of type 5 and those of the other serotypes than observed for *fnlB* and *fnlC* (data not shown).

Glycerol-1P (Gro-1P)—*gct*. The precursor of glycerol-1P is CDP-glycerol, which is synthesized by Gct (34) (Fig. 3), and the *gct* gene is present in 13 *cps* loci. An intact *gct* gene correlates with the presence of Gro-1P in the CPS of serogroup 18 and types 11A, 11C, and 45, whereas types 11B and 11F have a truncated *gct* gene and Gro-1P is not found in their CPS, being replaced by Rib-ol-1P.

Glycerol-2P (Gro-2P)—*gtp1*, *gtp2*, and *gtp3*. The *gtp1*, *gtp2*, and *gtp3* genes are present in the *cps* loci of serogroups 15, 23, and 28, and their products are thought to be responsible for the synthesis of NDP-2-glycerol (Fig. 3). Gro-2P is rarely present in bacteria and was originally only reported to be present in types 15A and 23F, with Cho-P replacing it in types 15F, 15B, and 15C (38). However, recent nuclear magnetic resonance (NMR) analysis indicates that type 15B contains glycerol and not choline (36). As the earlier NMR spectra of types 15B and 15C were superimposable, this most likely also applies to types 15C and 15F, the reported presence of Cho-P being due to contamination of the CPS with teichoic acid.

The sugar phosphate transferase for Gro-2P is proposed to be encoded by *wchX* (see below), and the presence of Gro-2P in the CPS of types 15A and 23F correlates with the presence of *wchX* and *gtp1* to *-3* in their *cps* loci (12, 55, 67). We consider the balance of the evidence to favor the presence of Gro-2P in the CPS of all four serotypes within serogroup 15, as all possess *wchX* and *gtp1* to *-3* and there are no differences in their *cps* loci that would account for the different side groups in the structures reported earlier.

Mannitol-6P (Man-ol-6P)—*mnp1* and *mnp2*. The presence of mannitol-6P in the CPS of type 35A correlates with the presence of two genes (*mnp1* and *-2*), which are also present in the *cps* loci of types 35C and 42, for which there are no available CPS structures. Mnp1 possesses the PF00483 family domain of NTP-transferases, i.e., enzymes that transfer nucleotides onto phosphosugars. Mnp2 possesses the PF01370 family domain of NAD-dependent epimerases/dehydratases. Mnp1 and −2 could be responsible for the synthesis of NDP-Man-ol-6P from D-fructose-6P by a two-step pathway parallel to that for CDP-ribitol-5P formation from ribulose-5P by TarJ and TarI (see above). The two steps are addition of the nucleoside monophosphate to give NDP-fructose-6P and reduction of the keto group on carbon 2. However, the order of the two reactions is not known and there is no biochemical support for this pathway as yet. The genes will be named *mnpA* and *mnpB* in order of function if this proposal is confirmed.

Arabinitol-1P (Ara-ol-1P)—*abp1* and *abp2*. Arabinitol-1P is only reported in the CPS of type 17F. The *abp1* and *−2* genes present in the type 17F *cps* locus encode proteins containing the PF00483 and PF01370 family domains of NTP-transferases and NAD-dependent epimerase/dehydratase family domains, respectively; thus, they are proposed to be involved in the biosynthesis of NDP-Ara-ol-1P. D-Xylulose-5P is produced from ribulose-5P (present as an intermediate of the pentose phosphate pathway) by a ribulose phosphate 3-epimerase (Rpe) (2), and the *rpe* gene (spr1797) is present in the *S. pneumoniae* R6 genome (GenBank accession number AE007317). NDP-Ara-ol-1P is proposed to be derived from D-xylulose-5P in a two-step pathway catalyzed by the products of the *abp1* and *abp2* genes. As there is no biochemical support for the pathway and the order of the two reactions of the *abp1* and *−2* gene products is not known, the genes will be renamed *abpA* and *abpB* in order of function if this proposal is confirmed.

The *abp1* and *abp2* genes are also present in the *cps* loci of serogroup 24 and types 13 and 48. Of these, only the structure of type 13 CPS is available and contains Rib-ol-5P but not Ara-ol-1P. However, the structure is old and there are no recent NMR data to confirm the presence of Rib-ol-5P rather than Ara-ol-1P. Alternatively, the absence of Ara-ol-1P could be attributed to the sequence divergence of the type 13 *abp1* and *abp2* genes compared to those present in the other *cps* loci, but new NMR data are required to address this issue.

Ribofuranose (Rib*f*)—*rbsF*. ADP-ribose (ADP-ribofuranose) is believed to be available from the recycling of NAD. It is a

Downloaded from jb.asm.org at Imperial College London on July 10, 2009

component of the *E. coli* O114 antigen (23), but no gene for ADP-ribose biosynthesis is present in this O-antigen gene cluster (24). However, in *S. pneumoniae* types 19B and 19C, the *rbsF* (also known as *cps19R*) gene was proposed to be responsible for the synthesis of NDP-ribose (57). RbsF possesses the PF01370 family domain of NAD-dependent epimerases/dehydratases. The *rbsF* gene is also present in the *cps* loci of types 7B, 7C, 24B, 24F, and 40. The CPS of type 7B and the constituents of type 24F also include Rib*f* (constituents of the other serotypes are not known), which adds support to the proposed function of RbsF.

Assignment of initial sugar transferases. Repeat unit synthesis mediated by the Wzy pathway is initiated by transfer of a sugar phosphate to a lipid carrier (Fig. 1) (77, 86). This reaction is catalyzed by an initial transferase (IT) and is reversible (66), whereas the reactions of the GTs that add the other sugars are, in practical terms, irreversible. This can be important in the allocation of gene function if an IT is able to transfer more than one sugar phosphate. In such cases, the specificity of the second sugar transferase determines which sugar is incorporated into the repeat unit, as this reaction is irreversible. Determination of the CPS repeat unit structure does not identify which glycosidic link is made by the Wzy polymerase, and repeat units are generally depicted chemically as circular molecules. Identification of the initial sugar (IT specificity) allows us to identify the linkage made by Wzy and to represent the repeat units as they are synthesized on the lipid carrier.

In most *S. pneumoniae cps* loci, the fifth gene appears to encode the proposed IT (WchA, WciI, WcjG, and WcjH; see below), as each possesses the same Pfam domain (PF02397) (12) (Table 1; see Fig. S1 in the supplemental material). In a few *cps* loci, the IT gene is not immediately downstream of the fourth regulatory gene (*wze*) because of the presence of an intervening gene or gene fragment and/or a rearrangement in the order of the first four *cps* genes. Thus, the gene order differs in the closely related *cps* loci of types 25A, 25F, and 38 and a defective transposase gene is between *wze* and *wzg*. Similarly, there is a defective transposase gene upstream of the IT gene (*wciI*) in the *cps* loci of types 12A and 46.

In *S. pneumoniae* type 1, there is no *cps* gene product harboring the Pfam domain PF02397 found in other ITs. Furthermore, none of the predicted gene products, other than Wzx and Wzy, have putative transmembrane segments, making it unlikely that the IT is encoded within the serotype 1 *cps* locus. The constituents of the type 1 CPS repeat unit are an AAT-Gal*p* and two α-D-Gal*p*A residues (38). The *cps* locus possesses the first four *cps* genes (*wzg*, *wzh*, *wzd*, and *wze*), the *wzy* and *wzx* genes, the *ugd* and *gla* genes for the biosynthesis of Gal*p*A, and a set of *rml* genes (as mentioned earlier, *rmlD* is frameshifted and therefore rhamnose is absent from the CPS) (12, 63). Type 1 CPS is acetylated, and accordingly, there is also a putative acetyltransferase gene (*wchC*). The products of the remaining genes (*wchB* and *wchD*) both possess the PF00534 domain found in GTs and are presumably involved in the assembly of the trisaccharide repeat unit. Thus, there is no candidate gene that could encode an IT.

As mentioned previously, WbgY in *S. sonnei* and WcfS in *B. fragilis* are thought to be the ITs for AAT-Gal*p* (21, 87), and in the *S. pneumoniae* genome, there is a WbgY/WcfS homolog (Spr1655; GenBank accession number AE007317) which possesses the PF02397 family domain found in ITs encoded within the other *cps* loci. Therefore, it is assumed that this chromosomally encoded IT provides the lipid-linked-AAT-Gal*p* required for repeat unit synthesis in serotype 1 pneumococci. Coyne et al. (21) suggested that the pneumococcal WbgY homolog makes lipid-linked-AAT-Gal*p* for initiating teichoic acid synthesis, but AAT-Gal*p* does not appear to be the initial sugar in teichoic acid repeat units (9). We assume that WbgY is present on the chromosome as there is some other noncapsular pneumococcal polysaccharide that has AAT-Gal*p* as the initial sugar.

WchA. The IT WchA has been shown to transfer Glc*p*-1P from UDP-Glc*p* to a lipid carrier in the *S. pneumoniae* capsules of types 8, 9V, and 14 (42, 65, 78). The data obtained for serotype 2 (15) made it nearly certain that UndPP is the lipid carrier for initiation of capsule synthesis in pneumococci, as shown for several other bacterial polysaccharides. WchA has been shown to be essential for CPS biosynthesis in *S. pneumoniae* serotype 8 as unencapsulated variants possess frameshift mutations within *wchA* (also known as *cap8E*) (81). Glucose is assumed to be the donor sugar used by WchA in all of the *S. pneumoniae* serotypes where this IT is present. This is supported by the observation that glucose is present in the repeat units of all serotypes that possess *wchA* (see Fig. S1 in the supplemental material). Where there are two glucose residues in the main chain of the repeat unit (serogroup 18, serogroup 32, and serotypes 2, 7B, 8, 20, 9A, 9L, and 9V), the initial sugar can be determined by GT-linkage correlation of the second sugar transferase (see GT section below.)

WciI. Serogroup 12 and serotypes 4, 5, 25A, 25F, 38, 44, 45, and 46 possess the putative IT gene *wciI* in their *cps* loci (see Fig. S1 in the supplemental material). Structures are known for types 4, 5, 12F, 12A, and 45, and for type 46 the sugar composition is known. L-Fuc*p*NAc is the only sugar common to all of these serotypes, but serotypes 25A, 25F, and 38 lack the L-Fuc*p*NAc pathway genes, suggesting that their CPS lacks this sugar. However, all but serotype 5 contain Gal*p*NAc and/or Glc*p*NAc (serotypes 4, 12F, and 45 contain only Gal*p*NAc, and only Glc*p*NAc is present in serotype 12A), and a possible scenario is that both Gal*p*NAc-P and Glc*p*NAc-P can be transferred by WciI. There is a precedent in the IT WecA of *E. coli* and *Y. enterolitica* (4, 50, 92) that transfers Glc*p*NAc-P to initiate synthesis of the repeat unit of the enterobacterial common antigen but also acts as the IT for O antigens in *E. coli* that contain Glc*p*NAc or Gal*p*NAc. In the case of WecA, the gene is outside of the O-antigen gene cluster and it is unlikely that WecA itself determines if Glc*p*NAc or Gal*p*NAc is the initial sugar. However, as already mentioned, the transfer of a sugar phosphate to UndP to give a pyrophosphate linkage is reversible, and it is envisaged that both UndPP-Glc*p*NAc and UndPP-Gal*p*NAc are formed, the next transferase determining which is permanently incorporated into the repeat unit. The reversibility of the first reaction ensures that no UndP is locked up as UndPP-Glc*p*NAc or UndPP-Gal*p*NAc if it is not used in further reactions.

Besides correlating the presence of a particular sugar in the CPS repeat unit with the presence of an IT gene (as above), it is in some cases possible to predict the initial sugar (and thus the polymerase linkage) by correlating the nature of the link-

Downloaded from jb.asm.org at Imperial College London on July 10, 2009

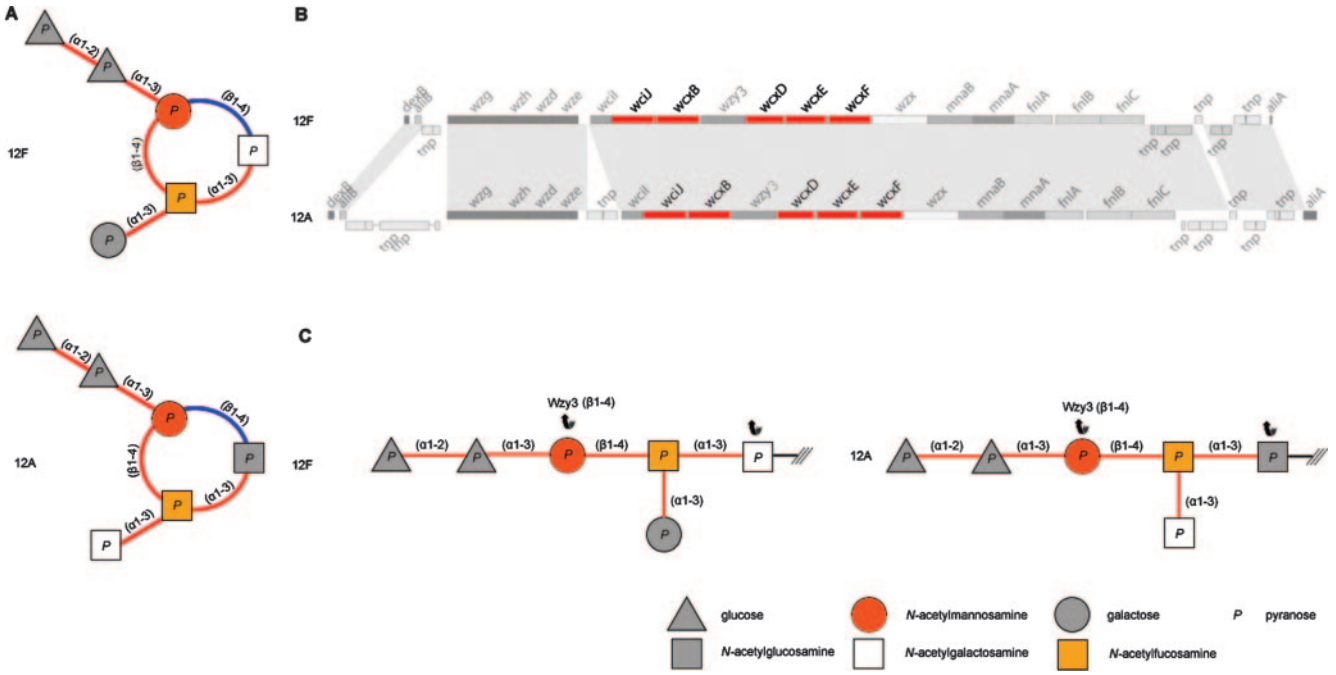


FIG. 4. Assignment of Wzy polymerase linkage and IT specificity. (A) Circularized representation of CPS repeat unit structures for serotypes 12F and 12A, where the polymerase linkage and the initial sugar are not indicated. Each sugar is represented symbolically as in the report by Bentley et al. (12), and the nature of the linkage is represented by red if it is retaining and blue if it is inverting. Each circular structure contains five retaining linkages and a single inverting linkage. (B) Excepting remnant transposases, the *cps* clusters of serotypes 12F and 12A are syntenic, and both encode five GTs, all of which are retaining based on CAZy GT family membership (colored red). Shading between the two loci represents TBLASTX amino acid sequence similarity viewed by using ACT. (C) We can infer that the single inverting linkage in the circular structure (A) corresponds to the linkage made by the polymerase, which also allows us to define the initial sugar and depict the structures biologically, with the lipid-linked initial sugar depicted to the right of the structure. ///, lipid carrier.

ages (either retaining or inverting) in the circularized repeat unit with the proposed nature of the GTs encoded by the *cps* locus from their CAZy family membership (see Materials and Methods). For example, the repeat units of serotypes 12F and 12A can be depicted as in Fig. 4A. The circular chemical structures each contain six linkages, three main chain and three side branch linkages. Of these six linkages, five are retaining and one is inverting. One of the main chain linkages is the Wzy linkage and if identified would also specify the initial sugar.

Excepting remnant transposases, the *cps* loci of serotypes 12F and 12A are syntenic and are predicted to encode five GTs, all of which are members of retaining GT families (Fig. 4B). We can therefore suggest that the single inverting linkage in the circular chemical representation of the repeat unit represents the Wzy polymerization linkage, allowing us to define the initial sugar (and hence WciI specificity) as Gal*p*NAc in 12F and Glc*p*NAc in 12A and also to define the polymerase linkages (Fig. 4C). This approach also suggests that Gal*p*NAc is the initial sugar for WciI in serotypes 4 and 45, which is further supported by the presence of Fuc*p*NAc as the second sugar (see below).

If Glc*p*NAc-P and Gal*p*NAc-P are the normal substrates for WciI, we have to account for serotype 5 CPS, which contains only Glc*p*, Fuc*p*NAc, Pne*p*NAc, and KDQNAc, one of which must be the substrate for WciI. As mentioned above, the specificity of the second sugar transferase can determine which sugar is incorporated into the repeat unit.

In all *cps* loci where *wciI* is the putative IT gene and Fuc*p*NAc is also presumed to be present in the CPS (due to the presence of the *fnlA*, *fnlB*, and *fnlC* genes in the *cps* loci), the sixth gene encodes the GT, wciJ. In serotypes 25F, 25A, and 38, where *wciI* is also the putative IT gene but the sixth gene is not *wciJ*, Fuc*p*NAc is not predicted to be present in the repeat unit. Thus, WciJ should be the Fuc*p*NAc transferase that transfers the second sugar in the repeat unit. This allows us to infer that the initial sugar in serotype 5 is KDQNAc. Clearly, biochemical studies of WciI are required to evaluate our proposals for the initial sugars in these serotypes.

WcjH. The *cps* loci of serotypes 29, 35F, 39, 43, and 47F possess *wcjH*, but only the CPS structure of type 29 is available and is similar to that of type 35B. There are differences in the acetylation patterns, due to the presence of an extra acetyltransferase gene (*wciG*) in type 35B, but apart from these differences, the CPS of types 29 and 35B differ in only one sugar (Gal*p* and Glc*p*, respectively) and their *cps* loci have different IT genes (*wcjH* and *wchA*, respectively); since WchA transfers Glc*p*, it seems clear from this comparison that in serotype 29 WcjH transfers Gal*p*.

WcjG. The putative IT gene *wcjG* is found in the *cps* loci of serogroup 10 and serotypes 31, 33C, and 47A (see Fig. S1 in the supplemental material), but there are biochemical structures only for 10A, 10F, and 31. Gal*p*, Gal*f*, Gal*p*NAc, and Rib-ol are the constituents of the pentasaccharide backbone of type 10A and 10F CPS, differing only in the anomeric status of Gal*p* (α in 10F and β in 10A). Gal*f*, Rha*p*, and Glc*p*A are the constituents of type 31 CPS (38), which might suggest that

Downloaded from jb.asm.org at Imperial College London on July 10, 2009

TABLE 2. GT genes and the linkages their gene products are predicted to catalyze[a]

GT gene	Pfam	CAZy	Mechanism	No. of occurrences of GT gene	Structure (donor, linkage, acceptor)	No. of occurrences where CPS structure known	Serotype(s) possessing GT gene	No. of occurrences, where structure known, that have the linkage	No. of occurrences of linkage where GT gene is absent	Note(s)
wchF	PF00534	4	RET	27	L-Rha*p* β1-4 β-D-Glc*p*	14	2, 7F, 7A, 7B, 17F, 18F, 18A, 18B, 18C, 22F, 23F, 27, 32F, 32A	14	0	A1
wciB				24	D-Gal*f* β1-3 α-D-Gal*p*	10	10F	1	0	A2, A3
					D-Gal*f* β1-3 β-D-Gal*p*		10A, 29	2	1	
					D-Gal*f* β1-3 β-D-Glc*p*		17A, 20, 33F, 35A, 35B	5	0	
					D-Gal*f* β1-3 α-D-Glc*p*		34	1	0	
					Structure uncertain		31			
wcrC	PF00534	4	RET	13	D-Gal*p* α1-2 D-Rib-ol	4	10F, 10A, 33B, 34	4	0	A4
wciF	PF00535	2	INV	13	D-Gal*p*NAc β1-3 α-D-Gal*p*	5	10F, 10A	2	0	A6
					D-Glc*p*NAc β1-4 D-Rib-ol		13	1	0	
					D-Gal*p*NAc β1-4 α-D-Gal*p* OR D-Gal*f* β1-3 β-D-Gal*p*NAc		33B	1	1	
					D-Gal*p* β1-3 α-D-Gal*p* OR D-Gal*f* β1-3 β-D-Gal*p*		33F	1	1	
wchJ	PF08660			12	GT enhancer					A7
wchK	PF04101	1	INV	12	D-Gal*p* β1-4 β-D-Glc*p*	10	11F, 11A, 11B, 11C, 13 14, 15F, 15A, 15B, 15C	10	0	A7
wcrD	PF00535	2	INV	10	D-Gal*f* β1-3 β-D-Gal*p*NAc	3	10A	1	1	A5
					D-Gal*f* β1-4 β-D-Glc*p*NAc		13	1	0	
					D-Gal*f* β1-3 α-D-Gal*p*		34	1	1	
wchO	PF03808	26	INV	9	D-Man*p*NAc β1-4 β-D-Glc*p*	8	9A, 9L, 9N, 9V, 19B, 19C	6	0	A8
					D-Man*p*NAc β1-4 α-D-Glc*p*		19F, 19A	2	0	
					D-Man*p*NAc β1-4 α-L-Rha*p*		19B, 19C	2	0	
wcrH			INV	9	D-Gal*f* β1-1 D-Rib-ol	4	29, 35B	2	0	A9
					D-Gal*f* β1-1 Man-ol		35A	1	0	
					D-Gal*f* β1-6 α-D-Gal*p*		10F	1	0	
wciJ	PF00534	4	RET	8	L-Fuc*p*NAc α1-3 β-D-Glc*p*NAc	5	12A	1	0	A10
					L-Fuc*p*NAc α1-3 α-D-Gal*p*NAc		4	1	0	
					L-Fuc*p*NAc α1-3 β-D-Gal*p*NAc		12F, 45	2	0	
					L-Fuc*p*NAc α1-3 β-D-KDQNAc		5	1	0	
wciE	PF05704	32	RET	7	D-Gal*p* α1-2 α-D-Galp	2	33F, 33B	2	0	A6
wciU	PF00534	4	RET	7	D-Glc*p*NAc α1-3 β-L-Rha*p*	4	18A	1	0	A11
					D-Glc*p* α1-3 β-L-Rha*p*		18F, 18B, 18C	3	1	
wciD	PF00535	2	INV	6	Assignment uncertain	3	20			A25
					D-Gal*p*NAc β1-4 α-D-Gal*p* OR D-Gal*f* β1-3 β-D-Gal*p*NAc		33B	1	1	A6
					D-Gal*p* β1-3 α-D-Gal*p* OR D-Gal*f* β1-3 β-D-Gal*p*		33F	1	1	
wchQ	PF00535	2	INV	6	L-Rha*p* α1-P-4 β-D-Man*p*NAc	6	19F, 19A, 19B, 19C	4	0	A8
					L-Rha*p* α1-P-2 β-D-Man*p*NAc		32A, 32F	2	0	

Downloaded from jb.asm.org at Imperial College London on July 10, 2009

wcxB	PF00534	4	RET	6	D-Gal*p* D-Gal*p*NAc	α1-3 α1-3	α-L-Fuc*p*NAc α-L-Fuc*p*NAc	3	12F, 45 12A	2 1	0 0	A12
wcyK	PF00534	4	RET	6	D-Gal*p*	α1-3	β-D-Gal*p*	4	11F, 11A, 11B, 11C	4	0	A15
wcwA	PF00534	4	RET	5	D-Gal*p* D-Glc*p*	α1-3 α1-3	β-L-Rha*p* β-L-Rha*p*	3	7F, 7A 22F	2 1	1 1	A13 A14
wcxD	PF00534	4	RET	5		Assignment uncertain		2	12F, 12A			A12
wcxE	PF00534	4	RET	5		Assignment uncertain		2	12F, 12A			A12
wcxF	PF00534	4	RET	5		Assignment uncertain		2	12F, 12A			A12
wcrL	PF04488	32	RET	5	D-Glc*p*NAc D-Glc*p*	α1-4 α1-4	α-D-Gal*p* α-D-Gal*p*	4	11F, 11B, 11C 11A	3 1	0 1	A15
wchL	PF00535	2	INV	5	D-Glc*p*NAc	β1-3	β-D-Gal*p*	5	14, 15F, 15A, 15B, 15C	5	0	A16
wchM	PF00535	2	INV	5	D-Gal*p*	β1-4	β-D-Glc*p*Nac	5	14, 15F, 15A, 15B, 15C	5	0	A16
wchN	PF05704	32	RET	5	D-Gal*p*	α1-2	β-D-Gal*p*	4	15F, 15A, 15B, 15C	4	0	A17
wciN	PF01501	8	RET	5	D-Gal*p*	α1-3 Nonfunctional	α-D-Glc*p*	3	6A, 6B 33B	2	0	A18 A6
wcjA	PF00534	4	RET	5		Assignment uncertain		4	9A, 9L, 9N, 9V			A19
wcjB				4		Assignment uncertain		4	9A, 9L, 9N, 9V			A19
wcjC	PF00534	4	RET	4		Assignment uncertain		4	9A, 9L, 9N, 9V			A19
wciV	PF00535	2	INV	4	D-Gal*p* D-Gal*p*	β1-4 β1-4	α-D-Glc*p*Nac α-D-Glc*p*	4	18A 18F, 18B, 18C	1 3	1 0	A11
wciW	PF05704	32	RET	4	D-Glc*p*	α1-2	β-D-Gal*p*	4	18F, 18A, 18B, 18C	4	0	A11
wcrP	PF00535	2	INV	4	D-Glc*p*A L-Rha*p*	β1-3 α1-4	β-D-Gal*f* OR β-D-Glc*p*A	2	17A, 31	1	0	A20
wcrR	PF00535	2	INV	4	D-Glc*p*A L-Rha*p*	β1-3 α1-4	β-D-Gal*f* OR β-D-Glc*p*A	2	17A, 31	1	1	A20
wcyS	PF05704	32	RET	4	D-Gal*p* D-Glc*p*	α1-4 α1-4	β-L-Rha*p* β-L-Rha*p*	3	27 32F, 32A	1 2	0 0	A21
wcrW		4	RET	3		Structure uncertain		1	31			
wcxU	PF00535	2	INV	3		Assignment uncertain		1	7B			A22
wcyA	PF00534	4	RET	3		No structure		0				
wcyB	PF00535	2	INV	3		No structure		0				
wcyC	PF00534	4	RET	3		No structure		0				
wcyD				3		No structure		0				
wcrQ	PF05704	32	RET	3	D-Glc*p*	α1-2	β-D-Gal*f*	1	17A	1	0	A23
wcrI	PF00535	2	INV	3	D-Gal*p*	β1-3	β-D-Gal*f*	1	35A	1	0	A24
wcrK	PF05704	32	RET	3		No structure		0				
wcxI	PF00535	2	INV	3		No structure		0				
wcxJ				3		No structure		0				
wcxN		1 and 2	INV/INV	3		No structure		0				
wciL	PF00534	4	RET	3		Assignment uncertain		3	4, 20, 45			A25

Continued on following page

Downloaded from jb.asm.org at Imperial College London on July 10, 2009

TABLE 2—*Continued*

GT gene	Pfam	CAZy	Mechanism	No. of occurrences of GT gene	Structure (donor, linkage, acceptor)			No. of occurrences where CPS structure known	Serotype(s) possessing GT gene	No. of occurrences, where structure known, that have the linkage	No. of occurrences of linkage where GT gene is absent	Note(s)
wcrG	PF02485	14	INV	3	D-Gal*p*	β1-6	β-D-Gal*p*NAc	1	10A	1	0	A5
wciP	PF00535	2	INV	3	L-Rha*p* L-Rha*p* L-Rha*p*	α1-2 α1-3 α1-4	D-Ara-ol D-Rib-ol D-Rib-ol	3	17F 6A 6B	1 1 1	0 0 0	A26
wciC	PF00534	4	RET	3	D-Gal*p*	α1-3	β-D-Gal*f*	1	33F	1	0	A6
wchV	PF00535	2	INV	3	D-Gal*p*	β1-4	β-L-Rha*p*	1	23F	1	2	A14, A27
wchW				3	L-Rha*p*	α1-2	β-D-Gal*p*	1	23F	1	0	A28
wcwI	PF00535	2	INV	3	Assignment uncertain			1	7B			A22
wcwL	PF00534	4	RET	3	Assignment uncertain			1	7B			A22
wcwX				2	Assignment uncertain			1	22F			A29
wcwV	PF00534	4	RET	2	Assignment uncertain			1	22F			A29
whaB	PF00535	2	INV	2	Assignment uncertain			1	22F	1	1	A29
wchR	PF00534	4	RET	2	D-Rib*f*	β1-4	α-L-Rha*p*	2	19B, 19C	2	1	A8
wchS	PF00535	2	INV	2	L-Rha*p*	α1-3	β-D-Man*p*NAc	2	19B, 19C	2	0	A8
wcxS	PF00535	2	INV	2	L-Rha*p*	α1-3	α-D-Gal*p*	1	45	1	0	A30
wcrM	PF00535	2	INV	2	D-Gal*p*NAc	β1-6	β-D-Gal*f*	2	29, 35B	2	0	A31, A9
wcrV	PF00535 PF05704	2 32	INV RET	2	D-Gal*p* D-Gal*p*	β1-4 α1-3	β-L-Rha*p* β-L-Rha*p*	2	17F, 17A 17F, 17A	2 2	1 2	A14
wcyE				2	No structure			0				
wcwD				2	D-Gal*p*	β1-2	α-D-Gal*p*	2	7F, 7A	1	0	A32
wcwF	PF00535	2	INV	2	D-Gal*p*NAc L-Rha*p*	β1-6 α1-4	α-D-Gal*p* OR β-D-Gal*p*NAc	2	7F, 7A	2	0	A33
wcwG				2	D-Gal*p*NAc L-Rha*p*	β1-6 α1-4	α-D-Gal*p* OR β-D-Gal*p*NAc	2	7F, 7A	2	0	A33
wcwH	PF00534	4	RET	2	D-Glc*p*NAc	α1-2	α-L-Rha*p*	2	7F, 7A	2	1	A34
wcyM	PF00534	4	RET	2	No structure			0				
wcyN	PF00535	2	INV	2	No structure			0				
wcrT				2	L-Rha*p*	β1-4	α-L-Rha*p*	2	17F, 17A	2	0	A35
wcyQ	PF00534	4	RET	1	Assignment uncertain			1	45			
wcyT	PF00535	2	INV	1	No structure			0				
wcyU	PF00535	2	INV	1	No structure			0				
wcyV				1	No structure			0				
wchU	PF00534	4	RET	1	D-Glc*p*	β1-6	β-D-Man*p*NAc	1	19C	1	0	A8
wchG	PF00535	2	INV	1	L-Rha*p*	α1-3	β-L-Rha*p*	1	2	1	0	A36
wchH	PF00534	4	RET	1	D-Glc*p* D-Glc*p*A	α1-2 α1-6	α-L-Rha*p* OR α-D-Glc*p*	1	2	1	0	A36

Downloaded from jb.asm.org at Imperial College London on July 10, 2009

wchI	PF00534	4	RET	1	D-Glc*p* D-Glc*p*A	α1-6 α1-6	α-L-Rha*p* OR α-D-Glc*p*		1	2	1	0	A36
wciS	PF00534	4	RET	1	D-Gal*p*	α1-4	β-D-Glc*p*A		1	8	1	0	A37
wciT	PF04488	32	RET	1	D-Glc*p*	α1-4	α-D-Gal*p*		1	8	1	1	A37
wchB	PF00534	4	RET	1	D-Gal*p*A D-Gal*p*A	α1-3 α1-3	AAT-α-D-Gal*p* OR D-Gal*p*A		1	1	1	0	A38
wchD	PF00534	4	RET	1	D-Gal*p*A D-Gal*p*A	α1-3 α1-3	AAT-α-D-Gal*p* OR D-Gal*p*A		1	1	1	0	A38
wciK				1	Assignment uncertain				1	4			A25
whaC	PF00535	2	INV	1	Assignment uncertain				1	5	1	0	A39
whaD	PF00534	4	RET	1	Assignment uncertain				1	5			A39
whaE				1	Assignment uncertain				1	5			A39
whaF	PF05704	32	RET	1	Assignment uncertain				1	20			A25
whaJ	PF00535	2	INV	1	Assignment uncertain				1	20			A25
whaK	PF00535	2	INV	1	D-Glc*p*NAc	β1-3	α-D-Gal*p*		1	27	1	0	A40
whaM	PF00535	2	INV	1	No structure				0				
wciQ	PF08660	1	INV	1	GT enhancer				1				A7, A37
wciR	PF04101	1	INV	1	D-Glc*p*A	β1-4	β-D-Glc*p*		1	8	1	0	A7, A37
wcyP				1	No structure				0				
wcwY				1	No structure				0				
wcxT	PF00535			1	No structure				0				

[a] Shown are the Pfam domains and CAZy families of the GT genes and the linkages the gene products are predicted to catalyze. GT genes are ordered according to their frequency. The Pfam and CAZy families assigned to their products are shown, along with the linkage(s) in the repeat unit structures that they are predicted to catalyze (where an assignment can be made). In some cases, different linkages are assigned to the same GT in different serotypes. The number of times the predicted linkage is found in the CPS of those serotypes where the GT gene is present is shown, as is the number of times the linkage is found in serotypes that do not have the GT gene. The notes in the rightmost column refer to Table S2 in the supplemental material, where the reasons for the GT assignments are discussed. INV, inverting; RET, retaining.

Downloaded from jb.asm.org at Imperial College London on July 10, 2009

WcjG transfers Gal*f*, but we are not confident in the latter CPS structure in the absence of more recent NMR data (8). Assignment of sugar specificity to the IT WcjH (see above) is based on a single monosaccharide difference between the two almost identical CPS structures of serotypes 29 and 35B. The second sugar in both of these structures is Gal*f* linked to the initial sugar by a β1-3 linkage. Within the *cps* loci where WchA is the IT, there are six serotypes where the second sugar is Gal*f* β1-3 linked to the initial sugar Glc*p* (17A, 20, 33F, 35A, 35B, and 34). Within the *cps* of these six serotypes, as well as serotype 29 (IT-WcjG) and serotypes 10F and 10A, the only putative GT gene in common is *wciB*, which in all cases is the sixth gene in the *cps* loci. Furthermore, *wciB* is not found in any other serotype where we have a different second sugar. We can therefore suggest that WciB transfers Gal*f* as the second sugar in the repeat units. Serotype 10F contains a single Gal*f*(β1-3) linkage to Gal*p*. Serotype 10A contains two Gal*f*(β1-3) linkages; however, one is a side branch (D-Gal*f*-(β1-3)-D-Gal*p*NAc) and not part of the circular main chain; therefore, we can suggest that Gal*f* is the second sugar and also propose with confidence that WcjG transfers Gal*p*.

Repeat unit polymerases (Wzy) and flippases (Wzx). The repeat unit polymerase (Wzy) and flippase (Wzx) define a polymerization/export process (Fig. 1), and in polysaccharide biosynthetic loci these genes almost invariably occur together (86). There is experimental evidence in *Y. enterocolitica* serotype O:8 that Wzy is strictly specific for the O unit to be polymerized, whereas Wzx has a relaxed specificity for the translocated polysaccharide (11). However, recently it has been reported that the Wzx flippases involved in biosynthesis of O antigen correlate with the first sugar of the O-specific lipopolysaccharide subunit (50) but otherwise seem to lack specificity at least in the situation where only one potential substrate is available.

There is enormous sequence divergence among the pneumococcal repeat unit polymerases (Wzy), but most have a best match to other Wzy proteins in a BLAST search and have about 10 putative transmembrane segments (12). The pneumococcal Wzy repeat unit polymerases fall into 40 HGs, which we refer to as Wzy-1 to Wzy-40. For their association with particular ITs, the putative linkages that they catalyze, and the serotypes in which they are present, see Table S1 in the supplemental material. The pneumococcal flippases (Wzx) are similarly divergent but are identified by their similarity to other Wzx proteins in BLAST searches and by having about 12 transmembrane segments (12). The pneumococcal Wzx flippases fall into 13 HGs (Wzx-1 to Wzx-13), although 40 *cps* loci encode a Wzx-1 (HG7) flippase, but in pneumococci there is no apparent association between the pneumococcal Wzx HGs and the IT HGs or the putative initial sugars (data not shown).

Following prediction of the polymerase linkage in the 52 known repeat unit structures that are synthesized by the Wzy-dependent pathway (see above), it can be seen that, in some cases, the same polymerase linkage is formed by members of different HGs (see Table S1 in the supplemental material). For example, the D-Glc*p*-(α1-3)-α-L-Rha*p* linkage is formed by Wzy-12 in types 6A and 6B and by Wzy-28 in type 19A, the D-Glc*p*-(β1-3)-α-D-Gal*p* linkage is formed by Wzy-14 in types 15A/15F and by Wzy-16 in types 17A/17F, and the D-Glc*p*-(β1-4)-β-D-Gal*f* linkage is formed by Wzy-6 in serogroup 18 and by Wzy-19 in type 23F. The IT is WchA, and Glc*p* is present in the CPS structures in all of these cases; therefore, we are very confident in the assignment of the above polymerization linkages.

Assignment of GTs. GTs catalyze the formation of glycosidic bonds between the lipid-linked glycan precursor as an acceptor and a nucleotide-activated sugar as a donor. Thus, GTs determine the sequence of components of the repeat units of polysaccharides and the characterization of the specificity of GTs would enable the prediction of the constituents of repeat units where these are unknown (Fig. 1) (35, 39).

Functional characterization of pneumococcal GTs is extremely limited, however, and has only been reported for types 8 and 14 (42, 69). The reaction specificities of the GTs may vary for the acceptor or donor monosaccharide residues and the linkage between them, and there are instances reported where a single nucleotide replacement can alter these specificities (41). For example, in pneumococci, a single amino acid difference in a GT has been reported to be responsible for the α-1,3 linkage of rhamnose to ribitol in serotype 6A but an α-1,4 linkage in serotype 6B (52). Prediction of the linkage catalyzed by a GT is therefore problematic and must be considered tentative until biochemical data are available. However, the sequences of the *cps* loci of all 88 *S. pneumoniae* serotypes that use the Wzy-dependent pathway, along with the 52 published CPS structures from these serotypes, present an opportunity to utilize comparative methods to tentatively assign ab initio the linkages within these structures catalyzed by the GTs (see Materials and Methods) and in many cases allow us to assign slightly different linkages to the same HG with some confidence.

Given the great diversity of sugars and sugar linkages among the pneumococcal CPS repeat units, it is not surprising that the GTs form a divergent group of 92 HGs (including three hypothetical proteins, WcwD, WcrT, and WcwX, which we predict to act as GTs), the majority of which fall into three Pfam families (PF00534, PF00535, and PF05704) (Table 2) . GTs have been classified into CAZy families based on high sequence similarity with one or more founding members of the family with experimentally demonstrated GT activity (20). Members of CAZy families appear to be either retaining GTs, forming glycosidic bonds with stereochemistry identical to that of the donor sugar (for example, UDP-glucose→α-glucoside) or inverting GTs which in forming the glycosidic linkage alter the stereochemistry of the donor sugar (for example, UDP-glucose→β-glucoside) (20, 46). Pfam families were highly correlated with CAZy families (Table 2), and CAZy family assignment proved a useful contributor to our putative functional assignments, as linkages that are inverting or retaining should be catalyzed by GTs that, from their CAZy family membership, are expected to be inverting or retaining.

The four main methods of assigning GTs to specific linkages are described in Materials and Methods. The linkages tentatively assigned to GTs are shown in Table 2; for the basis of these assignments, see the notes to Table S2 in the supplemental material. Using the GT-linkage correlation approach, the most prevalent GT gene in *cps* loci of serotypes whose CPS structures are known was *wchF* (14 occurrences). Among these 14 structures, there are 29 different triplet linkages in their repeat units but the only one in common is L-Rha*p*(β1-4)D-

Downloaded from jb.asm.org at Imperial College London on July 10, 2009

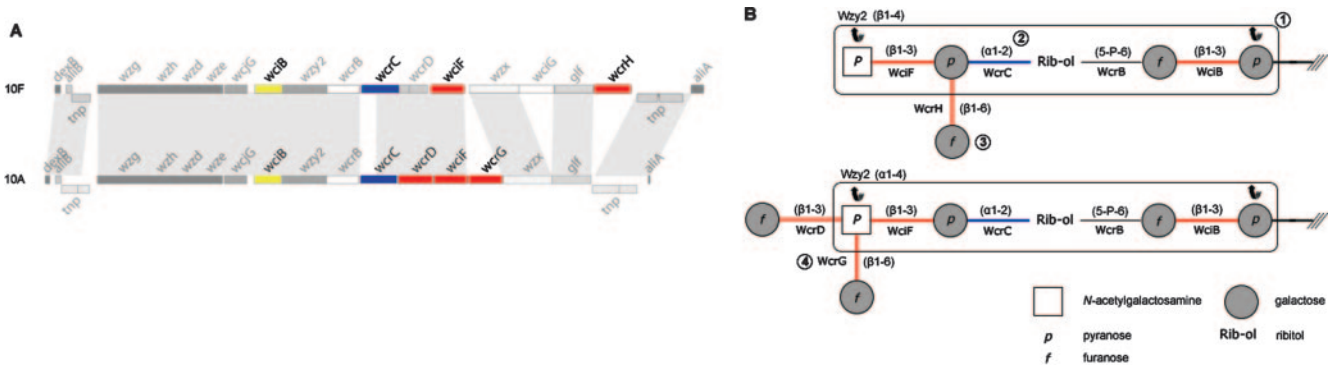


FIG. 5. Assignment of GT specificity. (A) *cps* loci of serotypes 10A and 10F are represented linearly with genes encoding GTs colored according to their CAZy membership—red for a GT with a predicted retaining mechanism, blue for an inverting mechanism, and yellow if the GT has not been assigned to a CAZy family. Shading between the two loci represents TBLASTX amino acid sequence similarity viewed by using ACT. (B) Structural representation of CPS repeat units. Glycosidic linkages are colored according to whether they require a retaining mechanism (red) or an inverting mechanism (blue). ///, lipid carrier. See the text for methods of GT assignment.

Glc*p*; as all 14 of the known structures contain this linkage, no other linkage in these structures correlates with the presence of *wchF*, and the above linkage is not found in any serotype that lacks *wchF*, we assign the synthesis of L-Rha*p*(β1-4)D-Glc*p* to WchF. There was also a perfect correlation between the presence of *wchK* (10 occurrences) and that of D-Gal*p*(β1-4)-β-D-Glc*p* and between the presence of *wchO* (8 occurrences) and that of D-ManpNAc(β1-4)-β-D-Glc*p*.

In other cases, GTs can be assigned on the basis of the predicted retaining or inverting status of the GTs encoded within the *cps* loci and the inverting and retaining linkages in the CPS. For example, the repeat units of serotypes 6A and 6B each contain two GT linkages, of which one is retaining and one is inverting. The *cps* loci each contain a single gene, *wciN*, encoding a retaining GT, which is therefore assigned to the retaining linkage, D-Gal*p*(α1-3)-α-D-Glc*p*, in both serotypes 6A and 6B. The inverting GT, WciP, can therefore be assigned to the inverting linkage, L-Rha*p*(α1-3)-D-Rib-ol in serotype 6A and L-Rha*p*(α1-4)-D-Rib-ol in serotype 6B, as also suggested by a completely different approach (52). Similarly, the repeat units of serogroup 15 each contain a single retaining linkage (D-Gal*p*-(α1-2)-β-D-Galp) and their *cps* loci each contain a single gene predicted to encode a retaining GT (WchN).

Gene content-linkage comparisons between similar serotypes can also be used to assign GT specificity. For example, the repeat unit structures of serotypes 19B and 19C differ only by the presence of a single D-Glc*p*(β1-6)β-D-Man*p*NAc side branch in 19C, which can be attributed to the single additional GT gene in the *cps* locus of this serotype (*wchU*; see the notes to Table S2 in the supplemental material). Similarly, the *cps* loci of serotypes 7F and 7A are syntenic and only differ in the presence of a frameshift mutation in the GT gene *wcwD* in serotype 7A. The CPS structures of both are known and differ only in the presence of an additional D-Gal*p*(β1-2)α-D-Gal*p* side branch in 7F, which allows us to attribute the side branch linkage to the action of WcwD.

Figure 5 depicts an example where a combination of methods can assign GT specificity and therefore propose a biosynthetic pathway for the repeat unit structures of serotypes 10F and 10A. Following the identification of the polymerase linkage and therefore the initial sugar (see above; Fig. 5B-1), the three GT genes *wciB*, *wcrC*, and *wciF* (and the ribitol phosphate transferase gene *wcrB*) are common to both *cps* loci (Fig. 5A) and presumably account for the three GT linkages (and the ribitol phosphate linkage, which is 5-P-5 linked in 10A but 5-P-6 linked in 10F; see below) in common in the main chain of the repeat units. The single retaining linkage in the structure, Gal*p*(α1-2)Rib-ol, can be assigned to the action of the single retaining GT, WcrC (Fig. 5B-2). WciB has been assigned as the Gal*f*(β1-3) transferase that adds the second sugar in the repeat unit through GT-linkage correlation, and so we can assign the action of WciF to the remaining D-Gal*p*NAc(β1-3)α-D-Gal*p* linkage in the main chain. The side branch linkage Gal*f*(β1-6)Gal*p* present in 10F but not in 10A can be accounted for by the presence in 10F, but not in 10A, of the GT gene *wcrH* (Fig. 5B-3). The two remaining side branch linkages in 10A can be attributed to the two additional GTs, encoded by *wcrD* and *wcrG* (Fig. 5B-4), which are not found or not intact (*wcrD* is frameshifted in type 10F) in the *cps* locus of 10F. The presence of *wcrD* in other serotypes correlates with a Gal*f*(β1-3) linkage, and therefore WcrD is assigned as catalyzing this linkage; WcrG, by a process of elimination, catalyzes the linkage of the other Gal*p*(β1-6) side branch.

Sometimes the evidence for GT specificity can be twofold. For example, the CPS structures of serogroup 11 contain three GT linkages, one inverting and two retaining. The *cps* loci each contain three GT genes which encode one inverting GT (WchK) and two retaining GTs (WcyK and WcrL). We can therefore assign WchK to the inverting linkage in the structure, and as we have already assigned this link to WchK based on GT-linkage correlation, our confidence in the assignment is further strengthened.

In some cases, the donor sugar appears to differ for GTs in the same HG. For example, WcxB uses two different sugars as donors, Gal*p*NAc in type 12A but Gal*p* in type 12F (12). Similarly, the donor sugar for WcrL is Glc*p*NAc in types 11F, 11B, and 11C but Glc*p* in type 11A, and WciU adds Glc*p* in types 18B, 18C, and 18F but Glc*p*NAc in type 18A (Table 2). There are also instances where the same linkage is present in the repeat units of CPS of different serogroups, but there is no common GT gene in the *cps* loci, suggesting that GTs of different HGs can catalyze the same linkage. For example, WciT

Downloaded from jb.asm.org at Imperial College London on July 10, 2009

of serotype 8 has been proposed to catalyze the D-Glc*p*-(α1-4)-α-D-Gal*p* linkage (69) and the same linkage is present in serotype 11A CPS, where we have suggested it is catalyzed by WcrL (Table 2; see the notes to Table S2 in the supplemental material). The donor sugar for WcrL is Glc*p* in type 11A but Glc*p*NAc in types 11F, 11B, and 11C.

Using the methods outlined above, we have been able to assign the GTs that catalyze 145 of the 200 glycosidic linkages in the available repeat unit structures, and in 32 of the 52 structures that use the Wzy-dependent pathway we have been able to assign all of the linkages (110 in total), with the remaining structures containing some linkages that could not be assigned to a specific GT (see Table S2 in the supplemental material) by any of the methods. In some cases, this was due to the presence of a greater number of GT genes than there are GT linkages in the repeat unit; for example, serotype 33B contains four GT linkages but the *cps* locus encodes five GTs. Conversely, there are instances where there are fewer GT genes in the *cps* locus than GT linkages in the repeat unit, for example, serotype 7B. In other cases, lack of assignment is due to ambiguous linkage correlation. For example, within the type 12F and 12A repeat units there are three unassigned retaining linkages and within the *cps* loci three remaining genes that encode retaining GTs, WcxD, WcxE, and WcxF. No CPS structures are available for the other serotypes that contain these genes and therefore there is no method of assigning their specificity.

For a few GTs, the assignments include linkages that are different and in some cases surprisingly so. These merit further study to establish if they are correct and, if so, the basis for the variation in linkage. For example, WcrV is credited with both a β(1-4) inverting linkage and an α(1-3) retaining linkage.

It has been suggested that the order in which GTs act to synthesize the repeat unit correlates with their order in the *cps* locus. This study allowed a more systematic examination of this suggestion and showed that the relationship held in 25/32 (78%) *cps* loci where we could assign all of the GTs (see Table S2 in the supplemental material), possibly suggesting that GTs might interact as part of a protein complex (22) that synthesizes the repeat unit. However, in this paper we chose not to use this correlation to assign GTs to linkages in those serotypes where this could not be done by any of the methods we have applied, and these unassigned GTs are in parentheses (see Fig. S1 in the supplemental material) to denote that we could not assign them to the linkages.

Assignment of sugar/polyalcohol phosphate transferases. Most GTs involved in CPS biosynthesis transfer a sugar from the NDP-linked precursor to the lipid-linked acceptor sugar, but there are several cases where there is a phosphodiester linkage and presumably a sugar phosphate is transferred. The polyalcohols Rib-ol, Ara-ol, Man-ol, Gro, and Cho are also incorporated into the repeat units via phosphodiester linkages. The *cps* loci in each of the above cases have a gene product that falls into one of three Pfam families (PF04464, PF04991, and PF01066), at least some of whose members are known to make phosphodiester linkages (Table 3). As most known CPS structures have a single phosphodiester linkage in the repeat unit and a single *cps* gene encoding a protein in one of the above Pfam families, we can often assign the sugar/polyalcohol transferase.

In some instances, phosphodiester linkages are present in the CPS structures but there were no *cps* gene products possessing the Pfam domains of sugar phosphate transferases. For example, the CPS structures of types 7B and 20 possess an α-Glcp-1P linkage and an α-GlcpNAc-1P linkage, respectively, and the *wcwK* gene is present in both *cps* loci, but neither WcwK nor any other *cps* gene product contains one of the above sugar phosphate transferase Pfam domains. The conserved hypothetical protein WcwK is also present in types 7C, 40, and 21, but their CPS structures are unknown. WcwK of type 21 is 95% similar to WefC encoded by the streptococcal receptor polysaccharide gene cluster of *S. gordonii*, which has been proposed to be the α-Gal-1P transferase that forms the phosphodiester linkage in the middle of the repeat unit (89). Therefore, we suggest that WcwK may function as a α-Glcp-1P or α-GlcpNAc-1P transferase.

There are two phosphodiester linkages in serogroup 32 CPS, the Cho-P-3-β-L-Rha*p* linkage and the α-L-Rha*p*-(1-P-2)-α-D-Gal*p* linkage. The incorporation of Cho is proposed to be catalyzed by WcrN, as in serotype 27, with amino acid differences in this protein accounting for the P-3 linkage in serogroup 32 compared to the P-2 linkage in serotype 27. There are no other gene products in serogroup 32 with the Pfam domains PF04464, PF04991, and PF01066, but there is a conserved hypothetical protein (WcyI) which could function as the second sugar phosphate transferase for the formation of the α-L-Rha*p*-(1-P-2)-α-D-Gal*p* linkage. The assignments of the sugar/polyalcohol phosphate transferases to the phosphodiester linkages in the CPS structures are summarized in Table 3.

Assignment of acetyltransferases. In this section, we consider only O-acetyl groups as, in general, *N*-acetyl sugars are transferred from NDP-*N*-acetyl precursors, each of which has a defined pathway (Fig. 2), whereas O-acetyl groups are commonly added to sugars by transferases independent of the NDP-sugar pathway. Several studies indicate that the O-acetyl groups in pneumococcal repeat units can be immunologically important, for example, in serotypes 1, 11A, 15F, and 34, whereas they appear not to influence the antibody response in other cases, for example, serotypes 9V and 20 (38, 53).

There are 78 putative acetyltransferases in the pneumococcal *cps* loci distributed into two Pfam families, PF01757 and PF00132 (Table 1). O-acetyl groups are present in many pneumococcal repeat units, but the patterns of acetylation differ considerably, and also the site or the percentage of acetylation for several structures has been revised in recent NMR studies (36, 45). There is no obvious correlation between the acetyltransferase HG and the pattern of acetylation (data not shown). Frameshifted, inserted, or deleted acetyltransferase genes are also often present in the pneumococcal *cps* loci, and frameshifting arising from variation in the number of TA repeats in the acetyltransferase gene *wciZ* has been shown to lead to frequent serotype switching between types 15B and 15C (79).

Assignment of pyruvyltransferases. Serotypes 4 and 27 have pyruvate moieties in their CPS (2,3-pyruvate and 4,6-pyruvate, respectively, in the S stereo configuration), and marked changes in immunological specificity have been observed by removing these moieties from the repeat units (38). Only two genes in the *S. pneumoniae cps* loci (Table 1) encode products, of different HGs, that possess the Pfam domain PF04230

Downloaded from jb.asm.org at Imperial College London on July 10, 2009

TABLE 3. Distribution of sugar/polyalcohol phosphate transferases into Pfam families and HGs and occurrence in serotypes and assignment of the putative linkages that they catalyze

Pfam family and gene	Gene frequency	Donor	Linkage	Acceptor	Serotype(s)
PF04991					
wcxG	5	D-Ara-ol	1-P-3	β-L-Rha*p*	17F
		No structure			24A, 24B, 24F, 48
wcrO	4	D-Rib-ol	5-P-3	β-D-Gal*f*	34
		No structure			33C, 35F, 36
whaI	4	No structure			39, 43, 47A, 47F
wcrB	4	D-Rib-ol	5-P-5	β-D-Gal*f*	10A
		D-Rib-ol	5-P-6	β-D-Gal*f*	10F
		No structure			10B, 10C
wchP	4	L-Rha*p*	1-P-4	β-D-Man*p*NAc	19A, 19B, 19C, 19F
wcxK	3	No structure			24A, 24B, 24F
wcrN	3	Cho	P-2	β-L-Rha*p*	27
		Cho	P-3	β-L-Rha*p*	32A, 32F
wcxQ	3	No structure			16F, 28A, 28F
wcxR	1	No structure			16A
wcyR	1	Gro	1-P-6	β-D-Glc*p*NAc	45
PF04464					
wchX	7	Gro	2-P-3	β-D-Gal*p*	15A, 15B, 15C, 15F, 23F
		No structure			23A, 23B
wcwU	5	Gro	1-P-4	α-D-Glc*p*	11A
		Gro	1-P-4	α-D-Glc*p*NAc	11C
		Rib-ol	1-P-4	α-D-Glc*p*NAc	11B, 11F
		No structure			11D
wciY	5	No linkage			14
		Gro	1-P-3	β-D-Gal*p*	18A, 18B, 18C, 18F
wcrJ	5	Rib-ol	5-P-4	β-D-Gal*p*NAc	29, 35B
		Man-ol	6-P-3	β-D-Gal*p*	35A
		No structure			35C, 42
wcxP	3	No structure			16F, 28A, 28F
whaG	1	Rib-ol	5-P-4	β-D-Gal*p*	13
PF01066					
wciO	4	Rib-ol	5-P-2	α-D-Gal*p*	6A, 6B
		Rib-ol	5-P-6	β-D-Glc*p*	33B
		No structure			33D
No Pfam hit					
wcwK	5	D-Glcp	1-P-6	α-D-Glc*p*	7B
		D-Glc*p*NAc	1-P-6	α-D-Glc*p*	20
		No structure			7C, 21, 40
wcyI	1	L-Rha*p*	1-P-2	*a*-D-Gal*p*	32A, 32F

(WciM and WhaL of types 4 and 27, respectively), which is found in other pyruvyltransferases involved in bacterial polysaccharide biosynthesis, e.g., CsaB in *Bacillus anthracis*, WcaK in *E. coli*, and AmsJ in *Erwinia amylovora*. It seems clear that WciM and WhaL are the expected pyruvyltransferases for type 4 and 27 CPS biosynthesis, respectively.

Transposases and similar elements within the *cps* loci. IS elements or IS-like sequences and several other transposable

Downloaded from jb.asm.org at Imperial College London on July 10, 2009

elements in the form of conjugative transposons have been previously characterized in pneumococci, but their role in population dynamics is still unclear (62, 71). They constitute a relatively large proportion (>3.5%) of the *S. pneumoniae* TIGR4 genome, where two full-length group II introns and a fragment of the streptococcal conjugative plasmid Tn*5252* are also present (75). In the *cps* loci studied here, IS*630*, IS*1167*, IS*1202*, IS*1381*, IS*1670*, and IS*1671* elements were identified (data not shown). Group II intron genes (Table 1) are also present in the 5′ region of serotype 19F and in the 3′ region of type 25A, 25F, and 38 *cps* loci. The presence of these elements at the 5′ end and often at the 3′ end of the *cps* locus suggests that they may (or did) have a role in the horizontal transfer of *cps* loci. However, in the few cases where recent changes of serotype have been examined this has occurred by homologous recombination, usually with crossover points in genes at the ends of, or flanking, the *cps* locus (19).

Concluding remarks. Biochemical studies of CPS biosynthesis in *S. pneumoniae* have been limited to serotypes 2, 3, 8, 14, and 37 (15, 16, 29, 42, 43, 47, 65, 69); thus, the vast majority of *cps* gene products have had to be assigned putative functions in the absence of experimental data. Assignment of function has been achieved by using bioinformatic procedures that group similar *cps* gene products into HGs, by seeking homologs with established functions in the databases, and by identifying domains shared by other proteins (Pfam families). The availability of the full set of *cps* sequences also allows the presence of particular HGs, members of Pfam families and CAZy families, to be correlated with the structures of the CPS repeat units, which can be particularly effective when comparing *cps* loci that are very closely related (51).

The assignment of a general function for *cps* gene products works well, and there were only 26 of the 1,999 gene products that had no significant database matches or which only matched other proteins of unknown function. The process of assigning specific reactions to the *cps* gene products also works well in many cases, and the products of biosynthetic pathway genes and most of the ITs, polymerases, sugar phosphate transferases, and pyruvyltransferases could be tentatively assigned specific functions by correlating gene distribution with repeat unit structures. The assignment of the linkages catalyzed by the large number of GTs is more problematic. Rather than simply making predictions based on database matches, we have attempted to develop a rational ab initio approach (see the notes to Table S2 in the supplemental material for the basis for the predicted assignments shown in Table 2) as a basis for stimulating further biochemical work that establishes the specific reactions catalyzed by these enzymes.

Assignments of GT specificity by correlation have a number of limitations. For example, they assume that the published CPS structures are correct, which in some instances may not be the case (for example, type 31), particularly with some of the older structures where contamination with other cell wall polysaccharides could have been a problem. Also, correlation of GTs and linkages relies on the assumption that members of a single CAZy family of GTs are invariably all retaining or inverting. See the supplemental material for outlines of cases where there are inconsistencies in functional assignments of GTs, which may reflect incorrect structures, incorrect prediction of mechanism from CAZy family membership, or unsuspected variation in the linkage catalyzed by different members of the same HG.

Our analysis highlights some areas of pneumococcal CPS biosynthesis where functional analysis is required or where efforts could be focused. Besides analysis of the specificity of the large numbers of GTs, these include studies of the biochemical reactions catalyzed by the putative GalE-like epimerase, Spr1460, and verification of the proposed AAT-Gal*p* biosynthetic pathway. It would also be valuable to examine whether deletion of the chromosomal gene encoding the pneumococcal WbgY homolog (Spr1655) results in loss of the ability to produce the serotype 1 capsule, to determine whether the substrate of the IT, WciI, is Glc*p*NAc-P/Gal*p*NAc-P and whether or not the substrate of the IT, WcjG, is Gal*p*/Gal*f*. Additionally, it would be valuable to determine the CPS structures for some of the 36 serotypes where these are unavailable, to reanalyze some of the older structures where the predicted functions of *cps* gene products do not correlate well with the existing structure (for example, serotype 31), and to reassess the presence of Cho-P in the CPS of serotypes 15F and 15C.

ACKNOWLEDGMENTS

This work was funded by the Wellcome Trust. B.G.S. is a Wellcome Trust Principal Research Fellow.

REFERENCES

1. **Abbott, J. C., D. M. Aanensen, K. Rutherford, S. Butcher, and B. G. Spratt.** 2005. WebACT—an online companion for the Artemis comparison tool. Bioinformatics **21:**3665–3666.
2. **Akana, J., A. A. Fedorov, E. Fedorov, W. R. Novak, P. C. Babbitt, S. C. Almo, and J. A. Gerlt.** 2006. D-Ribulose 5-phosphate 3-epimerase: functional and structural relationships to members of the ribulose-phosphate binding (beta/alpha)8-barrel superfamily. Biochemistry **45:**2493–2503.
3. **Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman.** 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25:**3389–3402.
4. **Amer, A. O., and M. A. Valvano.** 2002. Conserved aspartic acids are essential for the enzymic activity of the WecA protein initiating the biosynthesis of O-specific lipopolysaccharide and enterobacterial common antigen in *Escherichia coli*. Microbiology **148:**571–582.
5. **Apweiler, R., T. K. Attwood, A. Bairoch, A. Bateman, E. Birney, M. Biswas, P. Bucher, L. Cerutti, F. Corpet, M. D. Croning, R. Durbin, L. Falquet, W. Fleischmann, J. Gouzy, H. Hermjakob, N. Hulo, I. Jonassen, D. Kahn, A. Kanapin, Y. Karavidopoulou, R. López, B. Marx, N. J. Mulder, T. M. Oinn, M. Pagni, F. Servant, C. J. Sigrist, and E. M. Zdobnov.** 2000. InterPro—an integrated documentation resource for protein families, domains and functional sites. Bioinformatics **16:**1145–1150.
6. **Arrecubieta, C., E. García, and R. López.** 1995. Sequence and transcriptional analysis of a DNA region involved in the production of capsular polysaccharide in *Streptococcus pneumoniae* type 3. Gene **167:**1–7.
7. **Arrecubieta, C., R. López, and E. García.** 1996. Type 3-specific synthase of *Streptococcus pneumoniae* (Cap3B) directs type 3 polysaccharide biosynthesis in *Escherichia coli* and in pneumococcal strains of different serotypes. J. Exp. Med. **184:**449–455.
8. **Batavyal, L., and N. Roy.** 1983. Structure of the capsular polysaccharide of *Diplococcus pneumoniae* type 31. Carbohydr. Res. **119:**300–302.
9. **Behr, T., W. Fischer, J. Peter-Katalinic, and H. Egge.** 1992. The structure of pneumococcal lipoteichoic acid. Improved preparation, chemical and mass spectrometric studies. Eur. J. Biochem. **207:**1063–1075.
10. **Bender, M. H., R. T. Cartee, and J. Yother.** 2003. Positive correlation between tyrosine phosphorylation of CpsD and capsular polysaccharide production in *Streptococcus pneumoniae*. J. Bacteriol. **185:**6057–6066.
11. **Bengoechea, J. A., E. Pinta, T. Salminen, C. Oertelt, O. Holst, J. Radziejewska-Lebrecht, Z. Piotrowska-Seget, R. Venho, and M. Skurnik.** 2002. Functional characterization of Gne (UDP-*N*-acetylglucosamine-4-epimerase), Wzz (chain length determinant), and Wzy (O-antigen polymerase) of *Yersinia enterocolitica* serotype O:8. J. Bacteriol. **184:**4277–4287.
12. **Bentley, S. D., D. M. Aanensen, A. Mavroidi, D. Saunders, E. Rabbinowitsch, M. Collins, K. Donohoe, D. Harris, L. Murphy, M. A. Quail, G. Samuel, I. C. Skovsted, M. S. Kaltoft, B. Barrell, P. R. Reeves, J. Parkhill, and B. G. Spratt.** 2006. Genetic analysis of the capsular biosynthetic locus from all 90 pneumococcal serotypes. PLoS Genet. **2:**e31.
13. **Bergstrom, N., P. E. Jansson, M. Kilian, and U. B. Skov Sorensen.** 2000.

Downloaded from jb.asm.org at Imperial College London on July 10, 2009

Structures of two cell wall-associated polysaccharides of a *Streptococcus mitis* biovar 1 strain. A unique teichoic acid-like polysaccharide and the group O antigen which is a C-polysaccharide in common with pneumococci. Eur. J. Biochem. **267:**7147–7157.
14. **Bonofiglio, L., E. García, and M. Mollerach.** 2005. Biochemical characterization of the pneumococcal glucose 1-phosphate uridylyltransferase (GalU) essential for capsule biosynthesis. Curr. Microbiol. **51:**217–221.
15. **Cartee, R. T., W. T. Forsee, M. H. Bender, K. D. Ambrose, and J. Yother.** 2005. CpsE from type 2 *Streptococcus pneumoniae* catalyzes the reversible addition of glucose-1-phosphate to a polyprenyl phosphate acceptor, initiating type 2 capsule repeat unit formation. J. Bacteriol. **187:**7425–7433.
16. **Cartee, R. T., W. T. Forsee, and J. Yother.** 2005. Initiation and synthesis of the *Streptococcus pneumoniae* type 3 capsule on a phosphatidylglycerol membrane anchor. J. Bacteriol. **187:**4470–4479.
17. **Carver, T. J., K. M. Rutherford, M. Berriman, M. A. Rajandream, B. G. Barrell, and J. Parkhill.** 2005. ACT: the Artemis comparison tool. Bioinformatics **21:**3422–3423.
18. **Caspi, R., H. Foerster, C. A. Fulcher, R. Hopkinson, J. Ingraham, P. Kaipa, M. Krummenacker, S. Paley, J. Pick, S. Y. Rhee, C. Tissier, P. Zhang, and P. D. Karp.** 2006. MetaCyc: a multiorganism database of metabolic pathways and enzymes. Nucleic Acids Res. **34:**D511–D516.
19. **Coffey, T. J., M. C. Enright, M. Daniels, J. K. Morona, R. Morona, W. Hryniewicz, J. C. Paton, and B. G. Spratt.** 1998. Recombinational exchanges at the capsular polysaccharide biosynthetic locus lead to frequent serotype changes among natural isolates of *Streptococcus pneumoniae*. Mol. Microbiol. **27:**73–83.
20. **Coutinho, P. M., E. Deleury, G. J. Davies, and B. Henrissat.** 2003. An evolving hierarchical family classification for glycosyltransferases. J. Mol. Biol. **328:**307–317.
21. **Coyne, M. J., A. O. Tzianabos, B. C. Mallory, V. J. Carey, D. L. Kasper, and L. E. Comstock.** 2001. Polysaccharide biosynthesis locus required for virulence of *Bacteroides fragilis*. Infect. Immun. **69:**4342–4350.
22. **Dandekar, T., B. Snel, M. Huynen, and P. Bork.** 1998. Conservation of gene order: a fingerprint of proteins that physically interact. Trends Biochem. Sci. **23:**324–328.
23. **Dmitriev, B. A., V. Lvov, N. V. Tochtamysheva, A. S. Shashkov, N. K. Kochetkov, B. Jann, and K. Jann.** 1983. Cell-wall lipopolysaccharide of *Escherichia coli* 0114:H2. Structure of the polysaccharide chain. Eur. J. Biochem. **134:**517–521.
24. **Feng, L., W. Wang, J. Tao, H. Guo, G. Krause, L. Beutin, and L. Wang.** 2004. Identification of *Escherichia coli* O114 O-antigen gene cluster and development of an O114 serogroup-specific PCR assay. J. Clin. Microbiol. **42:**3799–3804.
25. **Finn, R. D., J. Mistry, B. Schuster-Bockler, S. Griffiths-Jones, V. Hollich, T. Lassmann, S. Moxon, M. Marshall, A. Khanna, R. Durbin, S. R. Eddy, E. L. Sonnhammer, and A. Bateman.** 2006. Pfam: clans, web tools and services. Nucleic Acids Res. **34:**D247–D251.
26. **Forsee, W. T., R. T. Cartee, and J. Yother.** 2006. Role of the carbohydrate binding site of the *Streptococcus pneumoniae* capsular polysaccharide type 3 synthase in the transition from oligosaccharide to polysaccharide synthesis. J. Biol. Chem. **281:**6283–6289.
27. **Frey, P. A.** 1996. The Leloir pathway: a mechanistic imperative for three enzymes to change the stereochemical configuration of a single carbon in galactose. FASEB J. **10:**461–470.
28. **Guo, H., L. Li, and P. G. Wang.** 2006. Biochemical characterization of UDP-GlcNAc/Glc 4-Epimerase from *Escherichia coli* O86:B7. Biochemistry **45:**13760–13768.
29. **Hardy, G. G., M. J. Caimano, and J. Yother.** 2000. Capsule biosynthesis and basic metabolism in *Streptococcus pneumoniae* are linked through the cellular phosphoglucomutase. J. Bacteriol. **182:**1854–1863.
30. **Hardy, G. G., A. D. Magee, C. L. Ventura, M. J. Caimano, and J. Yother.** 2001. Essential role for cellular phosphoglucomutase in virulence of type 3 *Streptococcus pneumoniae*. Infect. Immun. **69:**2309–2317.
31. **Henrichsen, J.** 1995. Six newly recognized types of *Streptococcus pneumoniae*. J. Clin. Microbiol. **33:**2759–2762.
32. **Hoskins, J., W. E. Alborn, Jr., J. Arnold, L. C. Blaszczak, S. Burgett, B. S. DeHoff, S. T. Estrem, L. Fritz, D. J. Fu, W. Fuller, C. Geringer, R. Gilmour, J. S. Glass, H. Khoja, A. R. Kraft, R. E. Lagace, D. J. LeBlanc, L. N. Lee, E. J. Lefkowitz, J. Lu, P. Matsushima, S. M. McAhren, M. McHenney, K. McLeaster, C. W. Mundy, T. I. Nicas, F. H. Norris, M. O'Gara, R. B. Peery, G. T. Robertson, P. Rockey, P. M. Sun, M. E. Winkler, Y. Yang, M. Young-Bellido, G. Zhao, C. A. Zook, R. H. Baltz, S. R. Jaskunas, P. R. Rosteck, Jr., P. L. Skatrud, and J. I. Glass.** 2001. Genome of the bacterium *Streptococcus pneumoniae* strain R6. J. Bacteriol. **183:**5709–5717.
33. **Ishiyama, N., C. Creuzenet, J. S. Lam, and A. M. Berghuis.** 2004. Crystal structure of WbpP, a genuine UDP-*N*-acetylglucosamine 4-epimerase from *Pseudomonas aeruginosa*: substrate specificity in UDP-hexose 4-epimerases. J. Biol. Chem. **279:**22635–22642.
34. **Jiang, S. M., L. Wang, and P. R. Reeves.** 2001. Molecular characterization of *Streptococcus pneumoniae* type 4, 6B, 8, and 18C capsular polysaccharide gene clusters. Infect. Immun. **69:**1244–1255.
35. **Jolly, L., J. Newell, I. Porcelli, S. J. Vincent, and F. Stingele.** 2002. *Lactobacillus helveticus* glycosyltransferases: from genes to carbohydrate synthesis. Glycobiology **12:**319–327.
36. **Jones, C., and X. Lemercinier.** 2005. Full NMR assignment and revised structure for the capsular polysaccharide from *Streptococcus pneumoniae* type 15B. Carbohydr. Res. **340:**403–409.
37. **Jones, C., C. Whitley, and X. Lemercinier.** 2000. Full assignment of the proton and carbon NMR spectra and revised structure for the capsular polysaccharide from *Streptococcus pneumoniae* type 17F. Carbohydr. Res. **325:**192–201.
38. **Kamerling, J. P.** 2000. Pneumococcal polysaccharides: a chemical view, p. 81–114. *In* A. Tomasz (ed.), *Streptococcus pneumoniae*: molecular biology and mechanisms of disease. Mary Ann Liebert Inc., Larchmont, NY.
39. **Kanehisa, M., S. Goto, M. Hattori, K. F. Aoki-Kinoshita, M. Itoh, S. Kawashima, T. Katayama, M. Araki, and M. Hirakawa.** 2006. From genomics to chemical genomics: new developments in KEGG. Nucleic Acids Res. **34:**D354–D357.
40. **Karlsson, C., P. E. Jansson, and U. B. Skov Sorensen.** 1999. The pneumococcal common antigen C-polysaccharide occurs in different forms: mono-substituted or di-substituted with phosphocholine. Eur. J. Biochem. **265:**1091–1097.
41. **Kawano, S., K. Hashimoto, T. Miyama, S. Goto, and M. Kanehisa.** 2005. Prediction of glycan structures from gene expression data based on glycosyltransferase reactions. Bioinformatics **21:**3976–3982.
42. **Kolkman, M. A., B. A. van der Zeijst, and P. J. Nuijten.** 1997. Functional analysis of glycosyltransferases encoded by the capsular polysaccharide biosynthesis locus of *Streptococcus pneumoniae* serotype 14. J. Biol. Chem. **272:**19502–19508.
43. **Kolkman, M. A., W. Wakarchuk, P. J. Nuijten, and B. A. van der Zeijst.** 1997. Capsular polysaccharide synthesis in *Streptococcus pneumoniae* serotype 14: molecular analysis of the complete *cps* locus and identification of genes encoding glycosyltransferases required for the biosynthesis of the tetrasaccharide subunit. Mol. Microbiol. **26:**197–208.
44. **Lazarevic, V., F. X. Abellan, S. B. Moller, D. Karamata, and C. Mauel.** 2002. Comparison of ribitol and glycerol teichoic acid genes in *Bacillus subtilis* W23 and 168: identical function, similar divergent organization, but different regulation. Microbiology **148:**815–824.
45. **Lemercinier, X., and C. Jones.** 2006. Full assignment of the ^{1}H and ^{13}C spectra and revision of the O-acetylation site of the capsular polysaccharide of *Streptococcus pneumoniae* type 33F, a component of the current pneumococcal polysaccharide vaccine. Carbohydr. Res. **341:**68–74.
46. **Liu, J., and A. Mushegian.** 2003. Three monophyletic superfamilies account for the majority of the known glycosyltransferases. Protein Sci. **12:**1418–1431.
47. **Llull, D., E. García, and R. López.** 2001. Tts, a processive beta-glucosyltransferase of *Streptococcus pneumoniae*, directs the synthesis of the branched type 37 capsular polysaccharide in Pneumococcus and other gram-positive species. J. Biol. Chem. **276:**21053–21061.
48. **Llull, D., R. Muñoz, R. López, and E. García.** 1999. A single gene (*tts*) located outside the *cap* locus directs the formation of *Streptococcus pneumoniae* type 37 capsular polysaccharide. Type 37 pneumococci are natural, genetically binary strains. J. Exp. Med. **190:**241–251.
49. **López, R., and E. García.** 2004. Recent trends on the molecular biology of pneumococcal capsules, lytic enzymes, and bacteriophage. FEMS Microbiol. Rev. **28:**553–580.
50. **Marolda, C. L., J. Vicarioli, and M. A. Valvano.** 2004. Wzx proteins involved in biosynthesis of O antigen function in association with the first sugar of the O-specific lipopolysaccharide subunit. Microbiology **150:**4095–4105.
51. **Mavroidi, A., D. M. Aanensen, D. Godoy, I. C. Skovsted, M. S. Kaltoft, P. R. Reeves, S. D. Bentley, and B. G. Spratt.** 2007. Genetic relatedness of the *Streptococcus pneumoniae* capsular biosynthetic loci. J. Bacteriol. **189:**7841–7855.
52. **Mavroidi, A., D. Godoy, D. M. Aanensen, D. A. Robinson, S. K. Hollingshead, and B. G. Spratt.** 2004. Evolutionary genetics of the capsular locus of serogroup 6 pneumococci. J. Bacteriol. **186:**8181–8192.
53. **McNeely, T. B., J. M. Staub, C. M. Rusk, M. J. Blum, and J. J. Donnelly.** 1998. Antibody responses to capsular polysaccharide backbone and O-acetate side groups of *Streptococcus pneumoniae* type 9V in humans and rhesus macaques. Infect. Immun. **66:**3705–3710.
54. **Mollerach, M., R. López, and E. García.** 1998. Characterization of the galU gene of *Streptococcus pneumoniae* encoding a uridine diphosphoglucose pyrophosphorylase: a gene essential for capsular polysaccharide biosynthesis. J. Exp. Med. **188:**2047–2056.
55. **Morona, J. K., D. C. Miller, T. J. Coffey, C. J. Vindurampulle, B. G. Spratt, R. Morona, and J. C. Paton.** 1999. Molecular and genetic characterization of the capsule biosynthesis locus of *Streptococcus pneumoniae* type 23F. Microbiology **145:**781–789.
56. **Morona, J. K., D. C. Miller, R. Morona, and J. C. Paton.** 2004. The effect that mutations in the conserved capsular polysaccharide biosynthesis genes *cpsA*, *cpsB*, and *cpsD* have on virulence of *Streptococcus pneumoniae*. J. Infect. Dis. **189:**1905–1913.
57. **Morona, J. K., R. Morona, and J. C. Paton.** 1999. Comparative genetics of

Downloaded from jb.asm.org at Imperial College London on July 10, 2009

capsular polysaccharide biosynthesis in *Streptococcus pneumoniae* types belonging to serogroup 19. J. Bacteriol. **181:**5355–5364.
58. **Morona, J. K., R. Morona, and J. C. Paton.** 2006. Attachment of capsular polysaccharide to the cell wall of *Streptococcus pneumoniae* type 2 is required for invasive disease. Proc. Natl. Acad. Sci. USA **103:**8505–8510.
59. **Morona, J. K., J. C. Paton, D. C. Miller, and R. Morona.** 2000. Tyrosine phosphorylation of CpsD negatively regulates capsular polysaccharide biosynthesis in *Streptococcus pneumoniae*. Mol. Microbiol. **35:**1431–1442.
60. **Mulrooney, E. F., K. K. Poon, D. J. McNally, J. R. Brisson, and J. S. Lam.** 2005. Biosynthesis of UDP-*N*-acetyl-L-fucosamine, a precursor to the biosynthesis of lipopolysaccharide in *Pseudomonas aeruginosa* serotype O11. J. Biol. Chem. **280:**19535–19542.
61. **Muñoz, R., R. López, M. de Frutos, and E. García.** 1999. First molecular characterization of a uridine diphosphate galacturonate 4-epimerase: an enzyme required for capsular biosynthesis in *Streptococcus pneumoniae* type 1. Mol. Microbiol. **31:**703–713.
62. **Muñoz, R., R. López, and E. García.** 1998. Characterization of IS1515, a functional insertion sequence in *Streptococcus pneumoniae*. J. Bacteriol. **180:** 1381–1388.
63. **Muñoz, R., M. Mollerach, R. López, and E. García.** 1997. Molecular organization of the genes required for the synthesis of type 1 capsular polysaccharide of *Streptococcus pneumoniae*: formation of binary encapsulated pneumococci and identification of cryptic dTDP-rhamnose biosynthesis genes. Mol. Microbiol. **25:**79–92.
64. **Park, I. H., D. G. Pritchard, R. Cartee, A. Brandao, M. C. Brandileone, and M. H. Nahm.** 2007. Discovery of a new capsular serotype (6C) within serogroup 6 of *Streptococcus pneumoniae*. J. Clin. Microbiol. **207:**1225–1233.
65. **Pelosi, L., M. Boumedienne, N. Saksouk, J. Geiselmann, and R. A. Geremia.** 2005. The glucosyl-1-phosphate transferase WchA (Cap8E) primes the capsular polysaccharide repeat unit biosynthesis of *Streptococcus pneumoniae* serotype 8. Biochem. Biophys. Res. Commun. **327:**857–865.
66. **Price, N. P., and F. A. Momany.** 2005. Modeling bacterial UDP-HexNAc: polyprenol-P HexNAc-1-P transferases. Glycobiology **15:**29R–42R.
67. **Ramirez, M., and A. Tomasz.** 1998. Molecular characterization of the complete 23F capsular polysaccharide locus of *Streptococcus pneumoniae*. J. Bacteriol. **180:**5273–5278.
68. **Reeves, P. R., M. Hobbs, M. A. Valvano, M. Skurnik, C. Whitfield, D. Coplin, N. Kido, J. Klena, D. Maskell, C. R. Raetz, and P. D. Rick.** 1996. Bacterial polysaccharide synthesis and gene nomenclature. Trends Microbiol. **4:**495–503.
69. **Saksouk, N., L. Pelosi, P. Colin-Morel, M. Boumedienne, P. L. Abdian, and R. A. Geremia.** 2005. The capsular polysaccharide biosynthesis of *Streptococcus pneumoniae* serotype 8: functional identification of the glycosyltransferase WciS (Cap8H). Biochem. J. **389:**63–72.
70. **Samuel, G., and P. Reeves.** 2003. Biosynthesis of O-antigens: genes and pathways involved in nucleotide sugar precursor synthesis and O-antigen assembly. Carbohydr. Res. **338:**2503–2519.
71. **Sánchez-Beato, A. R., E. García, R. López, and J. L. García.** 1997. Identification and characterization of IS*1381*, a new insertion sequence in *Streptococcus pneumoniae*. J. Bacteriol. **179:**2459–2463.
72. **Shepherd, J. G., L. Wang, and P. R. Reeves.** 2000. Comparison of O-antigen gene clusters of *Escherichia coli* (*Shigella*) sonnei and *Plesiomonas shigelloides* O17: sonnei gained its current plasmid-borne O-antigen genes from *P. shigelloides* in a recent event. Infect. Immun. **68:**6056–6061.
73. **Sohlenkamp, C., I. M. López-Lara, and O. Geiger.** 2003. Biosynthesis of phosphatidylcholine in bacteria. Prog. Lipid Res. **42:**115–162.
74. **Sulzenbacher, G., L. Gal, C. Peneff, F. Fassy, and Y. Bourne.** 2001. Crystal structure of *Streptococcus pneumoniae N*-acetylglucosamine-1-phosphate uridyltransferase bound to acetyl-coenzyme A reveals a novel active site architecture. J. Biol. Chem. **276:**11844–11851.
75. **Tettelin, H., K. E. Nelson, I. T. Paulsen, J. A. Eisen, T. D. Read, S. Peterson, J. Heidelberg, R. T. DeBoy, D. H. Haft, R. J. Dodson, A. S. Durkin, M. Gwinn, J. F. Kolonay, W. C. Nelson, J. D. Peterson, L. A. Umayam, O. White, S. L. Salzberg, M. R. Lewis, D. Radune, E. Holtzapple, H. Khouri, A. M. Wolf, T. R. Utterback, C. L. Hansen, L. A. McDonald, T. V. Feldblyum, S. Angiuoli, T. Dickinson, E. K. Hickey, I. E. Holt, B. J. Loftus, F. Yang, H. O. Smith, J. C. Venter, B. A. Dougherty, D. A. Morrison, S. K. Hollingshead, and C. M. Fraser.** 2001. Complete genome sequence of a virulent isolate of *Streptococcus pneumoniae*. Science **293:**498–506.
76. **Thompson, J. D., T. J. Gibson, F. Plewniak, F. Jeanmougin, and D. G. Higgins.** 1997. The CLUSTAL_X Windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res. **25:**4876–4882.
77. **Valvano, M. A.** 2003. Export of O-specific lipopolysaccharide. Front. Biosci. **8:**s452–s471.
78. **van Selm, S., M. A. Kolkman, B. A. van der Zeijst, K. A. Zwaagstra, W. Gaastra, and J. P. van Putten.** 2002. Organization and characterization of the capsule biosynthesis locus of *Streptococcus pneumoniae* serotype 9V. Microbiology **148:**1747–1755.
79. **van Selm, S., L. M. van Cann, M. A. Kolkman, B. A. van der Zeijst, and J. P. van Putten.** 2003. Genetic basis for the structural difference between *Streptococcus pneumoniae* serotype 15B and 15C capsular polysaccharides. Infect. Immun. **71:**6192–6198.
80. **Ventura, C. L., R. T. Cartee, W. T. Forsee, and J. Yother.** 2006. Control of capsular polysaccharide chain length by UDP-sugar substrate concentrations in *Streptococcus pneumoniae*. Mol. Microbiol. **61:**723–733.
81. **Waite, R. D., D. W. Penfold, J. K. Struthers, and C. G. Dowson.** 2003. Spontaneous sequence duplications within capsule genes *cap8E* and *tts* control phase variation in *Streptococcus pneumoniae* serotypes 8 and 37. Microbiology **149:**497–504.
82. **Waite, R. D., J. K. Struthers, and C. G. Dowson.** 2001. Spontaneous sequence duplication within an open reading frame of the pneumococcal type 3 capsule locus causes high-frequency phase variation. Mol. Microbiol. **42:** 1223–1232.
83. **Wang, L., S. Huskic, A. Cisterne, D. Rothemund, and P. R. Reeves.** 2002. The O-antigen gene cluster of *Escherichia coli* O55:H7 and identification of a new UDP-GlcNAc C4 epimerase gene. J. Bacteriol. **184:**2620–2625.
84. **Watson, D. A., D. M. Musher, and J. Verhoef.** 1995. Pneumococcal virulence factors and host immune responses to them. Eur. J. Clin. Microbiol. Infect. Dis. **14:**479–490.
85. **Weiser, J. N., A. A. Lindberg, E. J. Manning, E. J. Hansen, and E. R. Moxon.** 1989. Identification of a chromosomal locus for expression of lipopolysaccharide epitopes in *Haemophilus influenzae*. Infect. Immun. **57:**3045–3052.
86. **Whitfield, C.** 2006. Biosynthesis and assembly of capsular polysaccharides in *Escherichia coli*. Annu. Rev. Biochem. **75:**39–68.
87. **Xu, D. Q., J. O. Cisar, N. Ambulos, Jr., D. H. Burr, and D. J. Kopecko.** 2002. Molecular cloning and characterization of genes for *Shigella sonnei* form I O polysaccharide: proposed biosynthetic pathway and stable expression in a live *Salmonella* vaccine vector. Infect. Immun. **70:**4414–4423.
88. **Xu, D. Q., J. Thompson, and J. O. Cisar.** 2003. Genetic loci for coaggregation receptor polysaccharide biosynthesis in *Streptococcus gordonii* 38. J. Bacteriol. **185:**5419–5430.
89. **Yoshida, Y., S. Ganguly, C. A. Bush, and J. O. Cisar.** 2005. Carbohydrate engineering of the recognition motifs in streptococcal co-aggregation receptor polysaccharides. Mol. Microbiol. **58:**244–256.
90. **Yother, J.** 2004. Capsules, p. 30–48. *In* E. I. Tuomanen (ed.), The pneumococcus. ASM Press, Washington, D.C.
91. **Zhang, J. R., I. Idanpaan-Heikkila, W. Fischer, and E. I. Tuomanen.** 1999. Pneumococcal *licD2* gene is involved in phosphorylcholine metabolism. Mol. Microbiol. **31:**1477–1488.
92. **Zhang, L., J. Radziejewska-Lebrecht, D. Krajewska-Pietrasik, P. Toivanen, and M. Skurnik.** 1997. Molecular and chemical characterization of the lipopolysaccharide O-antigen and its role in the virulence of *Yersinia enterocolitica* serotype O:8. Mol. Microbiol. **23:**63–76.
93. **Zolli, M., D. J. Kobric, and E. D. Brown.** 2001. Reduction precedes cytidylyl transfer without substrate channeling in distinct active sites of the bifunctional CDP-ribitol synthase from *Haemophilus influenzae*. Biochemistry **40:** 5041–5048.

Downloaded from jb.asm.org at Imperial College London on July 10, 2009

JOURNAL OF BACTERIOLOGY, Nov. 2007, p. 7841–7855 Vol. 189, No. 21
0021-9193/07/$08.00+0 doi:10.1128/JB.00836-07
Copyright © 2007, American Society for Microbiology. All Rights Reserved.

Genetic Relatedness of the *Streptococcus pneumoniae* Capsular Biosynthetic Loci▽†

Angeliki Mavroidi,[1]‡ David M. Aanensen,[1]‡ Daniel Godoy,[1] Ian C. Skovsted,[2] Margit S. Kaltoft,[2] Peter R. Reeves,[3] Stephen D. Bentley,[4] and Brian G. Spratt[1]*

Department of Infectious Disease Epidemiology, Imperial College London, London, United Kingdom[1]; Statens Serum Institut, Copenhagen, Denmark[2]; School of Molecular and Microbial Biosciences, University of Sydney, Sydney, Australia[3]; and Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, United Kingdom[4]

Received 30 May 2007/Accepted 21 August 2007

Streptococcus pneumoniae **(the pneumococcus) produces 1 of 91 capsular polysaccharides (CPS) that define the serotype. The *cps* loci of 88 pneumococcal serotypes whose CPS is synthesized by the Wzy-dependent pathway were compared with each other and with additional streptococcal polysaccharide biosynthetic loci and were clustered according to the proportion of shared homology groups (HGs), weighted for the sequence similarities between the genes encoding the shared HGs. The *cps* loci of the 88 pneumococcal serotypes were distributed into eight major clusters and 21 subclusters. All serotypes within the same serogroup fell into the same major cluster, but in six cases, serotypes within the same serogroup were in different subclusters and, conversely, nine subclusters included completely different serotypes. The closely related *cps* loci within a subcluster were compared to the known CPS structures to relate gene content to structure. The *Streptococcus oralis* and *Streptococcus mitis* polysaccharide biosynthetic loci clustered within the pneumococcal *cps* loci and were in a subcluster that also included the *cps* locus of pneumococcal serotype 21, whereas the *Streptococcus agalactiae cps* loci formed a single cluster that was not closely related to any of the pneumococcal *cps* clusters.**

Many important bacterial pathogens produce a polysaccharide capsule that is believed to contribute to virulence by aiding bacterial survival in blood (47). Capsular polysaccharides (CPS) are surface exposed and immunogenic and provide a target for the host immune response. Consequently, there has been selection for mechanisms by which encapsulated bacteria evade the host immune system, and in most encapsulated species, this has been achieved by the generation of antigenic diversity, such that strains of the pathogen may express one of a number of different CPS (40).

In *Streptococcus pneumoniae*, 90 immunochemically distinct CPS types have been identified and sera that recognize the differences between these capsules provide a serological typing scheme, resolving pneumococci into individual serotypes or into serogroups, which include multiple immunologically related serotypes (16, 17). The serotyping scheme for pneumococci has been updated many times, and the 90 currently recognized serotypes probably represent a high proportion of the total capsule diversity in the species, although a new variant within serogroup 6 (serotype 6C) has recently been reported (33).

The biochemical structures of 54 capsular types are known (22), and the sequences of the CPS biosynthetic (*cps*) loci of several serotypes have been reported over the last decade (15, 18, 26, 51), but the sequences of the *cps* loci of all pneumococcal serotypes have only recently become available, allowing relationships among serology, genetics, and structure to be explored (5).

The CPS of 88 serotypes are known to be synthesized by the Wzy-dependent pathway, and the *cps* loci encoding this pathway in these serotypes are located at the same chromosomal location between *dexB* and *aliA* (5, 18, 36, 51); the capsules of the other two serotypes (3 and 37) are synthesized by the synthase pathway and are not considered further in this paper. The Wzy-dependent pathway is also found in several other streptococcal species (25). For example, it is found in CPS biosynthesis of *Streptococcus agalactiae* (10) and *Streptococcus suis* (39), in receptor polysaccharide synthesis (RPS) in the viridans group streptococci, *Streptococcus gordonii*, *Streptococcus mitis*, and *Streptococcus oralis* (12, 48–50) and exopolysaccharide synthesis in *Streptococcus thermophilus* (7, 8).

The evolution of these streptococcal loci is almost certainly very complex, with a long history of gene capture and loss and genetic rearrangements, and it is probably unrealistic to expect to be able to untangle their evolutionary history. Rather than take a phylogenetic approach, which seems inappropriate, we have explored the relatedness of the pneumococcal *cps* loci (and other streptococcal polysaccharide biosynthetic loci) by cluster analysis, incorporating similarities in both gene content and nucleotide sequence, and used this as a framework to relate differences in the *cps* loci of closely related serotypes to differences in their CPS structures.

MATERIALS AND METHODS

Polysaccharide biosynthetic sequences. A total of 116 streptococcal polysaccharide biosynthetic sequences were used in this study; those from strains of each of the 88 pneumococcal serotypes that are known to use the Wzy-dependent pathway (accession numbers CR931632, CR931633, CR931635, CR931637 to CR931708, and CR931710 to CR931722; 5), 17 pneumococcal *cps* loci sequenced

* Corresponding author. Mailing address: Department of Infectious Disease Epidemiology, Imperial College London, Room G22, Old Medical School Building, St. Mary's Hospital, Norfolk Place, London W2 1PG, United Kingdom. Phone: 44 (0)20 7594 3398. Fax: 44 (0)20 7402 3927. E-mail: b.spratt@imperial.ac.uk.

† Supplemental material for this article may be found at http://jb.asm.org/.

‡ A.M. and D.M.A. contributed equally to this work.

▽ Published ahead of print on 31 August 2007.

Downloaded from jb.asm.org at Imperial College London on July 10, 2009

by other groups (see Table S1 in the supplemental material), including the serotype 4 locus extracted from the TIGR4 genome (42), and 11 *S. agalactiae* or viridans group streptococcal sequences. The latter included sequences of the *cps* loci of *S. agalactiae* serotypes Ia and II to VIII (accession numbers AB028896, AY375362, AF163833, AF355776, AF349539, AF337958, AY376403, and AY375363, respectively; 10, 11) and the RPS loci of *S. oralis* strains 34 (AB181234; 50) and J22 (AB181235; 49) and of *S. mitis* NCTC12261.

The RPS locus of *S. mitis* strain NCTC12261 (nucleotides 215361 to 240666) was extracted from the genome sequence available at the J. Craig Venter Institute website (www.tigr.org). This sequence is not completely finished and includes some unresolved nucleotides; therefore, only the region between the *wzg* (*cpsA*) and *rmlD* (*rfbD*) genes was included. It should be noted that GenBank accession no. AB181235 was initially submitted as the sequence of the RPS biosynthetic locus of strain *S. mitis* J22, but this strain has subsequently been recharacterized as *S. oralis* (49). *S. pneumoniae* serotype 6C (33) could not be included in the cluster analysis, as its *cps* sequence was unavailable.

Bioinformatic methods and cluster analysis. TribeMCL (14) was run on the data set of 116 polysaccharide biosynthetic sequences with a TBLASTX cutoff of $1e^{-50}$ to assign the gene products to homology groups (HGs). The *cps* gene products were classified into Pfam families based on hidden Markov model profiles by using the Pfam database (4; http://www.sanger.ac.uk/Software/Pfam/). The pairwise nucleotide sequence similarities among *cps* genes whose products were assigned to the same HG were calculated by the Needleman-Wunsch algorithm by using the EMBOSS program NEEDLE (32).

Visual representation of the alignments using nucleotide similarities (BLASTN) and similarities of the gene products (TBLASTX) of the polysaccharide biosynthetic loci were performed by ACT version 6 and WebACT, an online version of the Artemis comparison tool (2, 9). The nucleotide differences between similar *cps* loci were calculated by the EMBOSS program Diffseq (MRC Rosalind Franklin Centre for Genomics Research, Hinxton, United Kingdom) and are presented elsewhere (see Table S1 in the supplemental material).

For clustering of the *cps* sequences, a modification of the method described by Tekaia et al. (41) was applied. A total of 99 polysaccharide biosynthetic loci that encode capsules synthesized by the Wzy-dependent pathway were included in the cluster analysis, i.e., 88 *S. pneumoniae cps* loci (5) and 11 other streptococcal sequences. A profile of each *cps* locus (gene content) was produced according to the presence or absence of each of the 254 HGs. The *dexB*, *aliA*, *aliB*, transposase, and group II intron genes were not included in the profiles. The products of the four *cps* genes *wzg*, *wzh*, *wzd*, and *wze* of all pneumococcal *cps* loci and those of the *S. mitis*, *S. oralis*, and *S. agalactiae* loci each fall into a single HG, and these genes also were not included in the *cps* profiles used for clustering.

Each polysaccharide biosynthetic locus was compared with all other loci, calculating the similarity scores in sequence-weighted gene content, and a 99-by-99 data table (T) resulting from the pairwise comparisons of the 99 polysaccharide biosynthetic loci was produced, where the similarity score (Tij) for each pair of biosynthetic loci is the sum of the percentage of nucleotide sequence similarities of the HGs shared between loci i and j divided by the total number of HGs present in locus j. It should be noted that Tij is normalized because it is divided by the total number of HGs in locus j (and is 100% for comparisons of each *cps* locus with itself), and the data table is not symmetrical since Tij is different from Tji if the loci i and j differ in the number of HGs. The data table was used as the input for cluster analysis by the XL-STAT software version 7.5.3 (Addinsoft, New York, NY). A distance (or dissimilarity) matrix was produced by the Pearson distance by transformation of Pearson's coefficient into dissimilarity values in a range of 0 to 1 (rows were clustered according to the columns). The distance matrix was used for cluster analysis by the unweighted-pair group method using arithmetic averages (19).

Biochemical structures and antigenic formulas. The CPS structures for 52 pneumococcal serotypes that use the Wzy-dependent pathway have been adopted from a report by Kamerling (22), incorporating the revised CPS structures of serotypes 15B, 17F, and 33F (20, 21, 24), and their symbolic representations were previously presented by Bentley et al. (5). The biochemical structures of the *S. agalactiae* CPS and the cell wall polysaccharides of *S. oralis* 34 (type 1Gn RPS) and *S. oralis* J22 (type 2G RPS) were from reports by Cieslewicz et al. (10) and Cisar et al. (12), respectively. The serotypes that react with each of the pneumococcal factor (typing) sera were obtained from data reported by Henrichsen (17).

RESULTS

Assignment of HGs. The products of the four *cps* genes (*wzg*, *wzh*, *wzd*, and *wze*) are relatively conserved in sequence in all 88 pneumococcal serotypes, but there are multiple highly divergent or nonhomologous groups of flippases, polysaccharide polymerases, initial transferases (ITs), and other *cps* gene products, including a large number of different groups of glycosyltransferases (GTs). Previously, we have classified the gene products of the pneumococcal *cps* loci into those that share significant homology (HGs) by using the program TribeMCL (5). Genes that encode proteins within the same HG have been given the same name, except for those that encode the polymerases and flippases, where the widely used generic gene names (*wzy* and *wzx*, respectively) have been retained (5, 38).

In this study, the gene products encoded within 28 additional streptococcal biosynthetic loci were added to those of the 88 pneumococcal *cps* loci encoding capsules synthesized by the Wzy-dependent pathway (see Materials and Methods) and TribeMCL was rerun on this combined set to allow these additional proteins to be assigned to pneumococcal HGs or to novel HGs. Table 1 shows the HGs that were common to both pneumococcal and other streptococcal polysaccharide biosynthetic loci. In all cases, the *cps* loci from different *S. pneumoniae* strains of the same serotype had identical HG profiles (data not shown) and almost identical sequences (see Table S1 in the supplemental material); therefore, only the pneumococcal sequences of Bentley et al. (5) were used in all further analyses.

Using the TBLASTX cutoff of $1e^{-50}$, the average amino acid sequence similarities of proteins assigned by TribeMCL to the same HG ranged from 42 to 100%, and 83% of the HGs included proteins with average sequence similarities of >70%. Members of the same HG should therefore correspond to proteins (and thus genes) with broadly similar functions.

Predicted function of *cps* gene products. A detailed analysis of the predicted functions and specificities of the products of the genes in the pneumococcal *cps* loci is presented elsewhere (1), and we have used these functional assignments (and gene names) when discussing *cps* loci that were shown to be similar by cluster analysis. The assignment of the initial sugars of the repeat units and the specificity of the GTs and other transferases are also discussed in detail elsewhere, together with the basis for these assignments (1). In many cases, the comparison of closely related *cps* loci and their CPS structures provides strength to these assignments of GT specificity (discussed below).

Cluster analysis. Figure 1 shows the clustering of the *cps* loci, based on a measurement of the extent of sharing of HGs and the degree of sequence similarity of the genes encoding the shared HGs as described in Materials and Methods. This approach was used as shared HGs could be encoded by genes that are almost identical, or that are divergent, in nucleotide sequence, and adjusting for sequence similarity results in tighter clustering when the genes encoding the shared HGs have very similar sequences than when they have less similar sequences. Clustering was used to identify those *cps* loci that possess a high level of similarity in sequence-weighted gene content, which forms a useful framework for discussing the relationships among the *cps* loci of the different pneumococcal serotypes.

The *cps* loci of the 88 pneumococcal serotypes fell into eight clusters at a distance of 0.3 (Fig. 1). The *cps* loci of pneumococcal serotypes within the same serogroup were invariably in

Downloaded from jb.asm.org at Imperial College London on July 10, 2009

TABLE 1. Products of other streptococcal polysaccharide biosynthetic loci that are within pneumococcal HGs

Name in *S. pneumoniae*[a]	Organism	Name in other species
WchA	*S. agalactiae* Ia	CpsIaE
	S. agalactiae II	CDS
	S. agalactiae III	CpsE
	S. agalactiae IV	CpsE
	S. agalactiae V	CpsE
	S. agalactiae VI	CpsE
	S. agalactiae VII	CDS
	S. agalactiae VIII	CDS
	S. mitis NCTC12261	CpsE
	S. oralis 34	WchA
	S. oralis J22	WchA
Glf	*S. mitis* NCTC12261	Glf
	S. oralis 34	Glf
	S. *oralis* J22	Glf
RmlB	*S. mitis* NCTC12261	RfbB
	S. oralis 34	RmlB
	S. oralis J22	RmlB
RmlD	*S. mitis* NCTC12261	RfbD
	S. oralis 34	RmlD
	S. oralis J22	RmlD
RmlA	*S. mitis* NCTC12261	RfbA
	S. oralis 34	RmlA
	S. oralis J22	RmlA
RmlC	*S. mitis* NCTC12261	SMT0231
	S. oralis 34	RmlC
	S. oralis J22	RmlC
Wzx-1	*S. mitis* NCTC12261	SMT0226
	S. oralis 34	Wzx
	S. oralis J22	Wzx
WchF	*S. agalactiae* VIII	CDS
	S. mitis NCTC12261	SMT0220
	S. oralis 34	WchF
	S. oralis J22	WchF
WchJ	*S. agalactiae* Ia	CpsIaF
	S. agalactiae II	CDS
	S. agalactiae III	CpsF
	S. agalactiae IV	CpsF
	S. agalactiae V	CpsF
	S. agalactiae VI	CpsF
	S. agalactiae VII	CDS
WciF	*S. oralis* 34	WefD
	S. oralis J22	WefG
WcrH	*S. mitis* NCTC12261	Eps6N
	S. oralis 34	WefE
	S. oralis J22	WefE
WcwA	*S. mitis* NCTC12261	SMT0221
	S. oralis 34	WefA
	S. oralis J22	WefA
WcwC	*S. mitis* NCTC12261	Eps9H
WchV	*S. agalactiae* VIII	CDS
Wzy-19	*S. agalactiae* VIII	CDS
WcwK	*S. mitis* NCTC12261	CpsY
	S. oralis 34	WefH
	S. oralis J22	WefF
Wzy-32	*S. oralis* 34	Wzy
	S. oralis J22	Wzy
WcwF	*S. mitis* NCTC12261	SMT0223

[a] Gene products that fall into the same HG as pneumococcal Wzg, Wzh, Wzd, and Wze were present in all of the *S. agalactiae* and viridans group streptococci listed.

the same cluster. In most cases, each cluster included *cps* loci of several different serogroups but the *cps* locus of serotype 8 was sufficiently different from those of the other *cps* loci to cluster alone, as were the *cps* loci of the serotypes within both serogroups 9 and 11. A ninth cluster included all of the *cps* loci of *S. agalactiae* but no pneumococcal *cps* loci. In contrast, the *S. mitis* and two *S. oralis* RPS loci clustered among pneumococcal *cps* loci within cluster 2 (Fig. 1).

Subclusters were defined at a distance of 0.05 to identify very closely related *cps* loci (Fig. 1). All *S. pneumoniae* serotypes within the same serogroup fell into the same subcluster, except for the serotypes within serogroups 7, 16, 17, 33, 35, and 47, which were split into two or three different subclusters. Cluster analysis identified nine pneumococcal subclusters that included the *cps* loci of completely different serotypes (Fig. 1), and in three cases these loci were fully syntenic (synteny is defined here as the same genes being present in the same order). Thus, the *cps* loci of serotypes 44 and 46 were syntenic with those of serogroup 12 (Fig. 2), as shown previously (5), and those of serotypes 35A, 35C, and 42 (see Fig. 4) were syntenic, as were those of serotypes 7B, 7C, and 40 (see Fig. 3). For seven of these nine subclusters, there were factor sera that cross-reacted with the different serotypes within the subclusters (Table 2).

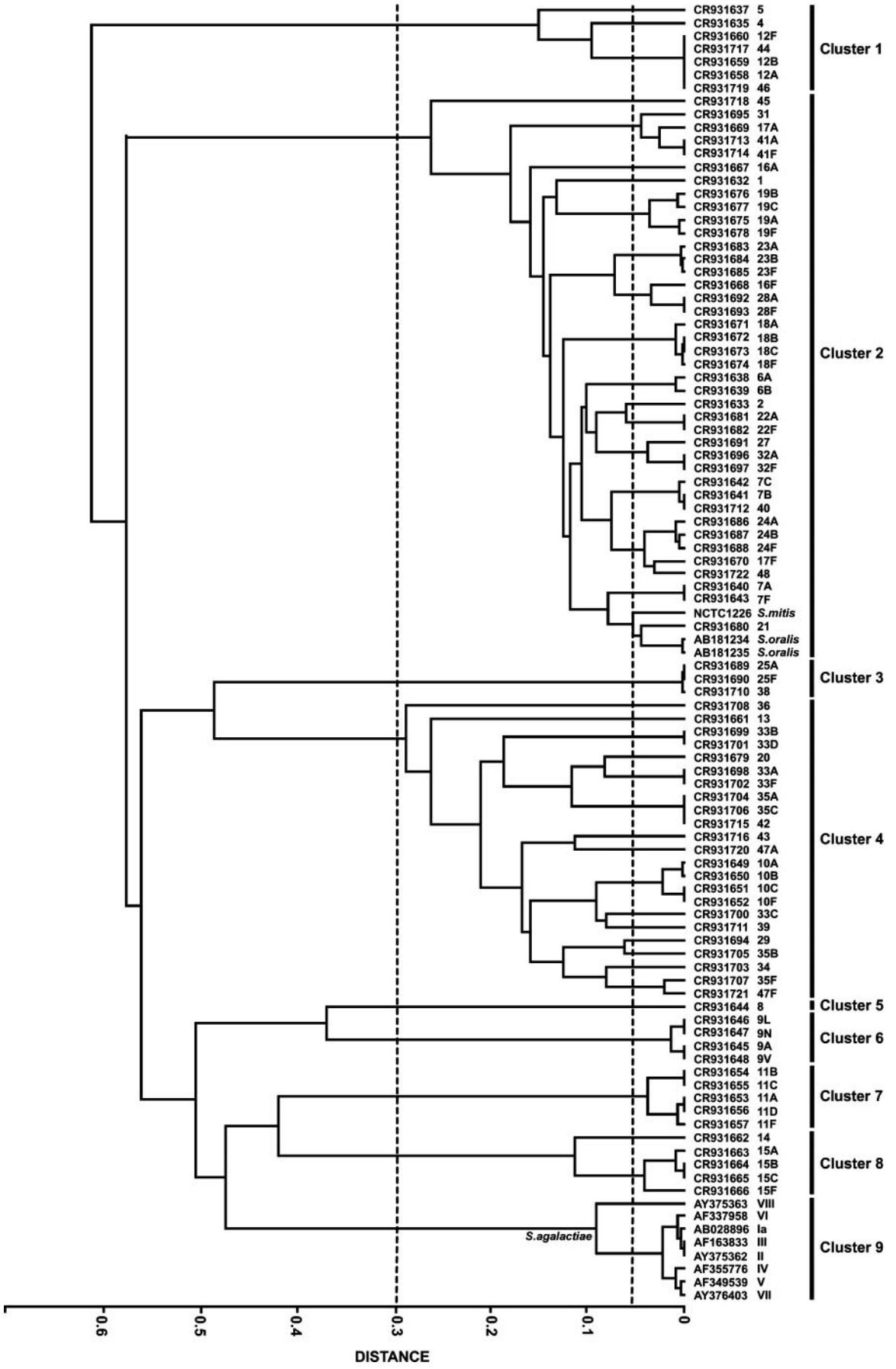
The similarities between the *cps* loci of different serotypes were sometimes greater than that between serotypes within the same serogroup (Fig. 1). For example, serotype 40 *cps* was more similar in sequence-weighted gene content to serotype 7B than 7B was to 7C (similarity scores of 97.5% and 90.3%, respectively) and was much more similar to the latter two serotypes, falling into the same subcluster, than these were to serotypes 7F and 7A (Fig. 1).

The serotypes within each major cluster are considered in the next sections, focusing on those that are sufficiently similar to be placed within the same subcluster (Fig. 1). The *cps* loci of those serotypes that are not within a subcluster are shown elsewhere (see Fig. S1 in the supplemental material).

Cluster 1. The *cps* loci of serogroup 12 and serotypes 44 and 46 form a subcluster and, ignoring transposon-related genes, are syntenic, presumably having diverged recently from a common ancestral *cps* locus (Fig. 2). The structures of serotypes 12F and 12A are known and differ only in the initial sugar and a side branch. As there is synteny, these differences in repeat unit structure are presumably due to the amino acid sequence variation in the IT WciI and the GT (WcxB) that is predicted to make this side branch (1, 5; see below). The structures of types 12B, 44, and 46 are unknown, but the synteny and immunological cross-reactivity between the serotypes in this subcluster (Table 2) suggest structural similarities with the CPS of serotypes 12F and 12A.

Serotype 4 and 5 *cps* loci are also members of cluster 1, but they are less closely related to the other members of the cluster (Fig. 1). Compared with serogroup 12 and types 44 and 46, the *cps* locus of type 4 shares the *wciI*, *wciJ*, *mnaA*, and *fnlA-C* genes and type 5 shares the *wciI*, *wciJ*, and *fnlA-C* genes, which explains why these serotypes cluster together (Fig. 2; see Fig. S1 in the supplemental material). The type 45 *cps* locus also shares the *wciI*, *wciJ*, *wcxB*, and *fnlA-C* genes but their sequences are distantly related to those of cluster 1 (data not

Downloaded from jb.asm.org at Imperial College London on July 10, 2009



Downloaded from jb.asm.org at Imperial College London on July 10, 2009

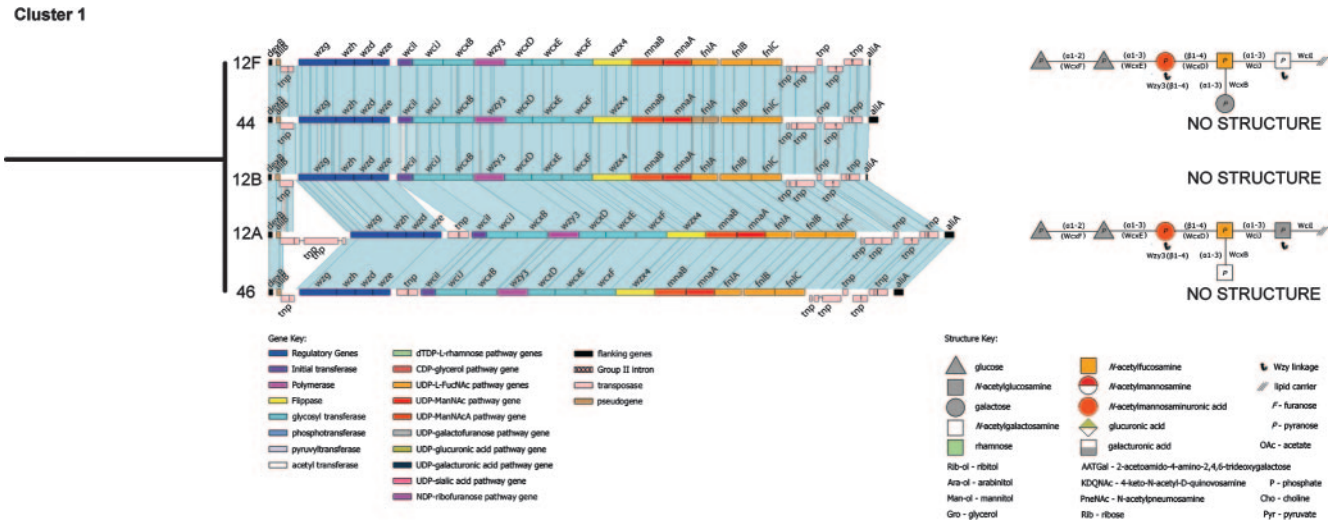


FIG. 2. Comparisons of those *cps* loci within cluster 1 assigned to the same subcluster. The relatedness of the *cps* loci in the subcluster is from Fig. 1. The *cps* loci and ACT comparisons based on the amino acid sequence similarities of the products of the *cps* loci are shown, along with the CPS repeat unit structures, where known. CPS structures are represented as in the report by Bentley et al. (5), with the initial sugar that is attached to the lipid carrier (///) on the right. The linkage between repeat units catalyzed by the Wzy polymerase is between the two sugars marked by arrows. The genes that are predicted to catalyze the linkages in the repeat units are shown on the structure. The basis of these gene assignments is discussed in detail by Aanensen et al. (1), and uncertain GT assignments are presented in parentheses. The color keys for the functional classes of genes in the *cps* loci and for the sugars (and other constituents) in the CPS repeat units are shown at the bottom.

shown) and it is placed in cluster 2 (Fig. 1; see Fig. S1 in the supplemental material).

The L-Fuc*p*NAc biosynthetic pathway genes (*fnlA-C*) are present in the *cps* loci of cluster 1, and L-Fuc*p*NAc is known to be present in the CPS of types 4, 5, 12A, and 12F, as well as the constituents of type 46. The type 44 *cps* locus has *fnlC* frameshifted; therefore, L-Fuc*p*NAc is not expected in its CPS structure. The type 5 CPS contains, in addition to L-Fuc*p*NAc, two sugars (Pne*p*NAc and KQDNAc) which are intermediates in the L-Fuc*p*NAc biosynthetic pathway (30) and within pneumococci are uniquely present in the type 5 CPS. The products of the *fnlA*, *fnlB*, and *fnlC* genes in type 5 are 76%, 98%, and 99% similar to the other members of the cluster. Therefore, the sequence divergence of FnlA may be implicated in the above difference in the L-Fuc*p*NAc pathway which results in these unusual sugars in the CPS of serotype 5 (1).

The Man*p*NAcA biosynthetic pathway genes *mnaA* and *mnaB* are both present in the *cps* loci of serogroup 12 and serotypes 44 and 46; therefore, Man*p*NAcA is present in the CPS structures of types 12A and 12F (and is predicted also to be present in types 12B, 44, and 46). In serotype 4, only *mnaA* is present and accordingly the type 4 CPS contains Man*p*NAc β-1,4 linked to L-Fuc*p*NAc. In the serotype 5 *cps* locus, neither of these genes is present and consequently the above sugars are absent from the type 5 CPS.

The IT gene *wciI* is present in the *cps* loci of cluster 1 and also in the *cps* loci of types 25A, 25F, 38, and 45, although the sequences of the latter are very divergent from those in cluster 1 (data not shown). WciI appears to have different specificities, as the initial sugar in cluster 1 varies (D-Glc*p*NAc in serotype 12A, D-Gal*p*NAc in serotypes 4 and 12F, and KQDNAc in serotype 5), although the sequences of WciI within this cluster are very similar (70 to 100% pairwise identity; average, 81.4%). This highlights the difficulties in predicting the enzymatic specificities of transferases from their amino acid sequence similarity.

The GT genes *wcxD*, *wcxE*, and *wcxF* are present only in serogroup 12 and types 44 and 46, and their products are predicted to form the linkages common to these serotypes. Inspection of the two known structures, types 12A and 12F, suggests that these GTs catalyze the formation of the common tetrasaccharide element D-Glc*p*-(α1-2)-D-Glc*p*-(α1-3)-D-Man-*p*NAcA-(β1-4)-L-Fuc*p*NAc. The GT gene *wciJ* is present in the *cps* loci of all serotypes of cluster 1 (and in type 45), and L-Fuc*p*NAc is α-1,3 linked to the repeat units but to different initial sugars (see above). Therefore, we have tentatively assigned *wciJ* as the putative α-1,3-L-Fuc*p*NAc transferase gene. As in the case of the IT of cluster 1, the specificities of the linkage made by similar members (average pairwise identity, 83.9%) of the same HG (WciJ) appear to vary. The GT gene *wcxB* is present in serogroup 12 and types 44 and 46 (and in type 45) and is suggested to encode the GT catalyzing the α-1,3 linkage of D-Gal*p*NAc to α-L-Fuc*p*NAc in the CPS of type 12A and of D-Gal*p* to α-L-Fuc*p*NAc in type 12F (and type 45).

FIG. 1. Cluster analysis of streptococcal polysaccharide biosynthetic loci. The relatedness of the loci was determined from the extent of sharing of HGs, adjusted for the percentage DNA sequence similarity between the genes encoding the shared HGs, as described in Materials and Methods. Clusters are defined by the dotted line at a distance of 0.3, and subclusters are defined by the dotted line at 0.05. Serotypes of *S. pneumoniae* (and *S. agalactiae*) or species of viridans group streptococci are shown at the end of each branch of the tree.

Downloaded from jb.asm.org at Imperial College London on July 10, 2009

TABLE 2. Serotypes that react with each pneumococcal typing serum

Serum[a]	Serotype(s)
1	1
2	2
3	3
4	4
5	5
6a	6A, 6B, 33D
6b	6A
6c	6B
7a	7F, 7A, 7B, 7C
7b	7F, 7A
7c	7A
7d	7B, 7C
7e	7B
7f	7C
7g	7C, 20, 40
7h	7B, 7C, 19B, 19C, 24F, 24B, 40
8	8
9a	9A, 9L, 9N, 9V
9b	9L, 9N
9c	9A, 9L, 9V
9d	9A, 9V
9e	9N, 36
9f	9L
9g	9V
10a	10F, 10A, 10B, 10C
10b	10F, 10B, 10C
10c	10A, 10B, 10C
10d	10A, 10B, 39
10e	10B
10f	10C
11a	11F, 11A, 11B, 11C, 11D
11b	11F, 11B, 11C, 11D
11c	11A, 11C, 11D
11d	11A, 11C, 16F
11e	11F, 11A, 11D
11f	11B, 11C
11g	11F, 11B
12a	12F, 12A, 12B
12b	12F, 12B, 44
12c	12A, 12B, 46
12d	12F, 12A, 44
12e	12B
13	13
13b	13, 29
14	14
15a	15F, 15A, 15B, 15C, 23A
15b	15F, 15B
15c	15F, 15A
15d	15A, 15B, 15C
15e	15B, 15C
15f	15F
15g	15A
15h	15B
16a	16F, 16A
16b	16F, 28F
16c	16A
17a	17F, 17A
17b	17F
17c	17A
18a	18F, 18A, 18B, 18C
18b	18F, 18A, 18B, 18C, 23F
18c	18F, 18C
18d	18A
18e	18B, 18C
18f	18F
18g	18B
19a	19F, 19A, 19B, 19C
19b	19F
19c	19A, 19B, 19C
19d	19F, 19A
19e	19B
19f	19C
20	20
20b	20, 31, 33A, 35A, 35C, 42
21	21
22a	22F, 22A
22b	22F
22c	22A
23a	23F, 23A, 23B
23b	23F, 23B
23c	23A
23d	23B, 28F, 28A
24a	24F, 24A, 24B
24b	24F, 24B
24c	24A
24c	24F, 24A
24e	24B
25a	25F, 25A
25b	25F, 38
25c	25A
27	27
27b	27, 32F, 32A
28a	28F, 28A
28b	28F
28c	28
29	29
29b	29, 35B
31	31
32a	32F, 32A
32b	32A
33a	33F, 33A, 33B, 33C, 33D
33b	33F, 33A
33c	33B, 33C, 33D
33d	33F, 33A, 33B, 33D
33e	33C
33f	33B, 33D
34	34
34b	34, 35F
35a	35F, 35A, 35B, 35C, 47F
35b	35F, 47F
35c	35A, 35B, 35C, 42
36	36
37	37
38	38
38a	25A, 38
39	39
40	40
41a	41F, 41A
41b	41F
42	42
42a	42, 35C
43	43
43b	43, 47A
44	44
44b	44, 46
45	45
46	46
47a	47F, 47A
48	48

[a] Type or factor serum.

Downloaded from jb.asm.org at Imperial College London on July 10, 2009

Cluster 2. The *cps* loci of 40 pneumococcal serotypes and the *S. oralis* and *S. mitis* polysaccharide biosynthetic loci (the latter two are discussed in a separate section below) are within this large cluster, and all *cps* loci within the cluster contain the rhamnose biosynthesis genes (*rml*). All of the available CPS structures of these serotypes contain rhamnose (except for types 1 and 24B, due to the presence of frameshift mutations within the *rmlC* and *rmlD* genes, respectively), and there is a perfect correlation between the presence of the linkage L-Rha*p*-(β1-4)-β-D-Glc*p* (49) in the repeat units where the *cps* locus possesses the putative rhamnosyltransferase gene *wchF*. It is therefore predicted that WchF is the GT that catalyzes the latter linkage. The other *cps* loci in this cluster have the *rml* genes present but in association with rhamnosyltransferase genes other than *wchF*, and where possible we have assigned these transferases (1).

Serotype 1 is the only case where there is no candidate within the *cps* locus for the IT gene. The serotype 1 repeat unit contains AAT-Gal*p*, which is a component of pneumococcal teichoic acid, and this unusual sugar is also present in *Bacteroides fragilis* CPS A, the *Shigella sonnei* form I antigen, and the *Plesiomonas shigelloides* serotype O17 antigen (1). In these latter species, an IT for AAT-Gal*p* has been identified and there is a homolog in the pneumococcal chromosome (13). It seems likely that this chromosomal IT synthesizes lipid-linked AAT-Gal*p*, which can be used for repeat unit synthesis in serotype 1, as further discussed by Aanensen et al. (1).

The serotypes within serogroups 6, 18, 19, 22, 23, 24, 28, 32, and 41 each fall within the same subcluster, suggesting that they are each descended from a common ancestral *cps* locus, whereas the *cps* loci of serogroups 7, 16, and 17 are split into different subclusters (Fig. 1). The *cps* loci of serotypes within serogroups 6, 22, 28, 32, and 41 are in each case syntenic, and differences in structure should be due to amino acid variation in one or more of the *cps* gene products (Fig. 3).

Serogroup 6. CPS of serotypes 6A and 6B differ only in the rhamnose linkage to Rib-ol (α-1,3 in 6A or α-1,4 in 6B), which correlates with a single nucleotide polymorphism in the rhamnosyltransferase gene *wciP* (27). The structure of the repeat unit of serotype 6C has been reported (33) and is most similar to that of 6A, containing the Rha*p*(α1-3)D-Rib-ol linkage. Serotype 6C differs from type 6A only in the presence of a Glc*p* residue as the second sugar, as opposed to Gal*p*, presumably due to amino acid differences in the GT, WciN (1).

Serogroup 7. The *cps* loci form two syntenic pairs (7F-7A and 7B-7C) that cluster apart from each other (Fig. 3), and the CPS structures of 7F, 7A, and 7B are known (22). Type 7A lacks only the side branch D-Gal*p*-(β1-2)-α-D-Gal*p* compared with type 7F, and this difference can be attributed to a frameshift mutation in the GT gene *wcwD*. The CPS of serotype 7B has little in common with those of types 7F and 7A. The CPS structures of types 7A, 7B, and 7F possess the L-Rha*p*-(β1-4)-β-D-Glc*p* linkage attributed to the presence of WchF, but L-Rha*p* is acetylated in types 7A and 7F, presumably due to the presence of the *wcwC* acetyltransferase gene in their *cps* loci. Additionally, the D-Glc*p*NAc-(α1-2)-α-L-Rha*p* linkage is present as a side branch in types 7A and 7F and in the main chain in type 7B and has been suggested to be the epitope common to all serotypes of serogroup 7 recognized by factor serum 7a (22), but there is no common GT as a candidate for forming the latter linkage. As mentioned previously, types 7B, 7C, and 40 constitute a separate subcluster, reflected in the serology (Table 2), with factor serum 7h cross-reacting with 7B, 7C, and 40, but not 7A or 7F, and 7g reacting with 7C and 40 (the CPS structures of types 7C and 40 are not known).

Serogroup 17. Serotype 17F and 17A *cps* loci are in different subclusters (Fig. 1) but share the IT gene *wchA*, three genes in the central *cps* region (the putative GT genes *wcrT* and *wcrV* and the acetyltransferase gene *wcrU*), and the polymerase gene *wzy-16*, although the order and the nucleotide sequences of the latter four shared genes differ substantially. The products of these shared genes should account for the synthesis of the common structural elements D-Gal*p*-(α1-3)-L-Rha*p*2Ac-(β1-4)-α-L-Rha*p* in the main chain of the repeat unit, the side branch D-Gal*p*-(β1-4)-L-Rha*p*2Ac (21), and the polymerase linkage D-Glc*p*(β1-3)D-Gal*p*, which are common to both serotypes 17F and 17A. WcrV contains two functional GT domains and is suggested to catalyze two transferase reactions (1). Except for these commonalities, serogroup 17 *cps* loci differ considerably and consequently type 17A clusters with serogroup 41 and type 31, whereas type 17F clusters with serogroup 24 and type 48 (Fig. 3).

Serogroup 18. Serotype 18F, 18A, 18B, and 18C *cps* loci form a separate subcluster; types 18B and 18C are syntenic, whereas type 18F has an extra acetyltransferase gene (*wcxM*) and type 18A lacks the acetyltransferase gene *wciX* (Fig. 3). As in serogroup 15, the O-acetylation state correlates with the presence of an intact acetyltransferase gene (*wciX*) in the *cps* loci of types 18F and 18C and accordingly the last Glc*p* residue of the repeat unit is acetylated, whereas it has a frameshift mutation in type 18B and is absent in type 18A. Type 18F additionally has the rhamnose residue acetylated due to the presence of the extra acetyltransferase gene *wcxM*.

Apart from the differences in the acetylation patterns, the CPS of types 18B, 18C, and 18F are identical; type 18A differs only in the third residue of the repeat unit (α-D-Glc*p*NAc instead of α-D-Glc*p*). Besides the rhamnosyltransferase gene *wchF*, serogroup 18 *cps* loci possess three shared GT genes (*wciU*, *wciV*, and *wciW*). WciV has homology to β-1,4-D-Gal transferases and is likely to catalyze the linkage of β-1,4-D-Gal*p* to α-D-Glc*p*NAc in type 18A but to α-D-Glc*p* in types 18B, 18C, and 18F. WciW possesses the same Pfam domain (PF05704) as WchN, which has been suggested to form the D-Gal*p*-(α1-2)-β-D-Gal*p* linkage in serogroup 15 (5). Serogroup 18 CPS structures do not have the latter linkage but do have a D-Glc*p*-(α1-2)-β-D-Gal*p* linkage. Thus, we propose that WciW functions as the α-1,2 D-Glc*p* transferase in serogroup 18. WciU would therefore, by a process of elimination, function as the putative D-Glc*p*NAc-(α1-3)-β-L-Rha*p* GT in type 18A or the D-Glc*p*-(α1-3)-β-L-Rha*p* GT in types 18F, 18B, and 18C. Serogroup 18 CPS contains glycerol-1-phosphate (Gro-1P), and accordingly, the sugar phosphate transferase *wciY* and the Gro-1P biosynthesis (*gct*) genes are present in their *cps* loci (5, 18).

Serogroup 19. A comparative analysis of serotype 19F, 19A, 19B, and 19C *cps* loci and the putative biosynthetic pathways has been presented previously (29). The *cps* loci of all four serotypes are in the same subcluster (Fig. 3), and those of types 19F and 19A are syntenic and their CPS structures differ only in the polymerization linkage between D-Glc*p* and α-L-Rha*p*

Downloaded from jb.asm.org at Imperial College London on July 10, 2009

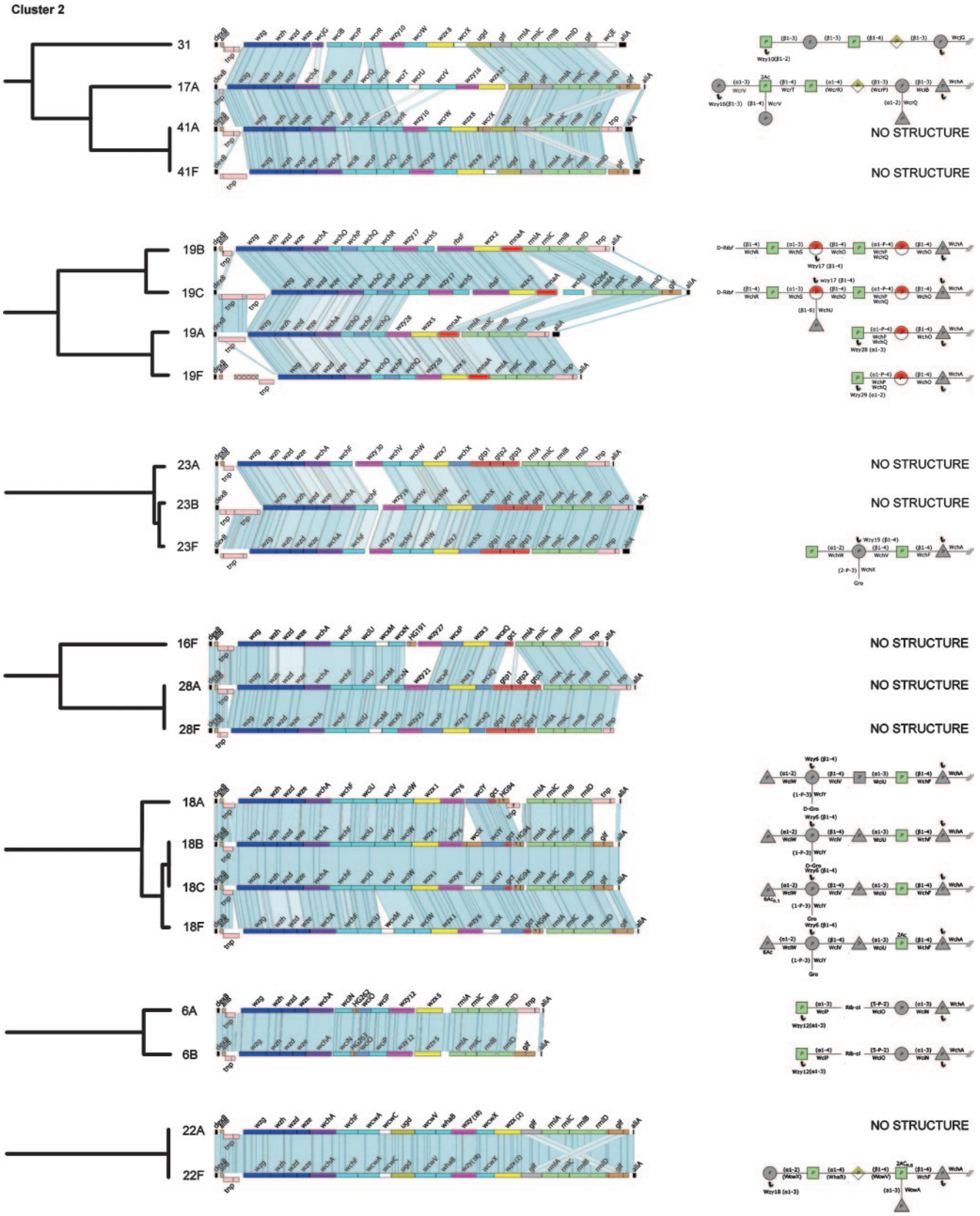


FIG. 3. Comparisons of those *cps* loci within cluster 2 assigned to the same subcluster. Details are as in Fig. 2.

Downloaded from jb.asm.org at Imperial College London on July 10, 2009

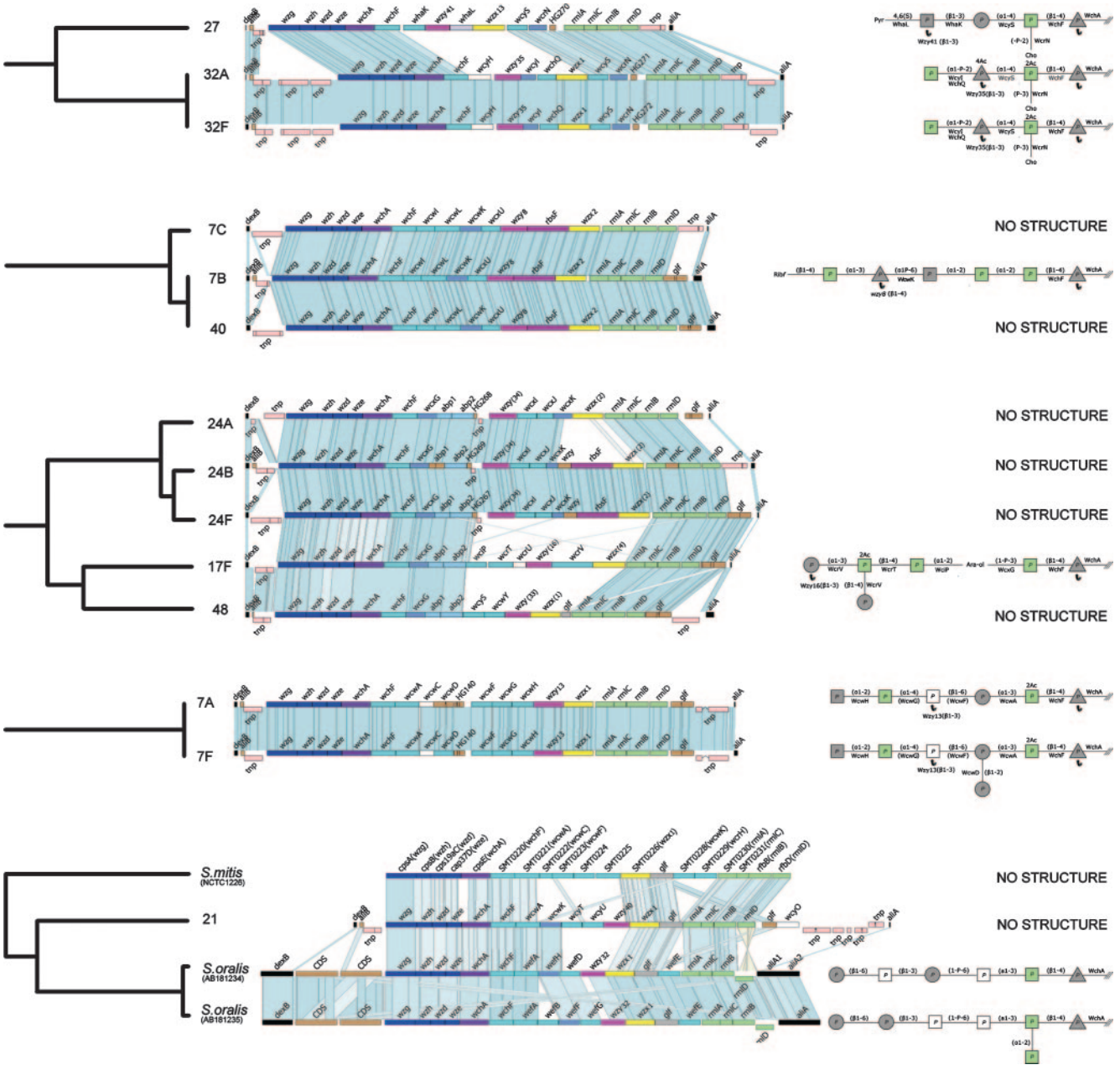


FIG. 3—*Continued*.

(α-1,2 in type 19F and α-1,3 in type 19A). Thus, it has been proposed that the differences between the products of the *wzy* genes should account for the different polymerization linkages (29). The products of the *wzy-28* genes of types 19F and 19A fall into the same HG, but they display 22% sequence divergence, which supports this view. The products of the *wzy* and *wzx* genes of types 19A and 19F fall into different HGs than those of types 19B and 19C.

Serogroup 22. The *cps* loci of serotypes 22F and 22A are syntenic (Fig. 3). The chemical differences between the CPS of these serotypes are not known since only the former structure is available.

Serogroup 23. The *cps* loci of serotypes 23F, 23A, and 23B are in the same subcluster (Fig. 3) and differ only in their polymerase genes; *wzy-19* in types 23F and 23B is replaced by *wzy-30* in type 23A, suggesting a different polymerase linkage (only the type 23F CPS structure is known). Type 23F CPS contains only one D-Glc*p* residue linked with a β-1,4 bond to β-D-Gal*p*. The IT gene *wchA* is present in the type 23F *cps* locus, and as this transfers Glc*p* to the lipid carrier, we are confident that the polymerization linkage is D-Glc*p*-(β1-4)-β-D-Gal*p*. Type 23F CPS contains glycerol-2-phosphate (Gro-2P), and as the *wchX* and *gtp1* and *-3* genes are also present in the *cps* loci of types 23A and 23B, it is assumed that they also contain Gro-2P (28).

Downloaded from jb.asm.org at Imperial College London on July 10, 2009

Serogroup 24. Serotype 24F, 24A, and 24B *cps* loci form a subcluster with those of types 17F and 48, the latter two serotypes being less related as they differ in gene content (Fig. 3). The *cps* loci of types 24B and 24F are syntenic and, compared with the type 24A *cps* locus, have an extra fragment of a polymerase gene and the putative ribofuranose biosynthetic gene (*rbsF*) that also are present in types 7B, 7C, and 40. In the type 24B *cps* locus, the *abp1* and *rmlC* genes are frameshifted; therefore, no arabinitol or rhamnose is expected in the CPS. The type 24A *cps* locus lacks the *rbsF* gene (and the *wzy* fragment), and therefore no ribofuranose is expected in the CPS. The sugar phosphate transferase gene *wcxG* and the arabinitol biosynthetic pathway genes (*abp1* and *-2*) are only found in the *cps* loci of serogroup 24 and types 17F and 48.

Serogroup 28. Serotype 28F and 28A and type 16F *cps* loci constitute a subcluster (Fig. 3), and their CPS structures (none of these are known) show immunological cross-reactions, factor serum 16b reacting with both serotypes 28F and 16F (Table 2). The type 16F *cps* locus is very similar to those of serotypes 28F and 28A, but it lacks the *gtp1* and *-3* genes and has an additional gene (*gct*) and a gene fragment (*HG191*). In addition, the *wzy* gene product of type 16F falls into a different HG. In contrast, the type 16A *cps* locus has relatively little in common with that of serotype 16F and is in a different subcluster (Fig. 1).

Serogroup 32. Serotype 32F, 32A, and 27 *cps* loci constitute a subcluster (Fig. 3), and the extensive similarity is reflected in the commonalities of their CPS structures and immunological cross-reaction with factor serum 27b (Table 2). The CPS structures of serogroup 32 differ only in the acetylation pattern (type 32A has an extra acetyl group at the α-D-Glc*p* residue), presumably due to sequence differences in the acetyltransferase WcyH (5% sequence divergence between types 32A and 32F). The *cps* locus of type 27 shows extensive similarity at the 5′ and 3′ regions, but the central *cps* region differs considerably. Following assignment of WchF to the catalysis of the L-Rhap(β1-4)-β-D-Glc*p* linkage of the repeat units, the GT WcyS is common and presumably catalyzes the D-Gal*p*-(α1-4)-β-L-Rha*p* linkage in type 27 CPS but the D-Glc*p*-(α1-4)-β-L-Rha*p* linkage in serogroup 32 CPS. The putative sugar phosphate transferase WcrN should catalyze the incorporation of choline via a P-2 linkage in type 27 CPS or a P-3 linkage in serogroup 32 CPS (1).

Serogroup 41. Serotype 41F, 41A, 17A, and 31 *cps* loci constitute a subcluster (Fig. 3). The *cps* loci of serogroup 41 differ only in the 3′ end, where type 41F has a second, nonfunctional, copy of *glf*, whereas type 41A has a defective transposase gene instead. The latter serotype also has a stop codon in the putative acetyltransferase gene *wcrX*; thus, there should be differences in the acetylation patterns (neither of the CPS structures is known). Serogroup 41 and type 31 *cps* loci differ in their IT genes (*wchA* in serogroup 41 and *wcjG* in serotype 31), and type 31 has a second copy of *glf*, which appears to be functional, and an extra acetyltransferase gene (*wcjE*) at the 3′ end, but it lacks the putative GT gene *wcrQ*.

The type 17A *cps* locus is syntenic with that of type 41F at the 5′ and 3′ ends. The formation of the similar trisaccharide backbone L-Rha*p*-(α1-4)-β-D-Glc*p*A-(β1-3)-β-D-Gal*f* of serotype 17A and L-Rha*p*-(β1-4)-D-Glc*p*A-(β1-3)-β-D-Gal*f* of serotype 31 (differing only in the anomeric status of L-Rha*p*) should be attributed to the presence of the GT genes *wcrP* and *wcrR* that are common to both *cps* loci. In all serotypes where *wciB* is present in the *cps* loci, there is a Gal*f* residue β-1,3 linked in the repeat units of the available CPS structures and we have assigned WciB as the GT that makes the β-1,3 linkage of D-Gal*f* to β-D-Glc*p* in type 17A (1). *wciB* is also present in serotype 31 and could be assigned to the Gal*f*-(β1-3)-β-L-Rha*p* linkage; however, we are not confident in the assignment of linkages in serotype 31, in the absence of more recent nuclear magnetic resonance data for the CPS structure (3).

Cluster 3. Serotypes 25A, 25F, and 38 are assigned to cluster 3, and all fall within the same subcluster (Fig. 1 and 4). Their *cps* loci are very similar and have the first four *cps* regulatory genes interrupted by a transposase gene and rearranged in order. The *cps* loci of types 25A and 25F are syntenic, and type 38 differs only in the presence of the GT WcyV instead of the truncated WcyE in the former types. None of the three structures are available, but the synteny and immunological cross-reactivity (factor sera 25b and 38a cross-react with CPS of types 25F and 38 and types 25A and 38, respectively; Table 2) suggest structural similarities and recent common ancestry.

Cluster 4. Cluster 4 includes the *cps* loci of 23 pneumococcal serotypes (Fig. 1 and 4). The presence of certain GT genes (for example, *wciB*, *wcrC*, *wciF*, *wcrD*, and *wciE*) and the commonalities in structure at the 3′ end of the *cps* loci of several of these serotypes appear to account for their inclusion in the same cluster. The cross-reactions of some typing factor sera with several serotypes of cluster 4 also reflect some structural commonalities, for example, 13b with types 13 and 29, 34b with types 34 and 35F, and 35a with types 35A, 35B, 35C, 35F, and 47F (Table 2). The *cps* loci of types 13 and 36 are not very closely related to any of the others in the cluster. The *cps* loci of the serotypes within serogroup 10 cluster together, but those of serogroups 33, 35, and 47 do not (Fig. 1). The *cps* loci of some serotypes in cluster 4 are much more similar to those of a serotype within another serogroup than the latter are to the other members of that serogroup, for example, types 35A, 35C, and 42 or 35F and 47F (Fig. 1).

Serogroup 10. Serotype 10F, 10A, 10B, and 10C *cps* loci are in the same subcluster (Fig. 4) and are divided into two syntenic pairs, types 10A and 10B and types 10C and 10F, but there are structures only for types 10A and 10F. The GT genes and the ribitol phosphate transferase gene that direct the synthesis of each linkage in the common backbone of their repeat units, and the different side branches, are discussed by Aanensen et al. (1). The putative acetyltransferase gene *wciG* is present in type 10C-10F loci but absent from type 10A-10B *cps* loci; nevertheless, type 10F CPS has not been reported to be acetylated.

Serogroup 33. There are also two syntenic pairs of *cps* loci in serogroup 33 (types 33A-33F and 33B-33D), but in this case the pairs are in different subclusters (Fig. 1 and 4), and the available CPS structures of types 33F and 33B are also quite different. The type 33C *cps* locus differs considerably from both of the above pairs and clusters separately from both of the above syntenic pairs (Fig. 1; see Fig. S1 in the supplemental material). The *cps* locus of type 33F, compared with the type 33A locus, differs only in the acetyltransferase gene (*wcjE*), which is frameshifted in type 33F but intact in 33A. The antigenic formulas of types 33F and 33A differ in an extra reaction

Downloaded from jb.asm.org at Imperial College London on July 10, 2009

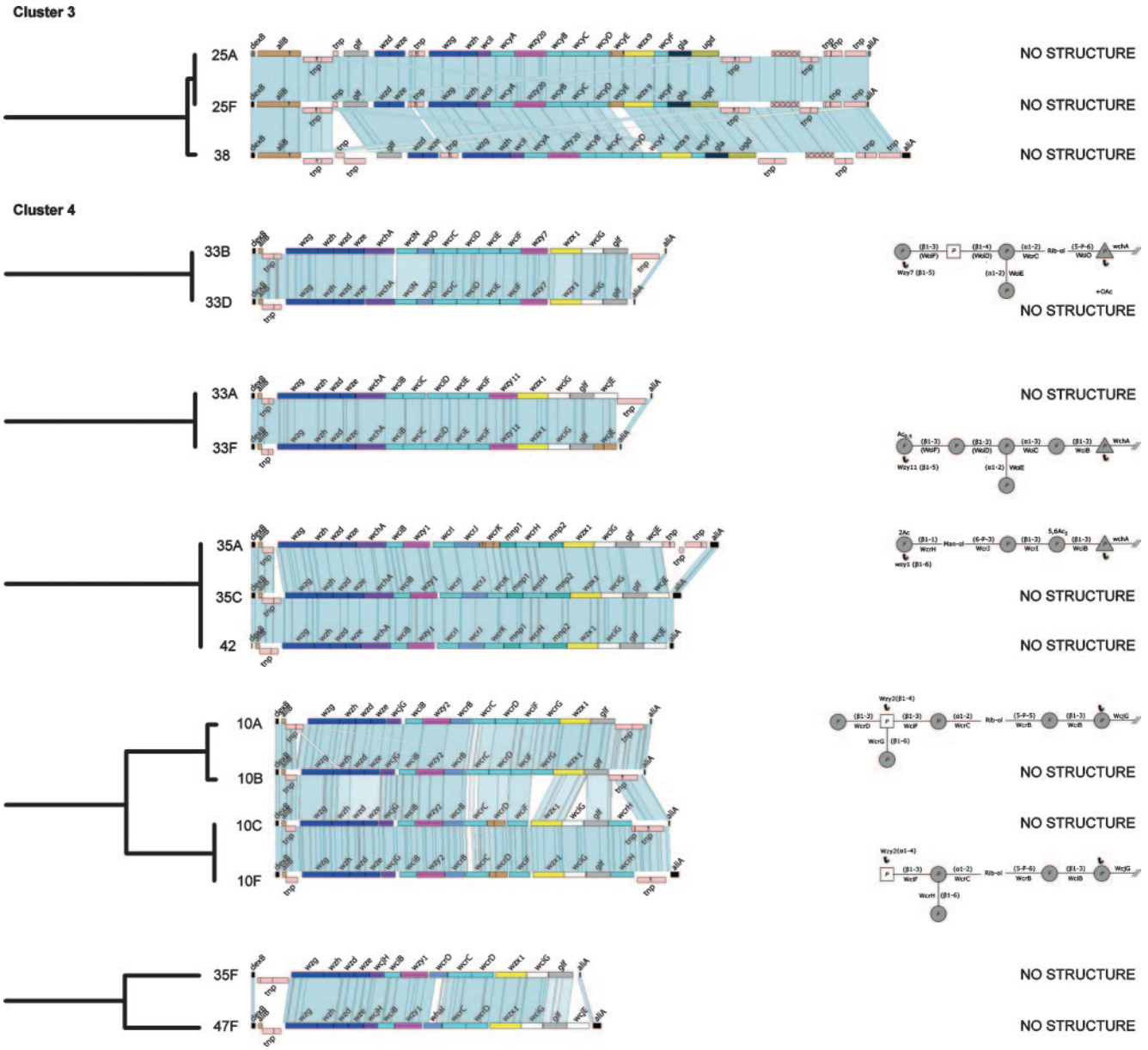


FIG. 4. Comparisons of those *cps* loci within clusters 3 and 4 assigned to the same subcluster. Details are as in Fig. 2.

of type 33A with factor serum 20b, and this serum might recognize a difference in acetylation (Table 2) (18). Type 33D reacts additionally with factor serum 6a compared to type 33B, but the difference in structure is not known as the 33D CPS structure is unavailable.

Serogroup 35. Serogroup 35, like serogroup 33, is unusual as it includes serotypes that fall into three different subclusters (Fig. 1); type 35A, 35C, and 42 *cps* loci (only the CPS structure of type 35A is known) are in the same subcluster (Fig. 4), whereas type 35F clusters apart and is very similar to serotype 47F (both CPS structures are unknown), and similarly, type 35B clusters apart and is most similar to type 29 (both structures are known). Although the serotypes within serogroup 35 fall into three subclusters, commonalities in their structures are reflected by the reactivity of factor serum 35a with all four serotypes (Table 2). The differences in the gene complements within serogroup 35 and the close similarity of some of the serotypes to unrelated serotypes are also reflected by serological cross-reactivity patterns (see below).

Except for the presence of defective transposase genes in the *cps* locus of type 35A, upstream of *aliA*, types 35A and 35C are syntenic with the type 42 *cps* locus (Fig. 4), and accordingly factor serum 35c cross-reacts with types 35A, 35C, and 42, 20b cross-reacts with types 35A and 35C, and 42a cross-reacts with types 35C and 42 (Table 2). Type 35A has a frameshift mutation in the GT gene *wcrK*, and this difference could explain the loss of reactivity with factor serum 42a.

Type 35B *cps* locus is also similar to 35A and 35C at the 5′ end (and cross-reacts with factor serum 35c) but resembles the type 29 locus in the central region (sharing *wcrJ*, *wcrM*, and *wcrH*) (Fig. 4;

Downloaded from jb.asm.org at Imperial College London on July 10, 2009

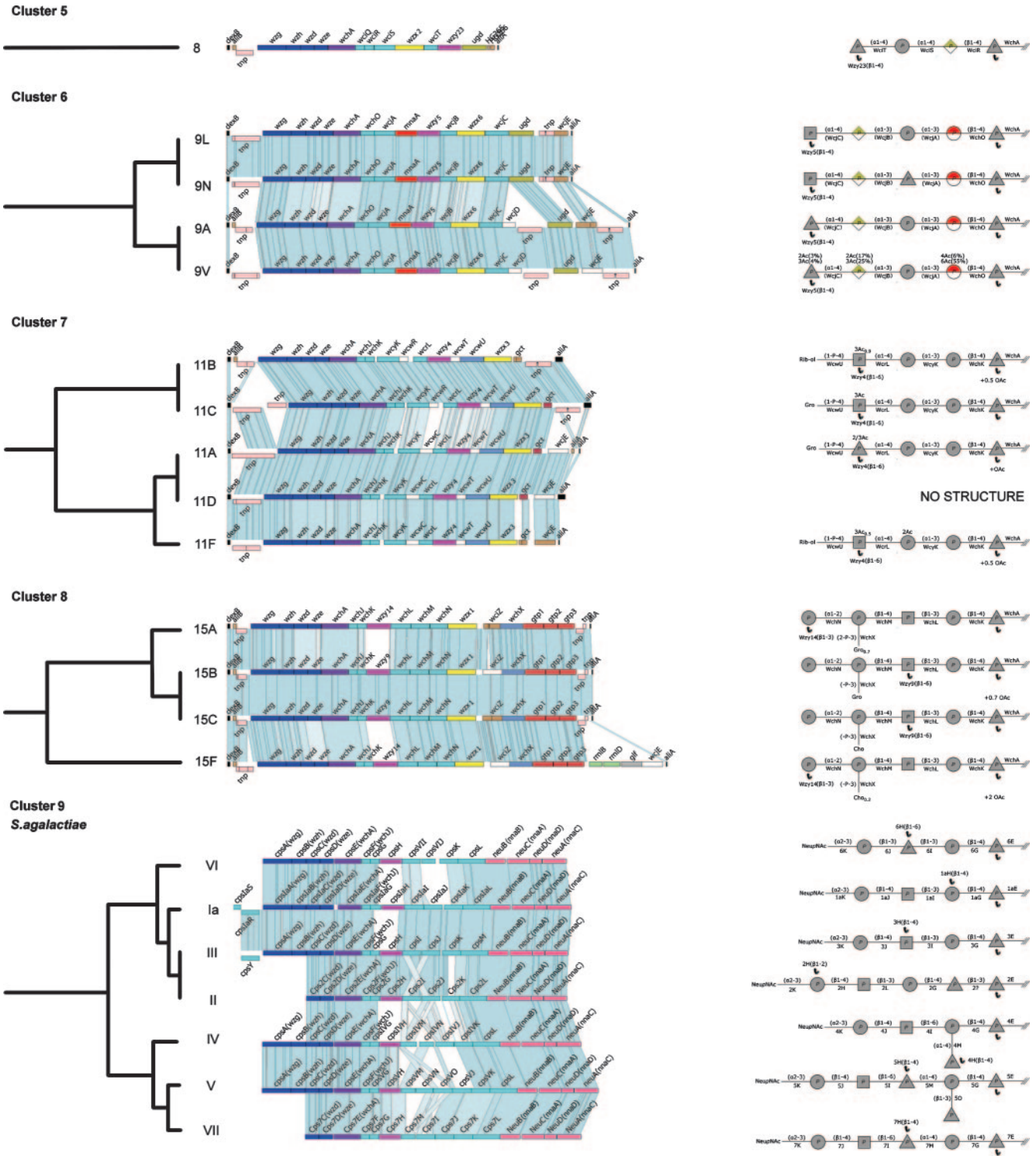


FIG. 5. Comparisons of those *cps* loci within clusters 5 to 9 assigned to the same subcluster. Details are as in Fig. 2.

see Fig. S1 in the supplemental material). The cross-reaction of types 35B and 29 with factor serum 29b is also reflected by the commonalities in their repeat units; they differ only in the initial sugar (Gal*p* transferred by WcjH and Glc*p* transferred by WchA, respectively), and the acetylation pattern of type 35B, due to the presence of an extra acetyltransferase gene (*wciG*).

Type 35F is similar at the 5′ and 3′ ends to the other serotypes within serogroup 35 but clusters with type 47F (Fig. 4), due to similarities in the central *cps* region, and accordingly, both 35F and 47F cross-react with factor sera 35a and 35b (Table 2); the CPS structures of both types are unknown. Types 35F and 47F differ only in a sugar phosphate transferase gene (type 35F has the

Downloaded from jb.asm.org at Imperial College London on July 10, 2009

wcrO gene replaced by the *whaI* gene in type 47F) and the presence of an extra acetyltransferase gene (*wcjE*) in type 47F.

Serogroup 47. As mentioned above, type 47F is more similar to type 35F than to 47A, which is in a different subcluster (Fig. 1). The central *cps* region of type 47A is similar to type 43, reflected by the cross-reaction of both serotypes with factor serum 43b; the CPS structures of both types are unknown.

Cluster 5. Serotype 8 is the only member of this cluster (Fig. 1), and the biochemical functions of the IT WchA and the GT WciS and a biosynthetic pathway have previously been suggested (31, 34, 37). WciS is an α-1,4 galactosyltransferase, WciQ and WciR are homologs of the N-terminal and C-terminal halves, respectively, of the characterized glucuronosyl β-1,4 transferase SpsK of *Sphingomonas* (35) and are suggested to catalyze the formation of the D-Glc*p*A-(β1-4)-β-D-Glc*p* linkage, and by a process of elimination, WciT is suggested to catalyze the D-Glc*p*-(α1-4)-α-D-Gal*p* linkage.

Cluster 6. The *cps* loci of the four serotypes within serogroup 9 fall into the same subcluster (Fig. 5), and they form two syntenic pairs (9A-9V and 9L-9N). A putative model of CPS biosynthesis has been suggested previously (5, 45).

Cluster 7. The five serotypes within serogroup 11 have very similar *cps* loci and fall into the same subcluster (Fig. 5). There are two syntenic groups of serogroup 11 *cps* loci, 11F-11A-11D and 11B-11C, differing only in their acetyltransferase genes, and presumably they have diverged from a recent common ancestral *cps* locus. All five serotypes possess the acetyltransferase gene *wcwT*, but types 11F, 11A, and 11D have two extra acetyltransferase genes (*wcjE* and *wcwC*) whereas types 11B and 11C have one extra acetyltransferase gene (*wcwR*). WcwU has been assigned as a glycerol phosphate transferase, and the presence of Gro-1P correlates with an intact *gct* gene in types 11A and 11C; *gct* is frameshifted in types 11F and 11B, and Rib-ol is present in the CPS instead of Gro.

The *wchJ* and *wchK* genes are present in the serogroup 11 *cps* loci. These genes are also present in serotype 14 *cps*, where biochemical studies have shown that WchK is a β-1,4-galactosyltransferase that catalyzes the linkage D-Gal*p*-(β1-4)-β-D-Glc*p*, whereas WchJ is suggested to act as an enhancer of WchK activity (23). It is therefore likely that the homologous gene products in serogroup 11 are responsible for the D-Gal*p*-(β1-4)-β-D-Glc*p* linkage found in their CPS. WcrL possesses the same Pfam domain (PF04488) as WciT of type 8, which, as mentioned above, has been suggested to catalyze the D-Glc*p*-(α1-4)-α-D-Gal*p* linkage (37). The last sugar of the repeat unit is D-Glc*p* in type 11A but D-Glc*p*NAc in types 11B, 11C, and 11F (the type 11D structure is unknown), linked in all cases via an α-1,4 bond to α-D-Gal*p*. Thus, we suggest that the last sugar of the repeat unit in serogroup 11 is transferred by WcrL and differences in amino acid sequence probably affect the specificity for the donor sugar. WcyK would therefore, by a process of elimination, catalyze the linkage D-Gal*p*-(α1-3)-β-D-Gal*p* in serogroup 11 CPS.

Cluster 8. Cluster 8 includes serogroup 15 and serotype 14, and the relatedness of their *cps* loci has been presented previously (5, 23, 44). The differences in *cps* profile and the sequences of the shared genes result in the serotype 14 *cps* locus being in a different subcluster from those of serogroup 15 (Fig. 1 and 5).

Cluster 9. Cluster 9 includes only the *cps* loci of the eight *S. agalactiae* serotypes (Fig. 1), with type VIII being the most divergent member of the cluster (see Fig. S1 in the supplemental material), as has been suggested previously (10). Only a partial sequence was available for the *cps* locus of *S. agalactiae* type Ib (46), and therefore it was not included in our study. All of these loci, except that of type VIII, fall into the same subcluster (Fig. 5). These loci are not very similar to any of the pneumococcal *cps* loci (Fig. 1), although some of the gene products of the *cps* loci of the two species fall into the same HGs, including *wzg*, *wzh*, *wzd*, *wze*, the IT gene *wchA*, and the GT gene *wchJ* (Table 1). It has been suggested that *S. agalactiae* type VIII and *S. pneumoniae* type 23F could have exchanged DNA segments due to the commonalities of their CPS structures and the sequence similarities in the central *cps* region (10, 28), and this observation was confirmed by our results. Thus, the products of the *wchF*, *wchV*, and *wzy-19* (*HG168*) genes present in the type 23F *cps* locus each fall into the same HG as the corresponding *S. agalactiae* type VIII gene products and the linkages catalyzed are also thought to be similar.

Relatedness of *S. mitis and S. oralis* RPS loci to the pneumococcal *cps* loci. The *wzg*, *wzh*, *wzd*, and *wze* genes are present in the *S. oralis* and *S. mitis* RPS biosynthetic loci (Fig. 3). They also possess the rhamnose biosynthetic genes (*rml*), although the direction of transcription of *rmlD* is reversed compared to that of the pneumococcal *cps* loci (48) and the rhamnosyltransferase gene *wchF*. The *glf* gene is also present in *S. mitis* and *S. oralis*, as are the IT gene *wchA* (the most common IT gene in pneumococci), the *wzx-1* (*HG7*) flippase gene (present in 40 pneumococcal serotypes), and several GT genes that also are found in pneumococcal *cps* loci (Table 1).

The *S. mitis* and *S. oralis* sequences differ somewhat in the central region (Fig. 3), and for example, the Wzy polymerase of *S. oralis* 34 is in the same HG as that of pneumococcal serotype 36 whereas that of *S. mitis* is unique. Overall, the RPS loci of *S. mitis* and *S. oralis* are most similar to the *cps* loci of pneumococcal serotypes 7A, 7F, and 21 (Fig. 1 and 3), and they fall in the same subcluster as the pneumococcal serotype 21 locus. The two *S. oralis* sequences are the most similar to the pneumococcal serotype 21 *cps* locus (Fig. 3) and have similar HG profiles, although the sequences of the genes encoding the shared HGs differ considerably.

DISCUSSION

Clustering of pneumococcal *cps* loci on the basis of shared gene content (shared HGs), adjusted for the sequence similarity among the shared genes, provides a useful way of identifying those *cps* loci that are the most similar. The assignment of serotypes to subclusters by using a cutoff of 0.05 identifies *cps* loci that are sufficiently similar to predict that they have diversified relatively recently from a common ancestor. In some cases, these serotypes are completely syntenic and differences in CPS structure must be due to the ability of variant forms of the same IT or GTs to catalyze different reactions, but in other serotypes in the same subcluster the *cps* loci are syntenic except for a single *cps* gene that is nonhomologous (for example, types 15A and 15B, types 23A and 23B, serogroup 25 and type 38, and types 35F and 47F). The mechanism that gave rise to this phenomenon is not understood, although gene acquisition or replacement by some illegitimate recombination event is the most likely candidate. In effect, the alternative nonhomologous genes constitute a form of polymorphism and strains of these

Downloaded from jb.asm.org at Imperial College London on July 10, 2009

serotypes should interconvert by recombination in the flanking common genes, as has been shown to occur for the interconversion of serotypes 6A and 6B, which differ by a single amino acid polymorphism (27).

Cluster analysis also provides a framework for correlating *cps* gene content, and gene function, of very similar *cps* loci with the known CPS structures and the immunochemical similarities inferred from cross-reactivity with factor typing sera. A single difference in gene content between *cps* loci, that correlates with a single difference in the CPS structure, provides a strong prediction of the specific reaction catalyzed by the products of these alternative genes, and in many cases prediction can be extended to situations where more that one gene differs between very similar *cps* loci, but in most cases predictions must be made by using comparative methods (1). The uncertainties in predicting the precise functions of HGs (particularly the specificity of the transferases) have been discussed elsewhere (1). However, there is little option but to predict function as extremely few of the gene products involved in pneumococcal CPS biosynthesis have been the subject of any biochemical study. We stress that many of our assignments of specific linkages to gene products are tentative (the basis for these assignments is presented by Aanensen et al. [1]), but we expect most to be correct. We provide them as they form a basis for further biochemical or structural work to firmly establish the linkages catalyzed by individual GTs.

The similarities between the genes of the *S. pneumoniae cps* loci and the polysaccharide biosynthetic loci of other streptococci have been reported previously (7, 25, 43, 51), and TribeMCL proved a useful tool to identify these homologies (Table 1). Cluster analysis allowed us to recognize the similarities between the *S. mitis* and *S. oralis* RPS loci and those of several pneumococcal serotypes, notably serotype 21, suggesting a history of interspecies exchange of polysaccharide biosynthetic loci.

In general, the immunological subdivision of pneumococci into serotypes is supported by the structures of the *cps* loci. Thus, all serotypes in the same serogroup were in the same cluster and in many cases were in the same subcluster. In those cases where serotypes in the same serogroup are syntenic, it is assumed that selection imposed by the host immune response has led to the accumulation of nonsynonymous substitutions (or in some case frameshift mutations) in one or more of the shared genes, which leads to a difference in their CPS structure, which is recognized by the typing sera. There are, however, several instances in which serotypes in the same serogroup have dissimilar *cps* loci and, conversely, where the *cps* loci of completely different serotypes are very similar. In these cases, a classification based on genetic similarity would be quite different from that which has been developed by serology. For example, serotypes 44 and 46 could be reassigned as serotypes within serogroup 12, and in other cases serotypes within the same serogroup could more logically be placed in different serogroups.

However, lumping different serotypes together, or splitting serogroups, would cause confusion and there are no strong arguments for realigning serotypes and serogroups based on genetic similarities or differences between *cps* loci. Indeed, in the context of pneumococcal conjugate vaccines (6), it is the extent of immunological cross-reactivity and the resulting presence or absence of cross-protection between serotypes that matter, and this is more likely to be captured by the current serotyping scheme than by genetic similarities and differences between *cps* loci. The serological cross-reactivity between serotypes has been defined in the rabbit, and cross-reactivity between serotypes may be different for antibodies produced in humans, but the genetic similarities between some completely different serotypes and common reactions with some typing sera suggest that there may be situations where antibodies induced in humans by CPS of one serotype might provide some cross-protection against CPS of completely different serotypes.

Conjugate vaccines that include the CPS from a limited number of serotypes (6) are likely to be the only effective way of protecting against pneumococcal disease until vaccines are developed that use antigens which can protect against disease caused by all pneumococcal strains. The availability of the sequences of the *cps* loci of all 90 serotypes and the analysis of the relatedness of the *cps* loci of the 88 serotypes that use the Wzy-dependent pathway will provide the basis to develop novel molecular serotyping methods to monitor changes in the serotypes of pneumococci following the introduction of conjugate vaccines. More generally, the clustering approach we used here can be applied to look at the relatedness of other genetic elements that share variable numbers of genes with various levels of sequence similarity.

ACKNOWLEDGMENTS

This work was funded by the Wellcome Trust. B.G.S. is a Wellcome Trust Principal Research Fellow.

REFERENCES

1. **Aanensen, D. M., A. Mavroidi, S. D. Bentley, P. R. Reeves, and B. G. Spratt.** 2007. Predicted functions and linkage specificities of the products of the *Streptococcus pneumoniae* capsular biosynthetic loci. J. Bacteriol. **189:**7856–7876.
2. **Abbott, J. C., D. M. Aanensen, K. Rutherford, S. Butcher, and B. G. Spratt.** 2005. WebACT—an online companion for the Artemis comparison tool. Bioinformatics **21:**3665–3666.
3. **Batavyal, L., and N. Roy.** 1983. Structure of the capsular polysaccharide of *Diplococcus pneumoniae* type 31. Carbohydr. Res. **119:**300–302.
4. **Bateman, A., L. Coin, R. Durbin, R. D. Finn, V. Hollich, S. Griffiths-Jones, A. Khanna, M. Marshall, S. Moxon, E. L. Sonnhammer, D. J. Studholme, C. Yeats, and S. R. Eddy.** 2004. The Pfam protein families database. Nucleic Acids Res. **32:**D138–141.
5. **Bentley, S. D., D. M. Aanensen, A. Mavroidi, D. Saunders, E. Rabbinowitsch, M. Collins, K. Donohoe, D. Harris, L. Murphy, M. A. Quail, G. Samuel, I. C. Skovsted, M. S. Kaltoft, B. Barrell, P. R. Reeves, J. Parkhill, and B. G. Spratt.** 2006. Genetic analysis of the capsular biosynthetic locus from all 90 pneumococcal serotypes. PLoS Genet. **2:**e31.
6. **Bogaert, D., P. W. Hermans, P. V. Adrian, H. C. Rumke, and R. de Groot.** 2004. Pneumococcal vaccines: an update on current strategies. Vaccine **22:**2209–2220.
7. **Bourgoin, F., A. Pluvinet, B. Gintz, B. Decaris, and G. Guedon.** 1999. Are horizontal transfers involved in the evolution of the *Streptococcus thermophilus* exopolysaccharide synthesis loci? Gene **233:**151–161.
8. **Broadbent, J. R., D. J. McMahon, D. L. Welker, C. J. Oberg, and S. Moineau.** 2003. Biochemistry, genetics, and applications of exopolysaccharide production in *Streptococcus thermophilus*: a review. J. Dairy Sci. **86:**407–423.
9. **Carver, T. J., K. M. Rutherford, M. Berriman, M. A. Rajandream, B. G. Barrell, and J. Parkhill.** 2005. ACT: the Artemis comparison tool. Bioinformatics **21:**3422–3423.
10. **Cieslewicz, M. J., D. Chaffin, G. Glusman, D. Kasper, A. Madan, S. Rodrigues, J. Fahey, M. R. Wessels, and C. E. Rubens.** 2005. Structural and genetic diversity of group B streptococcus capsular polysaccharides. Infect. Immun. **73:**3096–3103.
11. **Cieslewicz, M. J., D. L. Kasper, Y. Wang, and M. R. Wessels.** 2001. Functional analysis in type Ia group B Streptococcus of a cluster of genes involved in extracellular polysaccharide production by diverse species of streptococci. J. Biol. Chem. **276:**139–146.
12. **Cisar, J. O., A. L. Sandberg, G. P. Reddy, C. Abeygunawardana, and C. A.**

Downloaded from jb.asm.org at Imperial College London on July 10, 2009

Bush. 1997. Structural and antigenic types of cell wall polysaccharides from viridans group streptococci with receptors for oral actinomyces and streptococcal lectins. Infect. Immun. **65:**5035–5041.
13. **Coyne, M. J., A. O. Tzianabos, B. C. Mallory, V. J. Carey, D. L. Kasper, and L. E. Comstock.** 2001. Polysaccharide biosynthesis locus required for virulence of *Bacteroides fragilis*. Infect. Immun. **69:**4342–4350.
14. **Enright, A. J., S. Van Dongen, and C. A. Ouzounis.** 2002. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. **30:**1575–1584.
15. **García, E., D. Llull, R. Muñoz, M. Mollerach, and R. López.** 2000. Current trends in capsular polysaccharide biosynthesis of *Streptococcus pneumoniae*. Res. Microbiol. **151:**429–435.
16. **Heidelberger, M.** 1983. Precipitating cross-reactions among pneumococcal types. Infect. Immun. **41:**1234–1244.
17. **Henrichsen, J.** 1995. Six newly recognized types of *Streptococcus pneumoniae*. J. Clin. Microbiol. **33:**2759–2762.
18. **Jiang, S. M., L. Wang, and P. R. Reeves.** 2001. Molecular characterization of *Streptococcus pneumoniae* type 4, 6B, 8, and 18C capsular polysaccharide gene clusters. Infect. Immun. **69:**1244–1255.
19. **Jobson, J. D.** 1992. Applied multivariate data analysis. Volume II. Categorical and multivariate methods, p. 483–568. Springer-Verlag, New York, NY.
20. **Jones, C., and X. Lemercinier.** 2005. Full NMR assignment and revised structure for the capsular polysaccharide from *Streptococcus pneumoniae* type 15B. Carbohydr. Res. **340:**403–409.
21. **Jones, C., C. Whitley, and X. Lemercinier.** 2000. Full assignment of the proton and carbon NMR spectra and revised structure for the capsular polysaccharide from *Streptococcus pneumoniae* type 17F. Carbohydr. Res. **325:**192–201.
22. **Kamerling, J. P.** 2000. Pneumococcal polysaccharides: a chemical view, p. 81–114. *In* A. Tomasz (ed.), *Streptococcus pneumoniae*: molecular biology and mechanisms of disease. Mary Ann Liebert Inc., Larchmont, NY.
23. **Kolkman, M. A., B. A. van der Zeijst, and P. J. Nuijten.** 1997. Functional analysis of glycosyltransferases encoded by the capsular polysaccharide biosynthesis locus of *Streptococcus pneumoniae* serotype 14. J. Biol. Chem. **272:**19502–19508.
24. **Lemercinier, X., and C. Jones.** 2006. Full assignment of the ^{1}H and ^{13}C spectra and revision of the O-acetylation site of the capsular polysaccharide of *Streptococcus pneumoniae* type 33F, a component of the current pneumococcal polysaccharide vaccine. Carbohydr. Res. **341:**68–74.
25. **Llull, D., R. López, and E. García.** 2001. Genetic bases and medical relevance of capsular polysaccharide biosynthesis in pathogenic streptococci. Curr. Mol. Med. **1:**475–491.
26. **López, R., and E. García.** 2004. Recent trends on the molecular biology of pneumococcal capsules, lytic enzymes, and bacteriophage. FEMS Microbiol. Rev. **28:**553–580.
27. **Mavroidi, A., D. Godoy, D. M. Aanensen, D. A. Robinson, S. K. Hollingshead, and B. G. Spratt.** 2004. Evolutionary genetics of the capsular locus of serogroup 6 pneumococci. J. Bacteriol. **186:**8181–8192.
28. **Morona, J. K., D. C. Miller, T. J. Coffey, C. J. Vindurampulle, B. G. Spratt, R. Morona, and J. C. Paton.** 1999. Molecular and genetic characterization of the capsule biosynthesis locus of *Streptococcus pneumoniae* type 23F. Microbiology **145:**781–789.
29. **Morona, J. K., R. Morona, and J. C. Paton.** 1999. Comparative genetics of capsular polysaccharide biosynthesis in *Streptococcus pneumoniae* types belonging to serogroup 19. J. Bacteriol. **181:**5355–5364.
30. **Mulrooney, E. F., K. K. Poon, D. J. McNally, J. R. Brisson, and J. S. Lam.** 2005. Biosynthesis of UDP-*N*-acetyl-L-fucosamine, a precursor to the biosynthesis of lipopolysaccharide in *Pseudomonas aeruginosa* serotype O11. J. Biol. Chem. **280:**19535–19542.
31. **Muñoz, R., M. Mollerach, R. López, and E. García.** 1999. Characterization of the type 8 capsular gene cluster of *Streptococcus pneumoniae*. J. Bacteriol. **181:**6214–6219.
32. **Needleman, S. B., and C. D. Wunsch.** 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol. **48:**443–453.
33. **Park, I. H., D. G. Pritchard, R. Cartee, A. Brandao, M. C. Brandileone, and M. H. Nahm.** 2007. Discovery of a new capsular serotype (6C) within serogroup 6 of *Streptococcus pneumoniae*. J. Clin. Microbiol. **45:**1225–1233.
34. **Pelosi, L., M. Boumedienne, N. Saksouk, J Geiselmann, and R. A. Geremia.** 2005. The glucosyl-1-phosphate transferase WchA (Cap8E) primes the capsular polysaccharide repeat unit biosynthesis of *Streptococcus pneumoniae* serotype 8. Biochem. Biophys. Res. Commun. **327:**857–865.
35. **Pollock, T. J., W. A. van Workum, L. Thorne, M. J. Mikolajczak, M. Yamazaki, J. W. Kijne, and R. W. Armentrout.** 1998. Assignment of biochemical functions to glycosyl transferase genes which are essential for biosynthesis of exopolysaccharides in *Sphingomonas* strain S88 and *Rhizobium leguminosarum*. J. Bacteriol. **180:**586–593.
36. **Reeves, P. R., M. Hobbs, M. A. Valvano, M. Skurnik, C. Whitfield, D. Coplin, N. Kido, J. Klena, D. Maskell, C. R. Raetz, and P. D. Rick.** 1996. Bacterial polysaccharide synthesis and gene nomenclature. Trends Microbiol. **4:**495–503.
37. **Saksouk, N., L. Pelosi, P. Colin-Morel, M. Boumedienne, P. L. Abdian, and R. A. Geremia.** 2005. The capsular polysaccharide biosynthesis of *Streptococcus pneumoniae* serotype 8: functional identification of the glycosyltransferase WciS (Cap8H). Biochem. J. **389:**63–72.
38. **Samuel, G., and P. Reeves.** 2003. Biosynthesis of O-antigens: genes and pathways involved in nucleotide sugar precursor synthesis and O-antigen assembly. Carbohydr. Res. **338:**2503–2519.
39. **Smith, H. E., V. Veenbergen, J. van der Velde, M. Damman, H. J. Wisselink, and M. A. Smits.** 1999. The *cps* genes of *Streptococcus suis* serotypes 1, 2, and 9: development of rapid serotype-specific PCR assays. J. Clin. Microbiol. **37:**3146–3152.
40. **Spratt, B. G., W. P. Hanage, and A. B. Bruegemann.** 2004. Evolutionary and population biology of *Streptococcus pneumoniae*, p. 119–135. *In* E. I. Tuomanen (ed.), The pneumococcus. ASM Press, Washington, D.C.
41. **Tekaia, F., A. Lazcano, and B. Dujon.** 1999. The genomic tree as revealed from whole proteome comparisons. Genome Res. **9:**550–557.
42. **Tettelin, H., K. E. Nelson, I. T. Paulsen, J. A. Eisen, T. D. Read, S. Peterson, J. Heidelberg, R. T. DeBoy, D. H. Haft, R. J. Dodson, A. S. Durkin, M. Gwinn, J. F. Kolonay, W. C. Nelson, J. D. Peterson, L. A. Umayam, O. White, S. L. Salzberg, M. R. Lewis, D. Radune, E. Holtzapple, H. Khouri, A. M. Wolf, T. R. Utterback, C. L. Hansen, L. A. McDonald, T. V. Feldblyum, S. Angiuoli, T. Dickinson, E. K. Hickey, I. E. Holt, B. J. Loftus, F. Yang, H. O. Smith, J. C. Venter, B. A. Dougherty, D. A. Morrison, S. K. Hollingshead, and C. M. Fraser.** 2001. Complete genome sequence of a virulent isolate of *Streptococcus pneumoniae*. Science **293:**498–506.
43. **van Kranenburg, R., H. R. Vos, I. I. van Swam, M. Kleerebezem, and W. M. de Vos.** 1999. Functional analysis of glycosyltransferase genes from *Lactococcus lactis* and other gram-positive cocci: complementation, expression, and diversity. J. Bacteriol. **181:**6347–6353.
44. **van Selm, S., L. M. van Cann, M. A. Kolkman, B. A. van der Zeijst, and J. P. van Putten.** 2003. Genetic basis for the structural difference between *Streptococcus pneumoniae* serotype 15B and 15C capsular polysaccharides. Infect. Immun. **71:**6192–6198.
45. **van Selm, S., M. A. Kolkman, B. A. van der Zeijst, K. A. Zwaagstra, W. Gaastra, and J. P. van Putten.** 2002. Organization and characterization of the capsule biosynthesis locus of *Streptococcus pneumoniae* serotype 9V. Microbiology **148:**1747–1755.
46. **Watanabe, M., K. Miyake, K. Yanae, Y. Kataoka, S. Koizumi, T. Endo, A. Ozaki, and S. Iijima.** 2002. Molecular characterization of a novel beta1,3-galactosyltransferase for capsular polysaccharide synthesis by *Streptococcus agalactiae* type Ib. J. Biochem. (Tokyo) **131:**183–191.
47. **Watson, D. A., D. M. Musher, and J. Verhoef.** 1995. Pneumococcal virulence factors and host immune responses to them. Eur. J. Clin. Microbiol. Infect. Dis. **14:**479–490.
48. **Xu, D. Q., J. Thompson, and J. O. Cisar.** 2003. Genetic loci for coaggregation receptor polysaccharide biosynthesis in *Streptococcus gordonii* 38. J. Bacteriol. **185:**5419–5430.
49. **Yoshida, Y., S. Ganguly, C. A. Bush, and J. O. Cisar.** 2005. Carbohydrate engineering of the recognition motifs in streptococcal co-aggregation receptor polysaccharides. Mol. Microbiol. **58:**244–256.
50. **Yoshida, Y., S. Ganguly, C. A. Bush, and J. O. Cisar.** 2006. Molecular basis of L-rhamnose branch formation in streptococcal coaggregation receptor polysaccharides. J. Bacteriol. **188:**4125–4130.
51. **Yother, J.** 2004. Capsules, p. 30–48. *In* E. I. Tuomanen (ed.), The pneumococcus. ASM Press, Washington, D.C.

Downloaded from jb.asm.org at Imperial College London on July 10, 2009

W728–W733 *Nucleic Acids Research, 2005, Vol. 33, Web Server issue*
doi:10.1093/nar/gki415

The multilocus sequence typing network: mlst.net

David M. Aanensen* and Brian G. Spratt

Department of Infectious Disease Epidemiology, Imperial College London, St Mary's Hospital, London W2 1PG, UK

Received February 14, 2005; Revised and Accepted March 21, 2005

ABSTRACT

The unambiguous characterization of strains of a pathogen is crucial for addressing questions relating to its epidemiology, population and evolutionary biology. Multilocus sequence typing (MLST), which defines strains from the sequences at seven house-keeping loci, has become the method of choice for molecular typing of many bacterial and fungal pathogens (and non-pathogens), and MLST schemes and strain databases are available for a growing number of prokaryotic and eukaryotic organisms. Sequence data are ideal for strain characterization as they are unambiguous, meaning strains can readily be compared between laboratories via the Internet. Laboratories undertaking MLST can quickly progress from sequencing the seven gene fragments to characterizing their strains and relating them to those submitted by others and to the population as a whole. We provide the gateway to a number of MLST schemes, each of which contain a set of tools for the initial characterization of strains, and methods for relating query strains to other strains of the species, including clustering based on differences in allelic profiles, phylogenetic trees based on concatenated sequences, and a recently developed method (eBURST) for identifying clonal complexes within a species and displaying the overall structure of the population. This network of MLST websites is available at http://www.mlst.net

INTRODUCTION

Multilocus sequence typing (MLST) is a nucleotide sequence-based approach to the unambiguous characterization of strains of bacterial species, or other microbial species, via the Internet (1,2). MLST involves obtaining the sequences of internal fragments of seven house-keeping genes for each strain of a particular species. The sequences of each fragment are compared with all the previously identified sequences (alleles) at that locus and, thereby, are assigned allele numbers at each of the seven loci. The combination of the seven allele numbers defines the allelic profile of the strain and each different allelic profile is assigned as a sequence type (ST), which is used to describe the strain.

Nucleotide sequencing is relatively cheap, and easy to perform. The data produced by MLST are ideal for the characterization of strains of bacterial or fungal species via a web server. MLST is now widely used for molecular epidemiology as it allows strains studied by different groups to be compared and MLST schemes have been developed for ~20 bacteria (mostly pathogens) (3), and three fungi (4,5) and databases that can be queried have been available for several years (6). The MLST databases are currently hosted on two main web servers located at Imperial College London (http://www.mlst.net) and Oxford University [http://pubmlst.org; (7)]. The former web server acts as a gateway to a number of species-specific websites each of which contains tools for the analysis of allele sequences and STs, and a web interface for obtaining epidemiological information held on the increasing numbers of strains that are submitted by the user community.

Along with centrally available tools for those interested in starting their own MLST schemes, such as for defining alleles using non-redundant databases (NRDB), measuring linkage disequilibrium and an interface to Splits Tree (8), http://www.mlst.net provides a number of options to display the relatedness of query strains to those in the strain database.

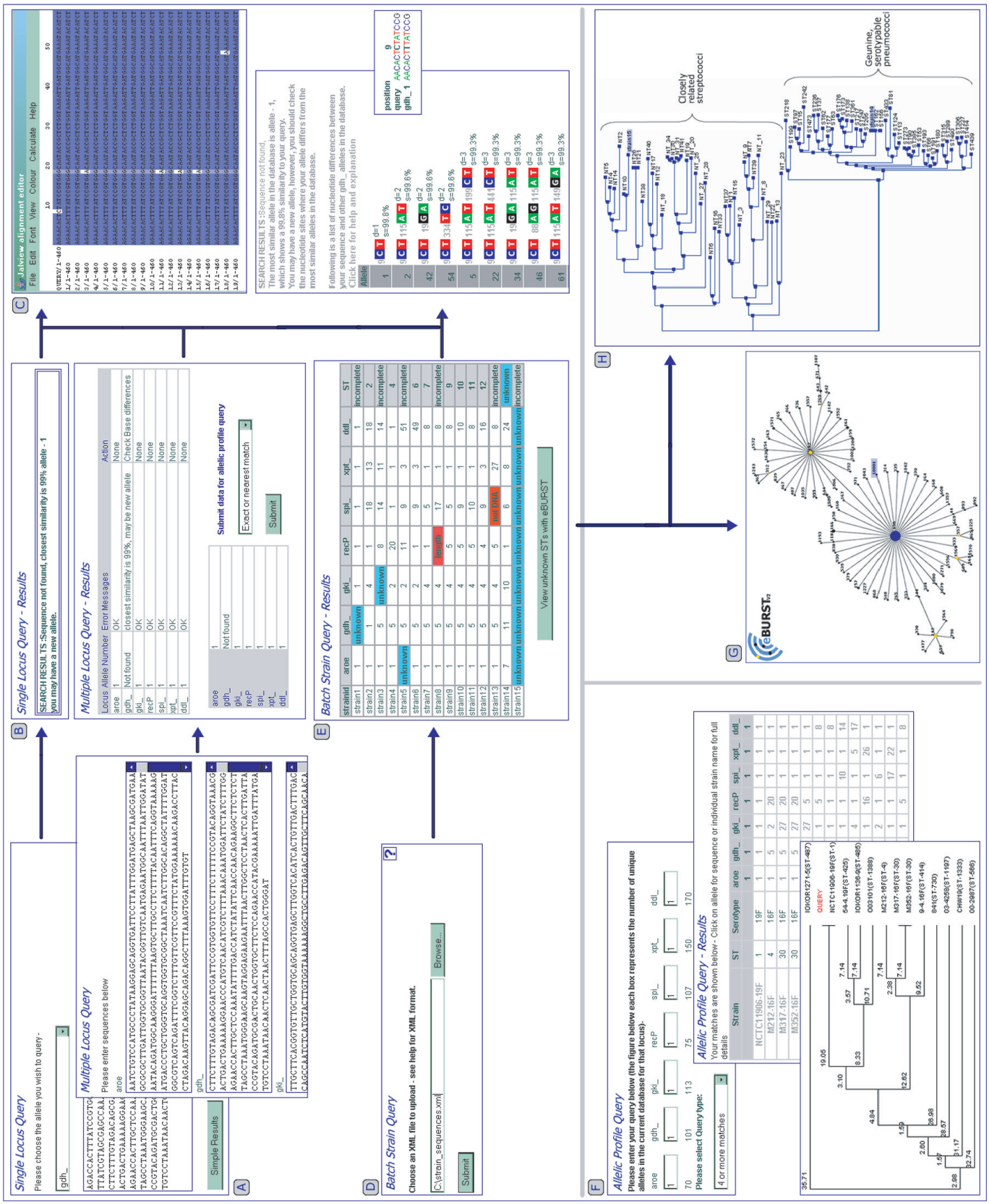
MLST WORKFLOW

Laboratories undertaking MLST can access species-specific information on each of the individual mlst.net species websites, including sequencing protocols and primer sequences, allowing a laboratory to begin producing data rapidly. Characterization of a strain requires the generation of the sequences of the seven gene fragments and, once these are available, they are used to query the appropriate mlst.net website, to assign the alleles at each locus and thereby to obtain the allelic profile of the query strain. Each MLST website holds the sequences of all known alleles at each of the seven loci, and all known allelic profiles, and through the curator assigns new allele numbers and STs. Every different sequence at each locus is assigned as a distinct allele and new alleles are assigned allele numbers by the curator and are entered in the allele database.

*To whom correspondence should be addressed. Tel: +44 0 20 7594 3825; Fax: +44 0 20 7594 3693; Email: d.aanensen@imperial.ac.uk

© The Author 2005. Published by Oxford University Press. All rights reserved.

The online version of this article has been published under an open access model. Users are entitled to use, reproduce, disseminate, or display the open access version of this article for non-commercial purposes provided that: the original authorship is properly and fully attributed; the Journal and Oxford University Press are attributed as the original place of publication with the correct citation details given; if an article is subsequently reproduced or disseminated not in its entirety but only in part or as a derivative work this must be clearly indicated. For commercial re-use, please contact journals.permissions@oupjournals.org


A
Single Locus Query
Please choose the allele you wish to query -
gdh_
Multiple Locus Query
Please enter sequences below
aroe
gdh_
gki_
recP
spi_
xpt_
ddl_
Simple Results
B
Single Locus Query - Results
SEARCH RESULTS :Sequence not found, closest similarity is 99% allele - 1
you may have a new allele.
Multiple Locus Query - Results
Locus Allele Number Error Messages Action
aroe 1 OK None
gdh_ Not found closest similarity is 99%, may be new allele Check Base differences
gki_ 1 OK None
recP 1 OK None
spi_ 1 OK None
xpt_ 1 OK None
ddl_ 1 OK None
Submit data for allelic profile query
Exact or nearest match
Submit
C
Jalview alignment editor
File Edit Font View Colour Calculate Help
SEARCH RESULTS :Sequence not found.
The most similar allele in the database is allele - 1,
which shows a 99.8% similarity to your query.
You may have a new allele, however, you should check
the nucleotide sites where your allele differs from the
most similar alleles in the database.
Following is a list of nucleotide differences between
your sequence and other gdh_ alleles in the database.
Click here for help and explanation
Allele
position 9
query A4CACTCTATCCG
gdh_1 A4CACTTATCCG
D
Batch Strain Query
Choose an XML file to upload - see help for XML format.
C:\strain_sequences.xml
Browse...
Submit
E
Batch Strain Query - Results
strainid aroe gdh_ gki_ recP spi_ xpt_ ddl_ ST
View unknown STs with eBURST
F
Allelic Profile Query
Please enter your query below (the figure below each box represents the number of unique alleles in the current database for that locus).
aroe gdh_ gki_ recP spi_ xpt_ ddl_
Please select Query type:
4 or more matches
Allelic Profile Query - Results
Your matches are shown below - Click on allele for sequence or individual strain name for full details
Strain ST Serotype aroe gdh_ gki_ recP spi_ xpt_ ddl_
NCTC11906-19F 1 19F 1 1 1 1 1 1 1
M212-16F 4 16F 1 5 2 20 1 1 1
M317-16F 30 16F 1 5 27 20 1 1 1
M352-16F 30 16F 1 5 27 20 1 1 1
G
eBURST
H
Closely related streptococci
Geunine, serotypable pneumococci

Each MLST species website offers a number of analysis steps for a user. First, alleles have to be assigned from the sequence data by one of three options (Figure 1A and D):

Single/batch locus query: allowing a single sequence or a batch of sequences for a single locus to be compared with all known alleles.

Multiple locus query: allowing the input of the sequences of all seven loci for a single strain.

Batch strain query: allowing input of the sequences of all seven loci for a batch of strains.

In all cases, the user's sequence is checked for correct length for that locus, and for the absence of unexpected characters, and is then queried against all other sequences in the species database. For *Candida albicans*, a diploid organism, the standard ambiguity codes are allowed and are used to assign heterozygous nucleotide sites (4).

If the user's sequence is found, the allele number is returned, whereas if the user has a novel sequence, the percentage identity to the closest allele in the database is returned and the user is advised to check carefully those nucleotide sites that differ from the most similar allele or alleles in the database (Figure 1B). This can be carried out using the Jalview alignment editor (9), or the nucleotide differences can be displayed between the query sequence and the most similar alleles, as in Figure 1C. The latter method allows the user to view the flanking sequence around each nucleotide difference between the query and the most similar alleles, allowing the trace files of their proposed new allele to be searched easily for any potential ambiguities or sequencing errors. If a user is confident that they have a new allele, the forward and reverse trace files are submitted to the MLST scheme curator, as a quality control check, before a new allele number is assigned by the curator for the novel sequence. The sequence of the new allele is then entered into the database.

Repeating this process for each locus provides the seven-digit allelic profile for the query strain. The seven-digit allelic profile can then be entered into the allelic profile query to discover if the strain is identical, or similar, in allelic profile to any of the strains already in the database (Figure 1F). The multiple locus query represents a batch processing method for a single strain, allowing all seven sequences for the query strain to be entered at once and for the allele numbers at each locus to be returned.

BATCH STRAIN QUERY

The repeated querying of single strains becomes very time consuming for laboratories undertaking MLST on many hundred strains of a particular species. The need to analyze sequences from multiple genes from a large number of strains at the same time precludes the use of standard sequence formats such as FASTA or MEGA. Therefore, we use a simple XML format that allows the batch processing of hundreds of strains at one time (Figure 2).

```
<dataroot>
<strain_sequences>
<strainid>STRAIN1</strainid>
<locus1>...GCCAGTGA...</locus1>
<locus2>...CCGAGTGA...</locus2>
<locus3>...GCGAGTGA...</locus3>
<locus4>...ACGGGTGA...</locus4>
<locus5>...TCGAGGGA...</locus5>
<locus6>...GCGAGTGA...</locus6>
<locus7>...CCGTGAGA...</locus7>
</strain_sequences>                    } STRAIN1
<strain_sequences>
<strainid>STRAIN2</strainid>
<locus1>...GCCAGTGA...</locus1>
<locus2>...CCGAGTGA...</locus2>
<locus3>...GCGAGTGA...</locus3>
<locus4>...ACGGGTGA...</locus4>
<locus5>...TCGAGGGA...</locus5>
<locus6>...GCGAGTGA...</locus6>
<locus7>...CCGTGAGA...</locus7>
</strain_sequences>                    } STRAIN2
  etc.....
</dataroot>
```

Figure 2. The XML format for batch querying multiple strains using mlst.net.

Formatting the input data with a basic XML wrapper around a set of seven sequences for each strain allows a user to produce a file, for an unlimited number of strains, that can be used for batch processing. To aid production of such a format, each of the MLST species subsites at http://www.mlst.net provides a modified Access database that allows users to store their sequence and strain information in one place, and allows the data to be exported in bulk in the correct XML format, without the need for a user to manually produce the document. Furthermore, for sequencing laboratories using the STARS (http://www.molbiol.ox.ac.uk/~paediat/stars/) platform for MLST, we provide a facility to convert the FASTA files generated by STARS into the XML format via a web form.

When a user uploads the generated XML file (Figure 1D), the sequences for each of the seven loci in all of the strains are checked for invalid characters and correct length. Each sequence is then queried against the appropriate allele database (Figure 1E). If found, the allele number is returned and, if unknown, the user can look further into the sequence differences between the query allele and the most similar alleles in the database (Figure 1C). If all the seven loci are found, the allelic profile of the strain is queried against a look-up table of STs within the database and, if a match is found, the ST number is returned. If the allelic profile is previously unknown this information is returned. The batch procedure, therefore, automatically returns a table with the alleles, allelic profiles

Figure 1. Schematic representation of a typical MLST workflow. (**A**) Sequences can be entered locus by locus (single locus query), or all seven loci from a single strain (multiple locus query), or (D) by uploading a XML file with a set of strains and their sequences (batch strain query). (**B**) For the single and multiple locus queries, sequences that are not in the database are identified, and can be compared with the sequences of all known alleles using Jalview (7), or (**C**) the nucleotide differences compared with the most similar alleles can be displayed. The batch strain query (**D**) returns a strain table (**E**), which shows the allele number for each locus if known and the allelic profile if all the seven alleles and the ST are known. Strains that have the most similar allelic profiles to query strains are displayed as a table or by cluster analysis (**F**), and further information about them can be obtained. For the pneumococcal example used here, the query strains can be compared with the reference set of pneumococcal strains and closely related streptococcal strains, to establish whether or not they are pneumococci, using the concatenated sequences to construct a neighbor-joining tree (**H**). The relationship of unknown strains to the whole population can also be investigated using eBURST (**G**).

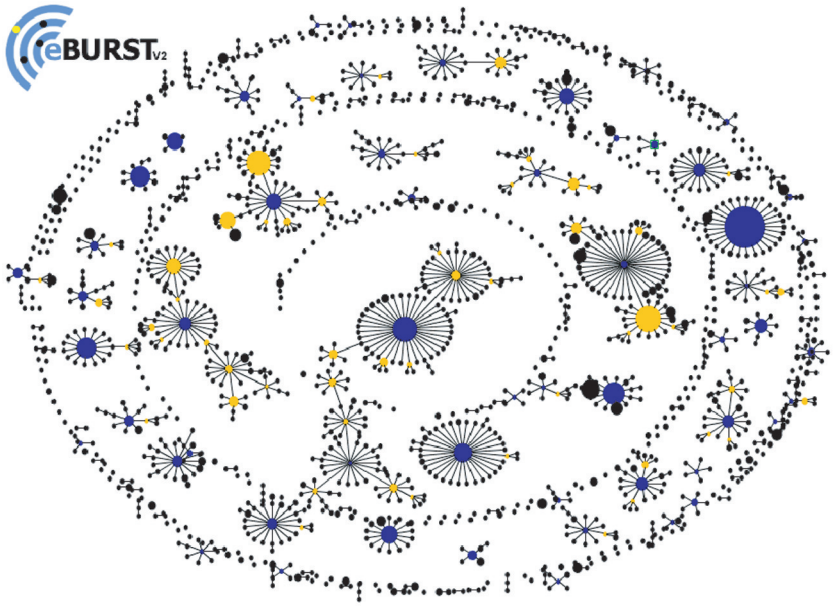


Figure 3. A population snapshot of the entire *S.pneumoniae* MLST database showing all major and minor clonal complexes viewed using eBURST.

and STs of all the input strains, flagging up those alleles and STs that are previously unknown (Figure 1E).

Comparing query strains to the database: clustering using allelic profiles

The simplest approach is to identify those strains in the database that have some minimum level of similarity in their allelic profile to each query strain (e.g. sharing alleles at ⩾4 of the seven loci), and to show the relationship of the query strain to those returned from the database query using a dendrogram, based on the matrix of pairwise differences between the allelic profiles of the strains (Figure 1F).

Comparing query strains to the database: using eBURST

Traditionally, dendrograms have been the method of choice for displaying the implied relationships between strains of a bacterial population or species. However, although dendrograms are good at visualizing the clusters of identical or very similar strains, the bifurcating process of lineage splitting implied by a dendrogram is a very poor representation of the way in which bacterial lineages emerge and diversify. A new algorithm, BURST (10), was recently introduced that does not impose a tree-like pattern of descent, but rather uses an appropriate model of recent bacterial evolution. In addition, it is very difficult to display the relatedness of all strains in a large MLST database, including thousands of STs, on a dendrogram, and better ways of displaying the relationships among all strains in large MLST databases are required.

Briefly, the model incorporated into BURST assumes that, due to selection or genetic drift, some genotypes will occasionally increase in frequency in the population and will then gradually diversify by the accumulation of mutation(s) and/or recombinational replacements, resulting in slight variants of the founding genotype. Initially, members of this emerging clone will be indistinguishable in allelic profile by MLST, however with time, the clone will diversify to produce a number of variants in which one of the seven MLST loci has been altered—single locus variants (SLVs). Further diversification will produce variants of the founder ST that differ at two out of the seven loci—double locus variants (DLVs). In this simple model, bacterial populations will consist of a series of clonal complexes (sets of variants of a founding genotype) that can be recognized from the allelic profiles of the strains within a MLST database (10).

An interactive implementation of the BURST algorithm, eBURSTv2 (10), is integrated within the MLST websites at http://www.mlst.net as a JAVA™ applet and can be used to explore the relationships among strains within the database and to explore the relationships of newly characterized strains to those in the database (Figure 1G). eBURST uses the STs and their associated allelic profiles as input and, using the default setting, divides the strains into groups in which all STs in the same group share ⩾6 out of 7 loci with at least one other member of that group, resulting in non-overlapping groups or clonal complexes. Of particular value is the ability to link back to the MLST database from the eBURST diagram of a clonal complex, and the ability to display all the STs in a large MLST database in a single diagram [(10); a population snapshot; and Figure 3], showing all the major and minor clonal complexes, and individual STs that are relatively distantly related to all other STs.

Comparing query strains to the database: using the concatenated sequences

The ability to concatenate the sequences at the seven loci, maintaining the correct reading frame, and to construct a neighbor-joining tree based on these sequences is provided, but needs to be used with considerable caution. A module from MEGA (11) provides the tree topology in Newick format which is then displayed using the ATV applet (12). Allelic changes at the MLST loci will occur (to a varying degree

depending on the species) by recombination, and in many cases the relative contribution of recombination and point mutation to the diversification of strains will be unknown (13). A long history of recombination will preclude the recovery of the true phylogenetic relationships between distantly related bacterial strains and even the relatedness between similar strains may be better represented on a tree based on differences in allelic profiles than one based on differences in the concatenated sequences. However, there are specific issues that can be usefully addressed by using the concatenated sequences. For example, the *Burkholderia pseudomallei* database includes strains of closely related species and the *B.pseudomallei* MLST website provides a facility to examine the position of a query strain on the tree constructed using concatenated sequences, which can establish whether the query strain is *B.pseudomallei* or something similar to, but distinct from, *B.pseudomallei* (14). Similarly, there is considerable confusion about whether strains that appear to be *Streptococcus pneumoniae*, but which cannot be assigned to a pneumococcal capsular serotype, are authentic pneumococci that do not produce a capsule or are members of a similar but distinct streptococcal population. The pneumococcal MLST website has a facility to examine whether a query strain clusters within a reference set of *S.pneumoniae* strains, or with the related population, using a tree based on concatenated sequences, which can resolve this issue in most cases (see the following section; Figure 1H). Trees based on concatenated sequences may also be useful for assigning *Haemophilus influenzae* strains to major lineages (15) or for *Staphylococcus aureus* where recombination appears to be rare (16).

Typical workflow for data entry using the batch strain query

Here, we consider the workflow of a user analyzing a number of recently sequenced strains using batch entry. As an illustrative example we focus on a single representative mlst.net species website, http://spneumoniae.mlst.net, the site for characterizing strains of *S.pneumoniae* (17).

The uploaded XML file of a batch of *S.pneumoniae* strains and their associated sequences results in a table of results (Figure 1E). Error messages (red) alert the user to the fact that some sequences are of the wrong length for that locus (strain 8) or contain unexpected characters (strain 13). In some strains, all the alleles are previously known and the allele numbers are returned in the results table. For some of these strains, the combinations of alleles at the seven loci (allelic profiles) are also known and the ST number is shown in the table (e.g. strain 4). In one case (strain 14) the alleles are all known but the combination of alleles is previously unknown. In other strains, one or more alleles are unknown and the ST must also be unknown (e.g. strain 3), and the ST is flagged as incomplete, as the new alleles have to be checked and assigned new allele numbers by the curator. Clicking on 'unknown' allele highlights the nucleotide differences in the new allele compared with the most similar alleles (Figure 1C).

None of the alleles in strain 15 are found in the *S.pneumoniae* database, and there is therefore some uncertainty whether this strain is a pneumococcus. To investigate the status of this strain further, the user can select the option to examine the phylogenetic status of the strain, by using the concatenated sequences to compare its position on a reference tree (Figure 1H), which includes a set of strains covering the known diversity of authentic pneumococci, and a set of closely related strains that are similar to but distinct from the authentic pneumococci (W. P. Hanage and B. G. Spratt, unpublished data). The sequences of the loci of the query strain are concatenated, and the sequence is added to a stored file containing the concatenated sequences of the reference strains, and a neighbor-joining tree is constructed (Figure 1H). Using this approach, strain 15, which has an unknown allelic profile but known alleles at all loci, clusters within the authentic pneumococci, but strain 14 with new alleles at all loci is clearly not a pneumococcus, as it clusters away from the pneumococci and within the more diverse set of related streptococcal strains.

From the results of the batch strain query, the user can also relate their unknown STs to all other strains in the MLST database using eBURST (Figure 1G). The unknown STs are assigned unique temporary ST numbers, to distinguish them from the STs in the database. In Figure 1G, strain 14 has been assigned the temporary ST10001 and by eBURST it can be seen to be a SLV of ST156 within one of the major pneumococcal clonal complexes. Any strain in the batch strain query (excepting those with alleles of incorrect length or with unexpected characters) can be compared with the MLST database as, using the eBURST option, new alleles, as well as new STs, are given temporary numbers allowing them to be analyzed by the program.

CONCLUSIONS

Websites for evaluating the taxonomic status of strains using 16S rRNA sequences are well established and, in recent years, several websites have been developed for molecular epidemiology and population genetics, to assign isolates of bacterial species to strains, lineages and clonal complexes, using data generated by MLST. We describe the set of MLST species websites within http://www.mlst.net, and the tools that allow users to identify query strains, and to explore their relationship with other strains in the database. MLST is being widely used and there is a need for new ways to input and query large sets of strains and to display the relatedness of the many thousands of strains within the larger MLST websites. Some progress has been made to achieve these aims and in future we envisage a fully automated procedure, with data flowing directly from sequencer to ST assignment. Those developing new MLST schemes for bacterial or fungal species can join http://www.mlst.net to take advantage of the features available at this site, and to have a consistency of format for the MLST websites. A slightly different common format for MLST websites is provided by those species sites (such as that for *Neisseria meningitidis*) hosted at http://pubmlst.org. Hosting of new MLST schemes at http://www.mlst.net allows the databases to be stored and backed up on servers at Imperial College London but with remote strain entry, ownership and curation, by the developer of the MLST scheme, using MLST curation software.

ACKNOWLEDGEMENTS

We acknowledge the contribution of Man-Suen Chan who developed the initial MLST website software, James Abbott

for the STARS XML conversion script, and Sudhir Kumar and Joel Dudley for providing the server-side MEGA module for concatenated tree drawing. The development of http://www.mlst.net was supported by the Wellcome Trust. Part of the publication cost for this article was provided by JISC. Funding to pay the Open Access publication charges for this article was provided by the Wellcome Trust.

Conflict of interest statement. None declared.

REFERENCES

1. Maiden,M.C.J., Bygraves,J.A., Feil,E., Morelli,G., Russell,J.E., Urwin,R., Zhang,Q., Zhou,J., Zurth,K., Caugant,D.A. *et al*. (1998) Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc. Natl Acad. Sci. USA*, **95**, 3140–3145.
2. Hanage,W.P., Feil,E.J., Brueggemann,A.B. and Spratt,B.G. (2004) Multilocus sequence typing: strain characterization, population biology, and patterns of evolutionary descent. In Persing,D.H., Tenover,F.C., Versalovic,J., Tang,Y., Unger,E.R., Relman,D.A. and White,T.J. (eds), *Molecular Microbiology: Diagnostic Principles and Practice*. American Society Press, Washington DC, pp. 235–243.
3. Urwin,R. and Maiden,M.C. (2003) Multi-locus sequence typing: a tool for global epidemiology. *Trends Microbiol*., **10**, 479–487.
4. Bougnoux,M.E., Aanensen,D.M., Morand,S., Theraud,M., Spratt,B.G. and d'Enfert,C. (2004) Multilocus sequence typing of *Candida albicans*: data exchange and applications. *Infect. Genet. Evol*., **6**, 243–252.
5. Dodgson,A.R., Pujol,C., Denning,D.W., Soll,D.R. and Fox,A.J. (2003) Multilocus sequence typing of *Candida glabrata* reveals geographically enriched clades. *J. Clin. Microbiol*., **12**, 5709–5717.
6. Chan,M.S., Maiden,M.C.J. and Spratt,B.G. (2001) Database-driven multi locus sequence typing (MLST) of bacterial pathogens. *Bioinformatics*, **17**, 1077–1083.
7. Jolley,K.A., Chan,M.S. and Maiden,M.C.J. (2004) mlstdbNet—distributed multi-locus sequence typing (MLST) databases. *BMC Bioinformatics*, **5**, 86.
8. Huson,D.H. (1998) SplitsTree: analyzing and visualizing evolutionary data. *Bioinformatics*, **14**, 68–73.
9. Clamp,M., Cuff,J., Searle,S.M. and Barton,G.J. (2004) The Jalview Java alignment editor. *Bioinformatics*, **3**, 426–427.
10. Feil,E.J., Li,B.C., Aanensen,D.M., Hanage,W.P. and Spratt,B.G. (2004) eBURST: inferring patterns of evolutionary descent among clusters of related bacterial genotypes from multilocus sequence typing data. *J. Bacteriol*., **186**, 1518–1530.
11. Kumar,S., Tamura,K., Jakobsen,I.B. and Nei,M. (2001) MEGA2: Molecular Evolutionary Genetics Analysis software. *Bioinformatics*, **17**, 1244–1245.
12. Zmasek,C.M. and Eddy,S.R. (2001) ATV: display and manipulation of annotated phylogenetic trees. *Bioinformatics*, **17**, 383–384.
13. Feil,E.J., Holmes,E.C., Bessen,D.E., Chan,M.S., Day,N.P., Enright,M.C., Goldstein,R., Hood,D.W., Kalia,A., Moore,C.E. *et al*. (2001) Recombination within natural populations of pathogenic bacteria: short-term empirical estimates and long-term phylogenetic consequences. *Proc. Natl Acad. Sci. USA*, **98**, 182–187.
14. Godoy,D., Randle,G., Simpson,A.J., Aanensen,D.M., Pitt,T.L., Kinoshita,R. and Spratt,B.G. (2003) Multilocus sequence typing and evolutionary relationships among the causative agents of melioidosis and glanders, *Burkholderia pseudomallei* and *Burkholderia mallei*. *J. Clin. Microbiol*., **5**, 2068–2079.
15. Meats,E., Feil,E.J., Stringer,S., Cody,A.J., Goldstein,R., Kroll,J.S., Popovic,T. and Spratt,B.G. (2003) Characterization of encapsulated and non-capsulated *Haemophilus influenzae*, and determination of phylogenetic relationships, using multilocus sequence typing. *J. Clin. Microbiol*., **41**, 1623–1636.
16. Feil,E.J., Cooper,J.E., Grundmann,H., Robinson,D.A., Enright,M.C., Berendt,A., Peacock,S., Maynard Smith,J., Murphy,M., Spratt,B.G. *et al*. (2003) How clonal is *Staphylococcus aureus*? *J. Bacteriol*., **185**, 3307–3316.
17. Enright,M.C. and Spratt,B.G. (1998) A multilocus sequence typing scheme for *Streptococcus pneumoniae*: identification of clones associated with serious invasive disease. *Microbiology*, **144**, 349–360.

BMC Biology

BioMed Central

Research article

Open Access

Assigning strains to bacterial species via the internet

Cynthia J Bishop[†1], David M Aanensen[†1], Gregory E Jordan[2], Mogens Kilian[3], William P Hanage[1] and Brian G Spratt*[1]

Address: [1]Department of Infectious Disease Epidemiology, Imperial College London, St. Mary's Hospital Campus, London, UK, [2]European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK and [3]Department of Medical Microbiology and Immunology, Aarhus University, DK-8000 Aarhus, Denmark

Email: Cynthia J Bishop - c.bishop@imperial.ac.uk; David M Aanensen - d.aanensen@imperial.ac.uk; Gregory E Jordan - greg@ebi.ac.uk; Mogens Kilian - kilian@microbiology.au.dk; William P Hanage - w.hanage@imperial.ac.uk; Brian G Spratt* - b.spratt@imperial.ac.uk

* Corresponding author †Equal contributors

Published: 26 January 2009

BMC Biology 2009, **7**:3 doi:10.1186/1741-7007-7-3

Received: 1 December 2008
Accepted: 26 January 2009

This article is available from: http://www.biomedcentral.com/1741-7007/7/3

© 2009 Bishop et al; licensee BioMed Central Ltd.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Background: Methods for assigning strains to bacterial species are cumbersome and no longer fit for purpose. The concatenated sequences of multiple house-keeping genes have been shown to be able to define and circumscribe bacterial species as sequence clusters. The advantage of this approach (multilocus sequence analysis; MLSA) is that, for any group of related species, a strain database can be produced and combined with software that allows query strains to be assigned to species via the internet. As an exemplar of this approach, we have studied a group of species, the viridans streptococci, which are very difficult to assign to species using standard taxonomic procedures, and have developed a website that allows species assignment via the internet.

Results: Seven house-keeping gene sequences were obtained from 420 streptococcal strains to produce a viridans group database. The reference tree produced using the concatenated sequences identified sequence clusters which, by examining the position on the tree of the type strain of each viridans group species, could be equated with species clusters. MLSA also identified clusters that may correspond to new species, and previously described species whose status needs to be re-examined. A generic website and software for electronic taxonomy was developed. This site http://www.eMLSA.net allows the sequences of the seven gene fragments of a query strain to be entered and for the species assignment to be returned, according to its position within an assigned species cluster on the reference tree.

Conclusion: The MLSA approach resulted in the identification of well-resolved species clusters within this taxonomically challenging group and, using the software we have developed, allows unknown strains to be assigned to viridans species via the internet. Submission of new strains will provide a growing resource for the taxonomy of viridans group streptococci, allowing the recognition of potential new species and taxonomic anomalies. More generally, as the software at the MLSA website is generic, MLSA schemes and strain databases for other groups of related species can be hosted at this website, providing a portal for microbial electronic taxonomy.

Background

The methods used to assign strains to bacterial species, and the definition of species, have been much debated [1-4]. There is widespread recognition that the current methods of defining prokaryotic species are no longer adequate and there has been little progress in developing an agreed concept of species that can guide a new definition of species [5-7]. The current approach for distinguishing species within a genus, and for defining new species, is based on polyphasic taxonomy, which incorporates all available phenotypic and genotypic data into a consensus classification [8]. The accepted genotypic method for defining species is based on overall genomic relatedness, such that strains which share approximately 70% or more relatedness using DNA-DNA hybridization, with 5 °C or less ΔT_m (the difference in the melting temperature between homologous and heterologous hybrids), under standard conditions, are considered to be members of the same species [9]. The cut-off value implied by this definition may not be appropriate for all bacterial populations and in recent years a move away from this cumbersome procedure to other ways of identifying and circumscribing bacterial species has been encouraged [10].

One approach is to observe the distribution of a large number of strains of closely related species in sequence space and to identify clusters of strains that are well resolved from other clusters. This approach has been developed by using the concatenated sequences of multiple core (house-keeping) genes to assess clustering patterns, and has been called multilocus sequence analysis (MLSA) [2,11] or multilocus sequence phylogenetic analysis [12]. MLSA has been used successfully to explore clustering patterns among large numbers of strains assigned to very closely related species by current taxonomic methods [12-18], to look at the relationships between small numbers of strains within a genus [19], or within a broader taxonomic grouping [20], and to address specific taxonomic questions [21,22]. More generally, the method can be used to ask whether bacterial species exist – that is, to observe whether large populations of similar strains invariably fall into well-resolved clusters, or whether in some cases there is a genetic continuum in which clear separation into clusters is not observed.

Although there are limitations of the MLSA approach [11], it is attractive since it should allow the assignment of unknown strains to species clusters via the internet (electronic taxonomy), in a similar way that isolates have been assigned to strains (sequence types) or clonal complexes within a species using multilocus sequence typing (MLST) [23]. Electronic taxonomy requires the recognition and circumscription of species by the generation of a large initial database (and associated MLSA data) of multiple strains from all of the recognized species within a genus, or part of a genus. It also requires a website and software that allow the sequences from the house-keeping genes of a new query strain to be entered via the internet, and for the sequence cluster (if any) within which the query strain falls to be returned, along with a species assignment for the strain. The approach therefore allows species assignment via the internet and also allows the recognition of the appearance in the database of new clusters, which can be assigned as new candidate species.

We describe an exemplar for electronic taxonomy (and a generic website) using the viridans group streptococci, since the taxonomy of this group has been, and to some extent remains, problematic [24]. The non-pyogenic streptococci are subdivided into the Mitis, Anginosus, Salivarius, Mutans and Bovis groups, of which the first three are often referred to as viridans streptococci. The Mitis group currently includes the important pathogen *S. pneumoniae* and 12 other validly described species, *S. australis*, *S. cristatus* (formerly *S. crista*), *S. gordonii*, *S. infantis*, *S. mitis*, *S. oligofermentans*, *S. oralis*, *S. parasanguinis* (formerly *S. parasanguis*), *S. peroris*, *S. pseudopneumoniae*, *S. sanguinis* (formerly *S. sanguis*) and *S. sinensis*. The Anginosus group includes three recognized species, *S. anginosus*, *S. constellatus* (including two subspecies *subsp. constellatus* and *subsp. pharyngis*) and *S. intermedius*, and the Salivarius group includes *S. salivarius*, *S. vestibularis*, and *S. thermophilus*. All viridans streptococci (with the exception of *S. thermophilus*, which is not associated with humans) are commensals of the human upper respiratory tract and, excepting *S. pneumoniae*, rarely cause disease in immune-competent individuals. They have been strikingly resistant to satisfactory classification, reflected in frequently changing nomenclature and significant problems of identification by phenotypic analysis and by sequencing of 16S rRNA genes [16,25-33].

There is extensive evidence that recombination occurs between closely related Mitis group species [34-38], which can lead to fuzzy species [14], and may explain some of the taxonomic confusion within this group, and could prevent the identification of well-resolved sequence clusters. However, initial studies of the Mitis group using MLSA have shown that sequence clusters corresponding to *S. pneumoniae*, *S. pseudopneumoniae*, *S. mitis* and *S. oralis* can be resolved using trees constructed from the concatenated sequences of six of the seven housekeeping genes used in the *S. pneumoniae* MLST scheme [11,39]. Similarly, concatenation of the sequences of four house-keeping genes has been shown to resolve the more distantly related species within the viridans streptococci [16], although with four genes there was poor resolution between the strains of the closely related species, *S. pseudopneumoniae* and *S. mitis* [18].

In this study, the sequences of a set of seven house-keeping gene fragments were obtained from 420 strains, including strains of all species within the Mitis, Anginosus and Salivarius group streptococci, to develop a viridans group MLSA scheme and strain database. A generic open-access MLSA website http://www.eMLSA.net was developed that allows query strains to be assigned via the internet to known species clusters, to as yet unnamed sequence clusters, or as single divergent genotypes.

Results

Amplification and sequencing of gene fragments from viridans group streptococci

The primers for the amplification of internal fragments of eight house-keeping gene fragments were selected as in Methods (Table 1). The primers successfully amplified the eight gene fragments from 326 of the 375 (87%) Mitis group strains that were examined. They also amplified the correct fragments from some streptococcal species outside the Mitis group, including all strains of *S. anginosus*, *S. vestibularis* and *S. pyogenes* that were examined. However, the *guaA* gene could not be amplified from some strains (mostly *S. sanguinis* and *S. cristatus*) and this gene was eliminated to produce the final seven-locus MLSA scheme. The seven selected house-keeping genes could be amplified from all viridans group streptococci that were examined and were at least 40 kb apart on the *S. pneumoniae* R6 chromosome.

For each strain, the gene fragments were sequenced on both strands and trimmed to defined start and end positions. For each locus, the trimmed sequences from all strains were the same length, indicating that indels in these genes were absent, as they were in the corresponding *S. agalactiae*, *S. mutans* and *S. suis* sequences, and are thus likely to be uncommon among the viridans group.

Identification of sequence clusters

Figure 1 shows a neighbour-joining tree for all 420 concatenated sequences (402 from viridans group strains, 17 from strains of *S. pyogenes* and one from *S. agalactiae*) of the seven loci used in the new seven-locus MLSA scheme, with the strains coloured where a clear species assignment was provided by the laboratory of MK. The tree identified a number of sequence clusters that were well resolved, with ≥ 95% bootstrap confidence values for the nodes. The strains identified as *S. pseudopneumoniae* (including the type strain) clustered together but were not well resolved from the *S. mitis* cluster (bootstrap value of 92% for the node separating the *S. pseudopneumoniae* and *S. mitis* clusters on the neighbour-joining tree, but only 46% on a minimum evolution tree; see below). The *S. mitis* cluster was unusual as it consisted of a set of closely related sub-clusters arising from the branch that separates the *S. pneumoniae* and *S. pseudopneumoniae* clusters from that of *S. oralis*.

Concordance between these clusters and the assigned species names was imperfect, in accord with the difficulties in assigning viridans group strains to species. However, almost all of the clusters contained only one designated type strain (Figure 1). Accordingly, the majority of the sequence clusters may be equated with recognized species, that is, *S. anginosus*, *S. australis*, *S. constellatus*, *S. cristatus*,

Table 1: Genes and PCR primers used in viridans group MLSA scheme

Primer	Gene product	Sequence (5'-3')*	Primer length (bp)	Trimmed fragment size (bp)	Annealing temperature (°C)
guaA-up	GMP synthase	ATYCARTTYCACCCMGAAGT	20	567	55
guaA-dn		CWGGNCCWGGRAATGGTTG	19		55
map-up	Methionine aminopeptidase	GCWGACTCWTGTTGGGCWTATGC	23	348	55
map-dn		TTARTAAGTTCYTTCTTCDCCTTG	24		55
pfl-up	Pyruvate formate lyase	AACGTTGCTTACTCTAAACAAACTGG	26	351	55
pfl-dn		ACTTCRTGGAAGACACGTTGWGTC	24		55
ppaC-up	Inorganic pyrophosphatase	GACCAYAATGAATTYCARCAATC	23	552	50
ppaC-dn		TGAGGNACMACTTGTTTSTTACG	23		50
pyk-up	Pyruvate kinase	GCGGTWGAAWTCCGTGGTG	19	492	50
pyk-dn		GCAAGWGCTGGGAAAGGAAT	20		50
rpoB-up	RNA polymerase beta subunit	AARYTIGGMCCTGAAGAAAT	20	516	50
rpoB-dn		TGIARTTTRTCATCAACCATGTG	23		50
sodA-up	Superoxide dismutase	TRCAYCATGAYAARCACCAT	20	378	50
sodA-dn		ARRTARTAMGCRTGYTCCCARACRTC	26		50
tuf-up	Elongation factor Tu	GTTGAAATGGAAATCCGTGACC	22	426	55
tuf-dn		GTTGAAGAATGGAGTGTGACG	21		55

*Sequences are shown using the IUPAC codes for sites where degeneracy was introduced.

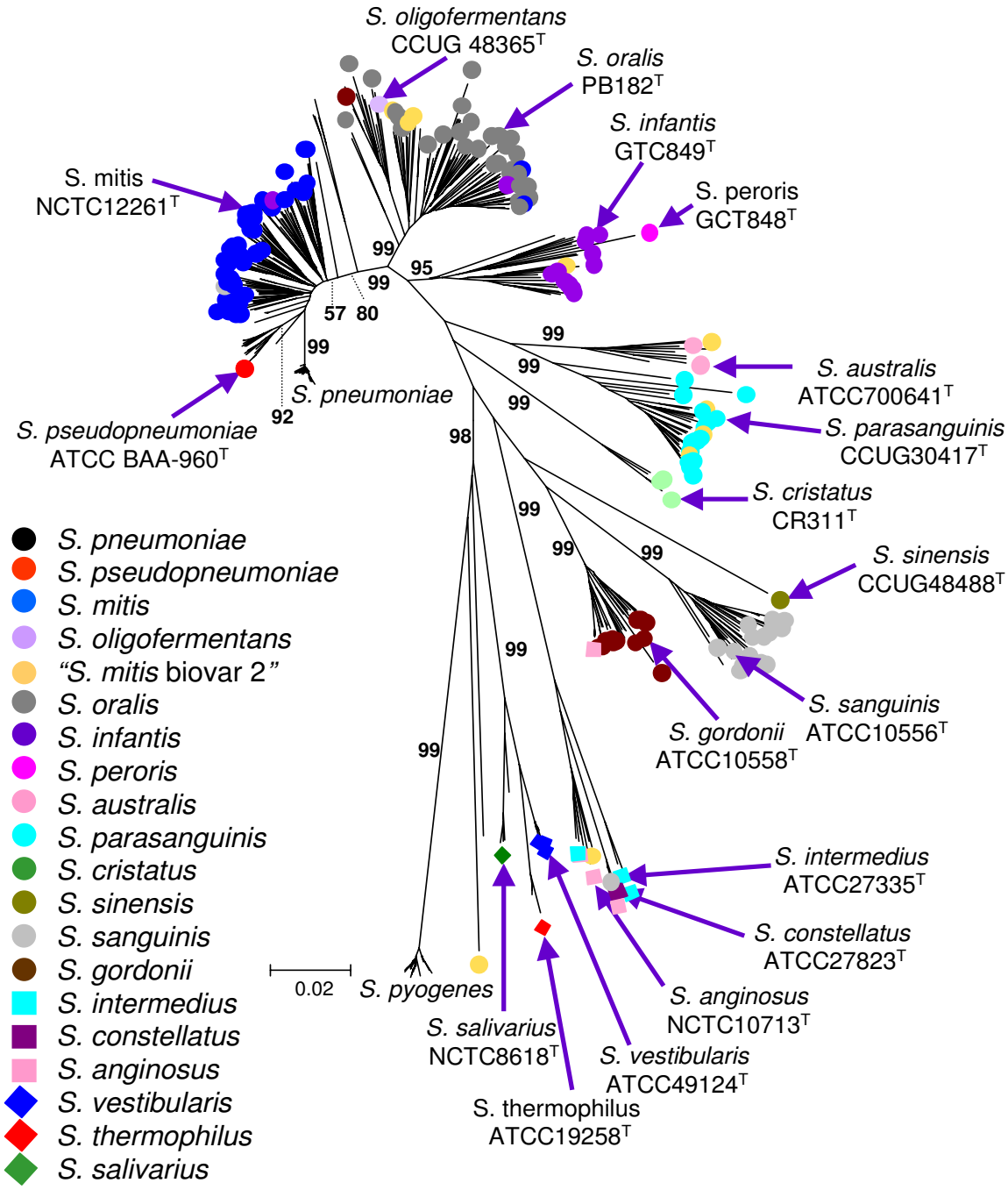


Figure 1

Tree showing the positions of well-characterized strains and type strains within sequence clusters. A neighbour-joining radial tree was constructed using the concatenated sequences of all 420 strains. The viridans group strains that are coloured were those assigned to species in the laboratory of MK; the positions on the tree of the type strains of viridans group species are indicated. Mitis, Anginosus and Salivarius group species are shown, respectively, as coloured circles, squares and diamonds. Bootstrap support values (%) for each of the nodes leading to the Mitis group sequence clusters are indicated; values for the clusters within the Anginosus and Salivarius groups are shown in Figure 2C. The colour key is ordered top to bottom according to the position of the clusters on the radial tree, from *S. pneumoniae* to *S. pyogenes*.

S. gordonii, *S. infantis*, *S. intermedius*, *S. mitis*, *S. oralis*, *S. parasanguinis*, *S. pseudopneumoniae*, *S. sanguinis*, *S. salivarius*, *S. thermophilus* and *S. vestibularis*. The only exceptions were the *S. peroris* type strain which, although on an unusually long branch, fell within a cluster that otherwise included the *S. infantis* type strain and several reference strains of that species, and *S. oligofermentans* that was within a subgroup of *S. oralis* (see below).

For several of the sequence clusters there was complete, or almost complete, concordance with species names assigned on the basis of phenotypic examination. However, for some species (for example, *S. mitis*, *S. infantis* and *S. oralis*) that have been affected by changing taxonomy, or difficulties in phenotypic differentiation, there was significant discordance between the clusters and previously assigned species names. The major anomaly was that strains historically assigned as *S. mitis* fell into a number of different closely related sequence clusters, particularly those otherwise consisting of *S. infantis* and *S. oralis*. This anomaly was due largely to strains initially assigned as '*S. mitis* biovar 2' being either part of the *S. oralis* cluster (see below), in agreement with a recent observation [18], or part of the subsequently described *S. parasanguinis* cluster [26]. Besides the anomalous clustering of some '*S. mitis*' strains, there were single strains assigned as *S. gordonii* and *S. infantis* within the *S. oralis* cluster, a single strain of '*S. mitis* biovar 2' within the *S. anginosus* cluster and a single strain of *S. anginosus* within the *S. constellatus* cluster.

The strains within the major sequence clusters correlated sufficiently well with the species names obtained through standard taxonomic procedures, and included the type strain of the species, that each of the sequence clusters was considered to represent a species cluster. Figure 2A shows a neighbour-joining tree for all 420 strains with the strains within each cluster coloured according to their assigned species names. Figure 2C shows the same tree with the Mitis group clusters collapsed to show better the resolution of the species clusters within the Anginosus and Salivarius groups. A few strains that were on the fringes of sequence clusters were not coloured, to represent uncertainty as to the species to which they belong. The tree shown in Figure 2A was used as the MLSA reference tree for assigning a query strain to a viridans group species from its location within one of the defined species clusters.

Comparing clustering patterns using different MLSA schemes and different tree-building methods

If MLSA is to be a reliable approach to identifying species as sequence clusters, the same clusters should be obtained for the same set of strains with different sets of house-keeping genes. Figure 3 shows a comparison of the trees produced from a set of 93 strains of *S. pseudopneumoniae*, *S. pneumoniae*, *S. mitis* and *S. oralis* that were characterized by both the new seven-locus MLSA scheme and the earlier scheme that used six of the pneumococcal MLST genes [11,15]. Although these schemes used completely different house-keeping genes, the trees were remarkably similar. The genetic distances between clusters were, however, less with the new MLSA scheme than with the earlier scheme (Figure 3). There were two minor differences between the trees. Firstly, two strains on the fringes of the *S. pseudopneumoniae* cluster using the MLST loci were allied to the *S. mitis* cluster using the seven-locus MLSA scheme and were re-assigned as uncertain species. Secondly, one strain (768-1) was an outlier of the *S. pneumoniae* sequence cluster in the six-locus tree but not in the seven-locus tree; however, its *guaA* gene was unusually divergent and it was also an outlier when this gene was added to the seven-locus MLSA scheme (Figure 4B).

The clustering patterns were not greatly influenced by the tree-building algorithm. Figure 2B shows the tree produced by minimum evolution (a method based on local optimization of an initial neighbour-joining tree), compared with that obtained using neighbour-joining (comparisons were restricted to these two methods since both are computationally efficient, as required for the rapid generation of trees on-line with sequences from large number of strains). The assignment of strains to the sequence clusters was the same, and the only significant difference was a slightly more marked sub-division of the *S. mitis* group into sub-clusters using minimum evolution. As expected for sequences that are closely related, using a genetic distance correction made no difference to the clustering patterns (data not shown).

Effect of adding an eighth gene to the MLSA scheme

To assess whether the addition of an eighth gene improved the resolution of sequence clusters (particularly of *S. mitis* and *S. pseudopneumoniae*), the 326 Mitis group strains in which the *guaA* gene could be amplified were used to produce an eight-locus tree, that could be compared with the seven-locus tree produced for all Mitis group strains (Figure 4). There were only minor differences between the eight and seven-locus trees and, excepting the location of the *S. peroris* type strain as either within, or an outlier of, the *S. infantis* cluster, none of the differences affected the resolution of the sequence clusters, and the seven-locus scheme was chosen for all further work.

Identification of potential new viridans group species

The MLSA scheme should be able to identify possible new species as groups of strains, or individual strains, which clearly do not fall within any of the known species clusters. We firstly checked whether strains that did not fall into known species clusters, or were outliers of species clusters, could be the result of using DNA from mixed cultures. In all cases, further sub-culturing of these strains (labelled in Figure 2A) and re-sequencing of the seven loci gave the same

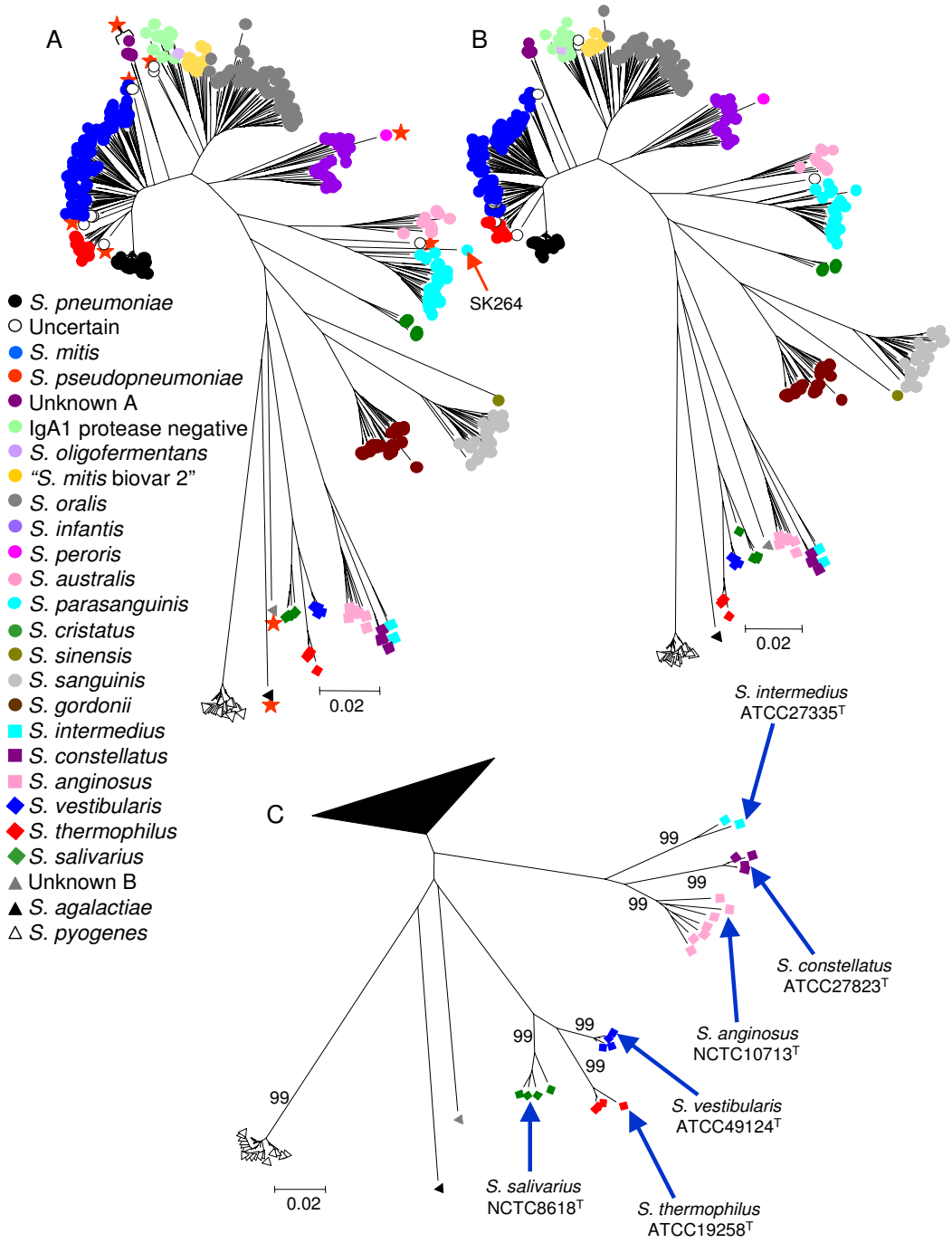


Figure 2
The MLSA reference tree. A. The neighbour-joining tree of Figure 1 was relabelled so that all strains within a sequence cluster were assigned to the species inferred from the positions on the tree of the type strain and other well-characterized strains. The limits of the *S. mitis* and *S. pseudopneumoniae* clusters are somewhat unclear and the strains on the flanks of this cluster are shown as white circles (uncertain species) to indicate this. Strains that were re-checked as they were outliers of species clusters, or were distinct from all other strains, are indicated by a red asterisk. B. The tree constructed from the same set of concatenated sequences using minimum evolution. C. The neighbour-joining tree with the Mitis group clusters collapsed, to show more clearly the sequence clusters within the Anginosus and Salivarius groups. The positions of the type strains and bootstrap values for the nodes are shown.

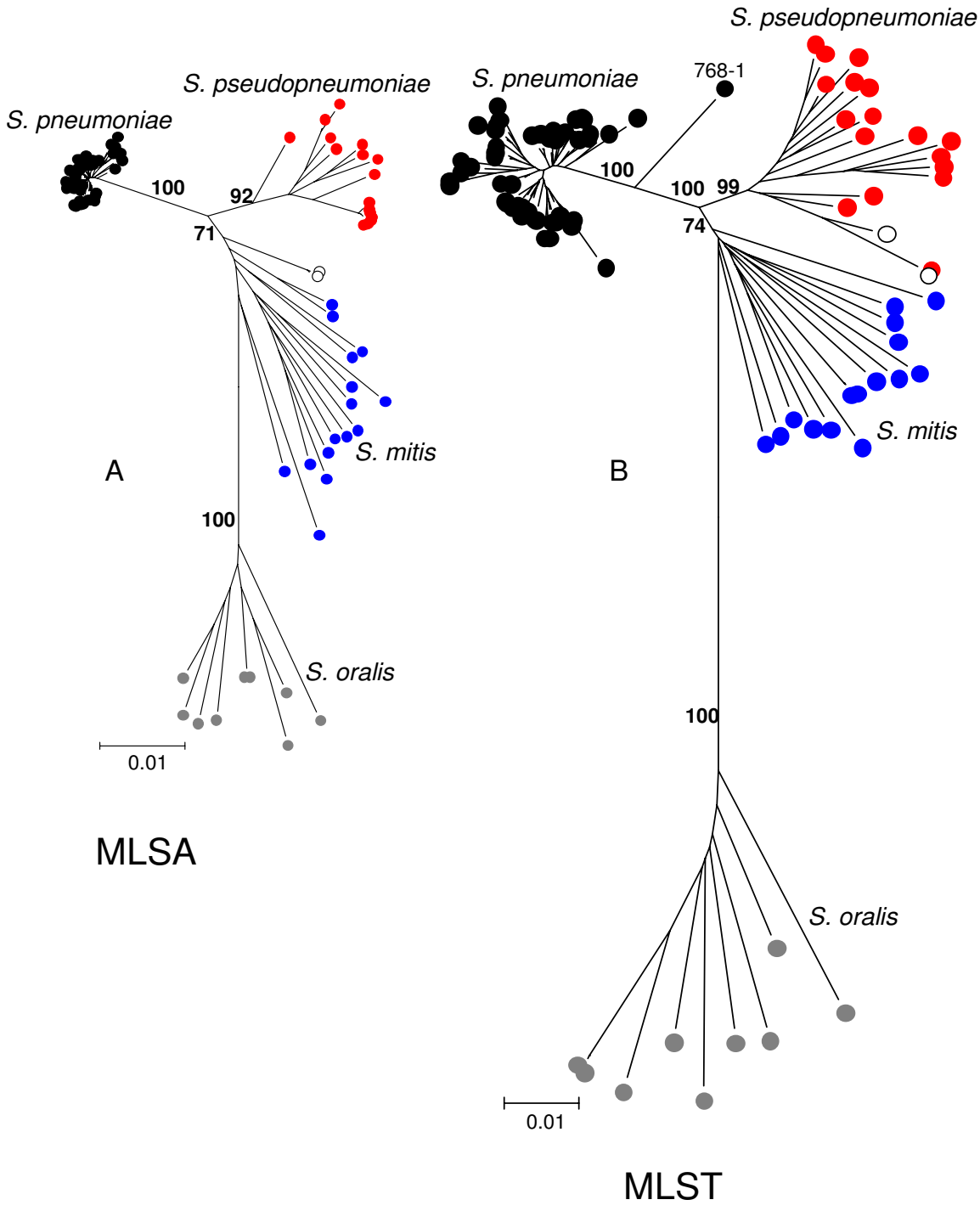


Figure 3

Comparison of the clustering of strains using two different MLSA schemes. The neighbour-joining tree obtained for a set of 93 strains of *S. pneumoniae*, *S. pseudopneumoniae*, *S. mitis* and *S. oralis* using the seven-locus MLSA scheme (A) was compared with the tree obtained from the same strains using the concatenated sequences of six of the loci used in the *S. pneumoniae* MLST scheme (B). The two trees are drawn to the same scale.

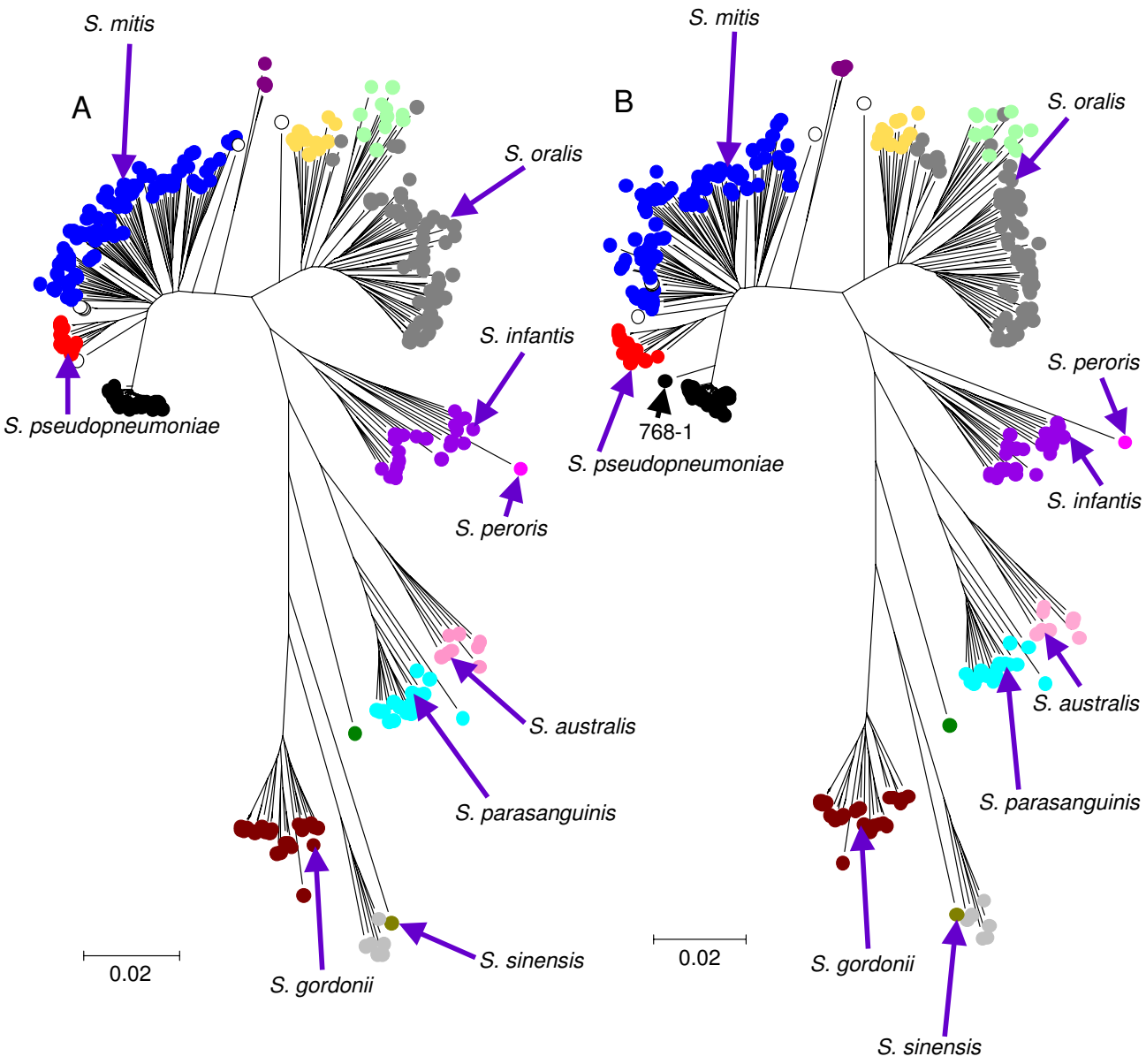


Figure 4

Effect on clustering patterns of adding an additional locus (*guaA*) to the seven-locus scheme. The *guaA* gene was successfully sequenced from 326 of the Mitis group strains and the neighbour-joining tree obtained for these strains using the seven-locus MLSA scheme (A) was compared with that obtained from the same strains by adding the *guaA* sequence to the concatenated sequences of the seven loci (B). The colour code for species clusters is as in Figure 2. The positions on the tree of the type strains are indicated.

results as obtained initially, ruling out the possibility that the position on the concatenated tree of these strains was due to the use of DNA from mixed cultures.

One example of an unidentified cluster is provided by four closely related strains (assigned by MK as *S. mitis*) that arose as a distinct lineage from the branch separating the *S. mitis* and *S. oralis* clusters (unknown A in Figure 2). These four strains were isolated from members of two families included in a study of the clonal diversity of *S. mitis* within individuals [40]. Although the strains cannot be distinguished from *S. mitis* by phenotypic analysis, they may constitute a separate species.

Likewise, the two major clusters that include the type strains of *S. mitis* and *S. oralis* both include several sub-clus-

ters, which may warrant recognition as separate species. As an example, 10 strains previously assigned to '*S. mitis* biovar 2' [25], formed a sub-cluster (including strains SK34, SK79, SK96 and others) within the major *S. oralis* cluster (Figure 5), in agreement with recent observations [18]. These strains are phenotypically distinct from the remaining part of the *S. oralis* cluster by being arginine hydrolysis and α-maltosidase positive, and by expressing the Lancefield group K cell wall carbohydrate antigen (data not shown). In addition, all strains of the '*S. mitis* biovar 2' subcluster lacked IgA1 protease activity (Figure 5).

Another sub-cluster within the major *S. oralis* cluster was also composed exclusively of strains that lacked IgA1 protease activity, while only six of the 59 *S. oralis* strains that were not in the above two sub-clusters were IgA1 protease negative (Figure 5). Surprisingly, the latter IgA1 protease-negative sub-cluster included the designated type strain of *S. oligofermentans*, in spite of its significantly different 16S rRNA sequences [41].

Apart from these examples, none of the other sub-clusters identified within the *S. mitis* and *S. oralis* clusters had distinguishing phenotypic properties, and their potential recognition as separate taxa should await further expansion of the eMLSA database and comprehensive phenotypic characterization of strains. A single strain (unknown B) was located on a long branch on the reference tree between *S. agalactiae* and the species clusters of the Salivarius group (Figure 2C) and its taxonomic status is unclear. However, from BLAST comparisons of the housekeeping genes of this strain to the nucleotide sequence databases this strain is almost certainly a member of the Bovis group streptococci.

Differences in diversity of species clusters

There were substantial differences in the extent of diversity within the species clusters and in their patterns of branching. For example, *S. pneumoniae* strains formed a tight sequence cluster on the tree and the average within-species diversity was only 0.3%, whereas the average diversity within several other clusters, including those of *S. anginosus*, *S. australis*, *S. infantis*, *S. mitis* and *S. oralis*, was 12 to 16 times greater (Table 2). Additionally, unlike *S. pneumoniae* and *S. pyogenes*, where there is a branching structure within the species clusters, with some strains being almost identical and others being more distantly related, the other species clusters typically included strains that were relatively distantly related to each other (for example, the *S. infantis* and *S. mitis* clusters), as implied by the long branch lengths from each strain in a cluster to its node (Figure 2A).

Clustering patterns obtained using single gene trees and identification of interspecies imports

Neighbour-joining trees were constructed from the sequences of the seven individual genes from all strains. These gene trees failed to resolve the species clusters obtained using the concatenated sequences and all individual trees showed anomalous clustering of some strains (Figure 6; Additional file 1). For example, *S. mitis* and *S. pseudopneumoniae* could only be resolved on the *map* and *pyk* gene trees (Figure 6B). Similarly, strains of *S. infantis* and *S. australis* could not be resolved on the *sodA* gene tree, and strains of neither of these species could be resolved from each other, or from *S. parasanguinis*, on the *ppaC* tree (Figure 6A).

Approximately 96% of strains assigned to a particular species by MLSA had alleles at all seven loci that were most similar to those of other strains of that species and, in individual gene trees, the sequences clustered with those of other strains of the species. However, 16 of the viridans group strains (4%) had alleles that were considered to be foreign (see Methods for criteria used to assign alleles as foreign). Table 3 shows the putative origins for each of the seven alleles in these 16 strains.

Assigning strains to species via the internet

The main features of the eMLSA.net website are described in the Methods section. Figure 7 is a screenshot from the species assignment window of http://viridans.eMLSA.net, showing the position of a query strain (*S. parasanguinis* SK264) on the reference tree, and the list of species assignments of the five strains in the database with the most similar concatenated sequences, and their sequence similarity to the query sequence. In this example the query strain is assigned as *S. parasanguinis* since the concatenated sequence of the seven MLSA loci clusters within this species cluster, and the top five most similar concatenated sequences are from strains assigned to this species.

Figure 8 shows a screenshot of the resident, resident compatible or foreign status assigned by eMLSA.net to each of the seven individual alleles of the above query strain. In this case, five of the individual loci are most similar to sequences from *S. parasanguinis* (resident alleles) or are resident compatible (for example, *ppaC*; Figure 6A). However, the *pyk* gene is within a cluster that is well resolved from the main *S. parasanguinis* cluster and which includes *S. infantis*, *S. australis* and *S. cristatus* (Figure 6B), and the *tuf* gene clusters within the well-resolved *S. australis* cluster (Figure 6C). These two alleles are therefore assigned by eMLSA.net as foreign, which may explain why this strain is an outlier within the *S. parasanguinis* cluster (Figure 2A).

Discussion

The aim of this paper was to produce a generic electronic taxonomy website and software that can be used for the assignment of bacterial strains to species via the internet and to develop a comprehensive dataset that can form a framework for establishing a more satisfactory approach to prokaryotic taxonomy. The viridans group streptococci

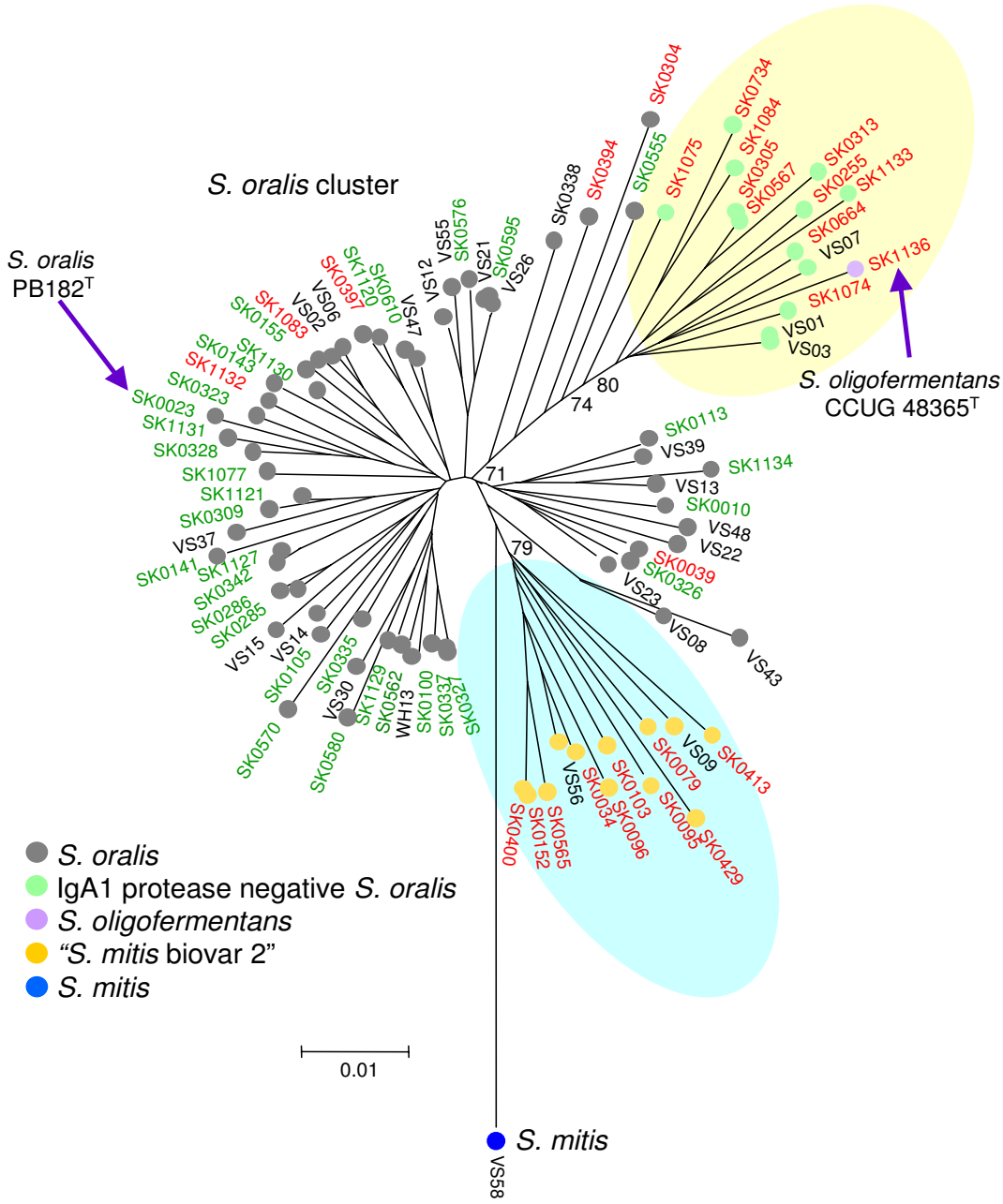


Figure 5
Phenotypically distinct sub-clusters within the *S. oralis* species cluster. The region of the neighbour-joining tree that includes strains within the *S. oralis* species cluster is shown in more detail. The positions on the tree of the type strains are indicated. The sub-cluster of phenotypically distinct strains that are arginine hydrolysis and α-maltosidase positive, and which express the Lancefield group K cell wall carbohydrate antigen, is indicated (highlighted in blue). A further subgroup, which included the *S. oligofermentans* type strain (SK1136), consisting exclusively of IgA protease-negative strains is also indicated (highlighted in yellow). Bootstrap values for relevant nodes are shown. Green strain names indicate IgA1 protease-positive strains whereas in red they are IgA1 protease negative. The IgA1 protease status of a few strains (black strain names) was unknown.

Table 2: Average nucleotide sequence diversity within species clusters

Species cluster	Average diversity
S. pneumoniae	0.003
S. vestibularis	0.006
S. thermophilus	0.009
S. constellatus	0.009
S. pseudopneumoniae	0.014
S. pyogenes	0.018
S. salivarius	0.020
S. cristatus	0.022
S. intermedius	0.027
S. gordonii	0.032
S. sanguinis	0.033
S. parasanguinis	0.034
S. anginosus	0.036
S. oralis	0.036
S. oralis – IgA1 protease negative	0.036
S. mitis	0.038
"*S. mitis* biovar 2"	0.039
Unknown A	0.043
S. australis	0.046
S. infantis	0.050

Species are listed in ascending order of intra-species sequence diversity.

were chosen as an exemplar of electronic taxonomy using the MLSA approach since their taxonomy by traditional means has been problematic, with poorly defined species, frequent revisions and changes of species names [24].

The seven-locus MLSA scheme resolved sequence clusters that were broadly consistent with those expected from the species assignments provided by standard taxonomic methods. Importantly, except for *S. peroris* and *S. oligofermentans*, the sequence clusters included the type strain of only a single species which, in a formal taxonomic sense, defines these as species clusters. The location of the *S. peroris* type strain on a long branch within the *S. infantis* cluster is consistent with the close relationship between these species in the tree based on four loci described by Hoshino et al [16], and with % DNA-DNA hybridization values as high as 62% between some strains of *S. infantis* and *S. peroris* [30]. The lack of any other strains assigned to the latter species precluded any further exploration of the validity of the distinction between these two very similar species, but the species status of *S. peroris* needs to be re-assessed.

The type strain of *S. oligofermentans* [41] was within the *S. oralis* species cluster. Similarly, six of the seven individual loci clustered with *S. oralis* strains; the other locus (*sodA*) clustered with *S. mitis*. Sequences of three of the MLSA genes (*sodA*; accession number DQ232554, *tuf*; EU156926 and *rpoB*; DQ232508) from the same strain of *S. oligofermentans* have been deposited in the nucleotide sequence databases. These genes have sequences identical to those we obtained, eliminating the possibility of any errors in our data on this strain, and the species status of *S. oligofermentans* also needs to be re-examined.

Otherwise, the major anomaly was the apparent difficulty that taxonomists have in recognizing strains of *S. mitis*; strains assigned on standard taxonomic grounds to this species were found among several clusters, although the major cluster included the *S. mitis* type strain. This problem is largely resolved by the recognition that '*S. mitis* biovar 2' strains are a sub-cluster within the major *S. oralis* cluster, as also suggested by Kilian et al [18].

MLSA schemes using different sets of house-keeping genes should give the same clusters and this requirement was fulfilled, although the resolution between *S. pseudopneumoniae* and *S. mitis* was less clear using the concatenated sequences of the seven genes than with the six MLST genes. The poorer resolution may be a consequence of the method used to ensure that the gene fragments could be amplified from all Mitis, Anginosus and Salivarius group species (see Methods), which may favour genes that evolve more slowly than the average house-keeping gene. This was apparently the case as the genetic distances between clusters was clearly less with the new MLSA scheme compared with the earlier scheme (Figure 3). However, the new scheme is favoured as it allows a much wider group of species to be studied, allowing strains to be assigned via the internet to all Mitis, Anginosus and Salivarius group species (and to some other non-viridans streptococci including *S. agalactiae* and *S. pyogenes*) using the initial database of 420 strains. The concatenated sequences of some of these species differed at over 20% of nucleotides and it is likely that the degenerate primers will amplify the seven loci from other closely related streptococcal species, allowing the MLSA scheme to assign strains to these additional species.

As new strains are added to the database, electronic taxonomy allows the identification of new sequence clusters that can be assigned as new species. Additionally, the mapping of phenotypic characteristics onto sequence clusters provides an important way of exploring the consistency of taxonomic assignments. In some cases this may produce evidence supporting the splitting of a cluster into more than one species. This was evident among the *S. oralis* species cluster where one sub-cluster, corresponding to '*S. mitis* biovar 2', had clearly distinctive biochemical properties, which may be an argument for recognizing it as a separate species.

The individual gene trees failed to identify the sequence clusters that were resolved with the concatenated sequences, a phenomenon observed previously in an

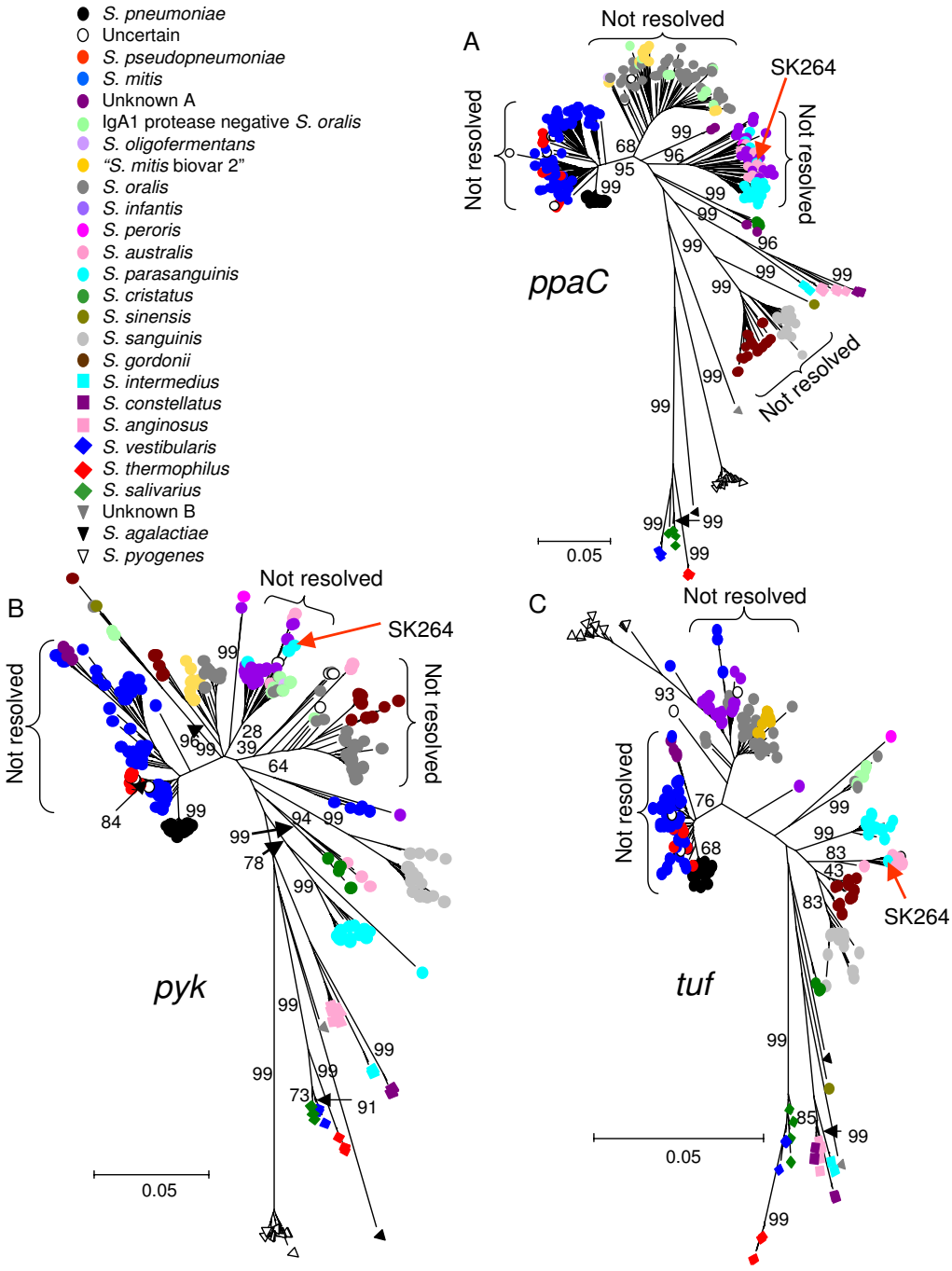


Figure 6

Examples of individual gene trees produced from the sequences of all 420 strains. A-C shows the *ppaC, pyk and tuf* gene trees, which each fail to resolve some of the species clusters. Strain SK264 (assigned as *S. parasanguinis*; labelled in Figure 2A) was used as the query strain in Figures 7 and 8. The *ppaC* allele of this strain is assigned as resident compatible as it falls within an unresolved cluster that includes the other *S. parasanguinis* sequences. For both *pyk* and *tuf*, the sequences from SK264 fall within clusters that are well resolved (bootstrap values of ≥ 80%) from the cluster that includes all (or the great majority) of the other *S. parasanguinis* sequences and are assigned as foreign alleles. For *pyk* the source of the foreign allele is unclear as its sequence falls within a cluster that includes those from several species. For *tuf*, the sequence appears to have been introduced from an *S. australis* strain. The major unresolved clusters in the trees are indicated.

Table 3: Strains possessing alleles from other species

Species	Strain	*map*	*pfl*	*ppaC*	*pyk*	*rpoB*	*sodA*	*tuf*
Unknown A	SK638	resident*	resident	resident	resident	resident	*S. mitis*	resident
S. gordonii	SK121	resident	resident	resident	uncertain**	resident	resident	resident
IgA1 protease negative *S. oralis*	SK313	resident	resident	resident	uncertain	resident	resident	resident
IgA1 protease negative *S. oralis*	SK734	resident	resident	resident	resident	resident	*S. australis*/*S. infantis*	resident
IgA1 protease negative *S. oralis*	SK1084	resident	resident	resident	resident	resident	*S. australis*/*S. infantis*	resident
IgA1 protease negative *S. oralis*	SK1133	resident	resident	resident	uncertain	resident	resident	resident
S. oralis	SK304	resident	resident	resident	uncertain	resident	resident	IgA1 protease negative *S. oralis*
S. oralis	SK309	resident	resident	resident	resident	resident	*S. australis*/*S. infantis*	resident
S. oralis	SK394	resident	resident	resident	resident	resident	resident	IgA1 protease negative *S. oralis*
S. oralis	SK555	resident	resident	resident	resident	resident	resident	IgA1 protease negative *S. oralis*
S. oralis	VS37	resident	resident	resident	resident	resident	*S. australis*/*S. infantis*	resident
S. parasanguinis	SK236	resident	resident	resident	*S. infantis*/*S. oralis*/ IgA1 protease negative *S. oralis*	resident	resident	resident
S. parasanguinis	SK264	resident	resident	resident	*S. infantis*/*S. oralis*/ IgA1 protease negative *S. oralis*	resident	resident	*S. australis*
S. parasanguinis	SK438	resident	resident	resident	*S. infantis*/*S. oralis*/ IgA1 protease negative *S. oralis*	resident	resident	resident
S. pseudopneumoniae	IOPR1711	resident	resident	resident	resident	*S. pneumoniae*	resident	resident
S. thermophilus	19258	resident	uncertain	resident	resident	resident	resident	resident

Foreign alleles are underlined. *Resident or resident compatible allele. **Foreign allele but donor species uncertain.

MLSA study of *S. pneumoniae*, *S. pseudopneumoniae*, *S. mitis* and *S. oralis* [11,15]. The most likely reasons for the differences in the individual gene trees are either that there were insufficient informative sites to resolve some species using a single gene fragment, or that a history of interspecies recombination has distorted the relationships between species clusters. There was, however, sufficient signal in the individual gene trees to resolve the species clusters using the concatenated sequences of all seven genes.

Only 16 strains clearly had foreign alleles, which presumably had been introduced by interspecies recombination from another viridans group species. However, the approach used would not detect recombination between very similar species that are not resolved on individual gene trees (which are likely to be the most common events; see below) and, together with the conservative high bootstrap value (≥ 80%) used to define clusters as well resolved, this may have lead to a considerable underestimate of the extent of inter-species recombination within this group. The efficiency of recombination falls off as the logarithm of the sequence divergence between the donor and recipient strains [42] and, as expected from this relationship, where the source of alleles could be clearly assigned, most foreign alleles were from donor species that clustered on the concatenated tree close to the recipient species. For example, the source of foreign alleles in *S. pseudopneumoniae* was *S. pneumoniae*, and for *S. oralis* it was *S. australis*/*S. infantis* andIgA1 protease-negative S. *oralis*.

Interspecies recombination could lead to a strain clustering on the tree away from other strains of the species if the recombinational replacement is from an unusually divergent donor, or there is more than one locus that has received a recombinational replacement. This phenomenon has been found with *Salmonella enterica* var Typhi strains, where six of the genes used in MLSA clustered within those of other strains of *S. enterica*, whereas the *purE* sequence was about 6% divergent such that, using the concatenated sequences, the Typhi strains clustered slightly apart from the *S. enterica* cluster [43]. This potential problem can be recognized using the MLSA website by examination of the individual gene trees and the resident or foreign status of the alleles (and the percentage diver-

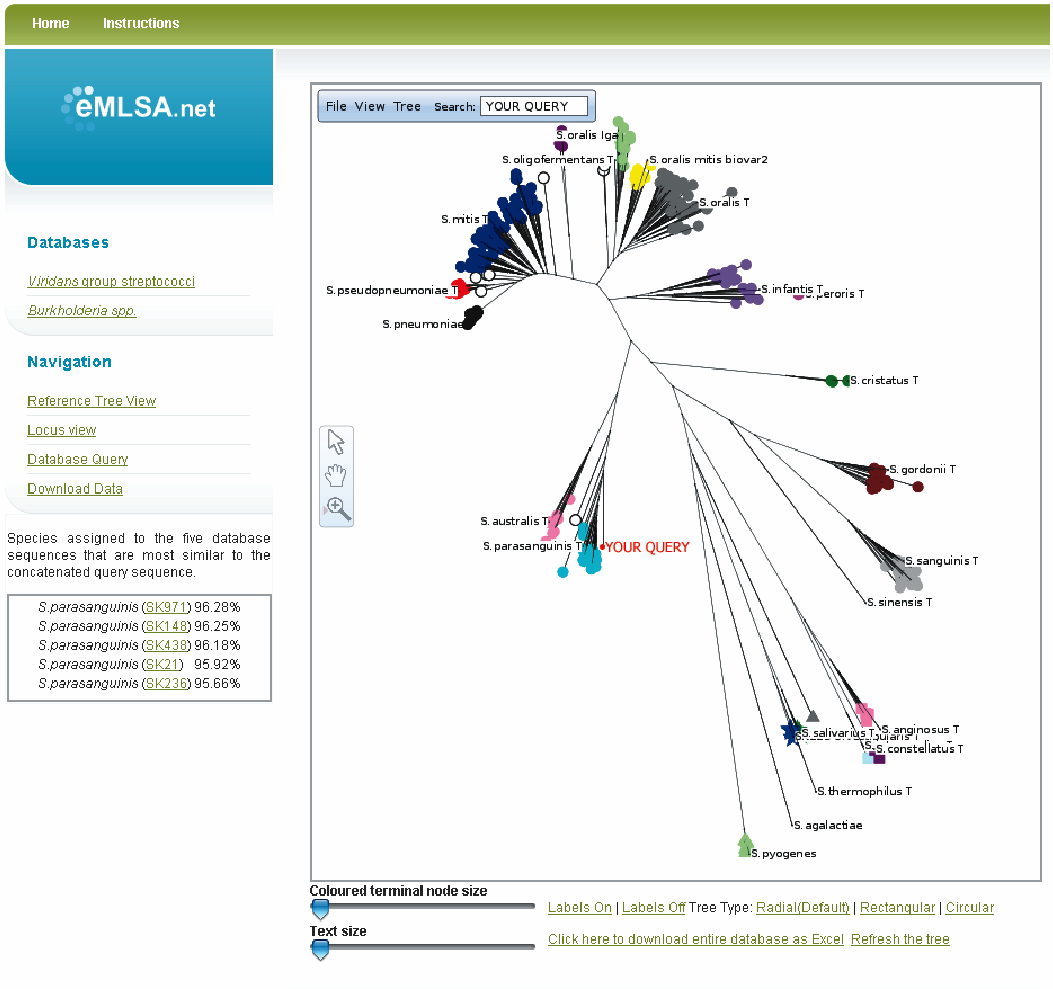


Figure 7
Species assignment on the internet. The eMLSA.net page returned after entering the seven gene sequences of a query strain and requesting a species assignment. The species assigned to the five most closely matching concatenated sequences are returned (left) along with an unrooted neighbour-joining radial tree indicating the position of the query strain. In this case, the query strain (SK264) is assigned as *S. parasanguinis* as the five most similar concatenated sequences are all from this species and the query strain falls within the *S. parasanguinis* species cluster on the tree.

gence) in strains that are outliers of known species clusters.

How do the results described here inform the way bacterial species should be defined? Firstly, the patterns of clustering observed in the viridans group have relevance to any method that attempts to define bacterial species using fixed cut-off values, whether these are based on percent DNA-DNA hybridization [9] or average percent nucleotide differences between all shared genes [44]. The very different levels of diversity among the different species clusters (Table 2) highlight the difficulties in defining bacterial species using a fixed cut-off value. The advantage of electronic taxonomy is that species clusters can be identified and circumscribed by visual inspection of the tree, without regard to preconceptions about the extent of diversity a bacterial species should have. The diversity of genotypes within some species of this group, and recombination among similar species, could make it unlikely that there will be a single phenotypic characteristic, common to all strains of a species but not to those of other similar species, which can be used to recognize each viridans group species. The requirement for a defining phenotypic characteristic for the official recognition of new species can in any case be questioned as outmoded and unnecessary if these can be identified simply by their clustering on a tree using MLSA [11]. Finally, as shown by the unexpected clustering of the *S. oligofermentans* and *S. peroris* type strains, the results highlight how MLSA can evaluate the taxonomic status and validity of proposed or accepted new species and can identify taxonomic anomalies.

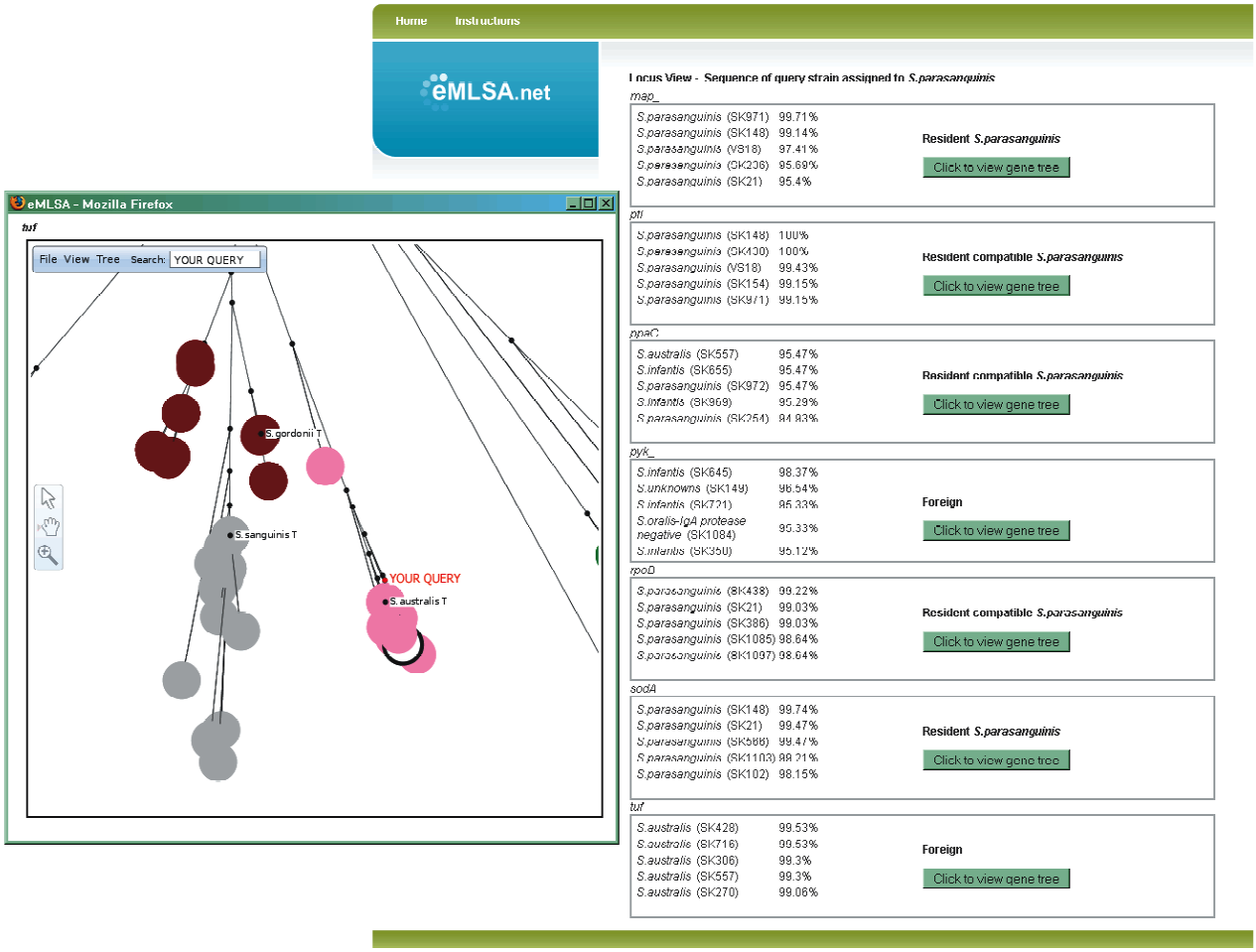


Figure 8

Online assignment of alleles as resident or foreign. Having assigned a query strain to a species (Figure 7), the locus view page shows the assignment of each of the seven individual sequences as resident to that species (or compatible with being resident) or foreign. In the example, the sequences of five of the genes from strain SK264 are assigned by eMLSA.net as resident (that is, *S. parasanguinis*) or resident compatible, but the *pyk* and *tuf* genes are assigned as foreign. The locus view page allows the individual gene trees (in this case, the *tuf* tree) to be displayed to explore why the algorithm considers the sequences of an allele of a query strain to be resident, resident compatible or foreign (see Figure 6 for details). The colour codes for species clusters are as in Figure 2.

The number of house-keeping loci needed for electronic taxonomy using MLSA is not yet clear, and resolution between species clusters should improve with increasing numbers of loci. However, in practice, the number will be a balance between that needed for good resolution of species clusters and that which allows the generation of large databases and the uptake of the method by taxonomic laboratories. The number is likely to vary depending on the extent of sequence diversity within the group and the extent of interspecies recombination. In groups of species where recombination is very rare, resolution between clusters will depend on the number of polymorphic sites, and the aim of MLSA is to use sufficient genes to obtain enough polymorphic sites to adequately resolve species clusters. For such highly clonal species three or four house-keeping genes may be adequate to resolve clusters. In groups where interspecies recombination occurs, multiple genes are required both to provide adequate numbers of polymorphic sites and to overcome the distorting effects of interspecies recombination at individual loci, by extracting adequate taxonomic signal from the weak signal present in each of the individual genes. Seven-locus MLSA schemes appear to be adequate for resolving closely related species that undergo relatively frequent recombination, such as the viridans streptococci.

It is tempting with a bacterial group that has been well studied to select for MLSA schemes those genes for which the individual trees give the best approximation of the expected relationships among strains. However, for a general approach to discerning clustering patterns this is unwise, as the expected clusters will not be known when poorly studied bacteria are examined. Initial MLSA studies should therefore use the concatenated sequences of all selected genes to build a tree to identify the sequence clusters. It is advisable when developing MLSA for a poorly studied group of species to select more genes than are needed for the final MLSA scheme (for example, eight or nine genes for a final seven-gene scheme), so that the least informative can be eliminated from the final scheme. The best gene(s) to eliminate can be identified by using the tree based on the concatenated sequences of all genes as the reference tree, and examining the maximum likelihood of each of the single gene trees as a fit to the reference tree, eliminating those with the lowest likelihood.

Although we have used the term electronic taxonomy, the assignment of strains to known species and the identification and acceptance of new species cannot realistically be automated as it requires the experience, knowledge and judgement of taxonomists. The assignment of species using MLSA for electronic taxonomy should therefore be a community activity, in which those interested in a particular taxonomic group can share their experience and knowledge to provide a consensual approach to deciding whether new sequence clusters should be assigned as new species. As with the MLST databases, eMLSA.net is an actively curated open-access website and submissions of MLSA data on further strains of those species that currently are represented by only a small number of strains will help to circumscribe these species. Similarly, the submission of strains that are difficult to speciate by standard methods may identify new sequence clusters, and submission of sets of strains of other closely related streptococcal species will expand the breadth of streptococcal species that can be assigned to species clusters. Furthermore, those who consider they have identified, using standard taxonomic methods, a new species within or closely related to the viridans group streptococci can perform MLSA and submit the strain(s) to the database to see whether it matches any already in the database or whether, once submitted, new strains that cluster with that of the proposed new species appear as the database grows. Developers of MLSA schemes for other closely related bacterial (or other microbial) species are encouraged to host their data at http://www.eMLSA.net to take advantage of the generic analysis software and to provide a single portal for the electronic taxonomy of many groups of microbial species. A site for the on-line speciation of vibrios is also available http://www.taxvibrio.lncc.br, although at present it only identifies the most similar species to a query strain using BLAST, rather than generating the more informative sequence clusters [22].

Conclusion

The identification of species by electronic taxonomy using the MLSA approach provides a way of identifying species that is as least as valid as any other current taxonomic approach. The clustering patterns from MLSA provide a framework for species definition that can be further validated or revised by mapping on additional phenotypic, ecological or genomic data. Finally, this approach to electronic taxonomy, and our generic MLSA website, are not restricted to identifying and circumscribing bacterial species and could be applied to the Archaea and some eukaryotic microorganisms.

Methods

Bacterial strains

The Mitis group streptococci studied here included strains (numbers of strains are shown in parentheses) assigned by standard taxonomic procedures (or, subsequently, from their clustering patterns) as *S. australis* (10), *S. cristatus* (5), *S. gordonii* (28), *S. infantis* (24), *S. mitis* (88), *S. oligofermentans* (1), *S. oralis* (84), *S. parasanguinis* (25), *S. peroris* (1), *S. pneumoniae* (51), *S. pseudopneumoniae* (20), *S. sanguinis* (25) and *S. sinensis* (1). Strains of *S. anginosus* (7), *S. constellatus* (4) and *S. intermedius* (2) (Anginosus group streptococci), *S. salivarius* (4) and *S. vestibularis* (4) (Salivarius group streptococci) and *S. pyogenes* (17) (Pyogenic group streptococci) and *S. agalactiae* (1) were also studied. Sequences of the MLSA loci of the third member of the Salivarius group, *S. thermophilus*, were obtained from the type strain (ATCC 19258^T) and from the genome sequences of strains LMD-9, CNRZ1066 and LMG 18311 (GenBank: CP000421, CP000024 and CP000023). Two strains of *S. pseudopneumoniae* (one of which is the type strain ATCC BAA-960^T) were from the study of Arbique et al [45], who proposed this new Mitis group species. Another 21 strains were considered to be this species as they (and the above two strains) formed a well-resolved sequence cluster using the MLSA scheme based on six MLST loci [11]. Three of these were subsequently reassigned as uncertain species (see Results).

The type strains of all of the viridans species, except *S. pneumoniae*, were included in the strain collection. Strains assigned, or tentatively assigned, to the above viridans species (except *S. pneumoniae* and *S. pseudopneumoniae*) were selected from the collection of MK and were sent (in most cases without the species name assigned by the laboratory of MK) to Imperial College London for MLSA. In addition, a number of strains that were difficult to assign to the recognized species were included, as well as 52 viridans streptococci recovered in a UK hospital from infections in neutropenic patients [46]. The *S. pyogenes* and *S.*

pneumoniae strains were from the collections of WPH and BGS. In total there were 420 sets of sequences, corresponding to those obtained from 399 viridans group strains (plus three extracted from *S. thermophilus* genome sequences), 17 obtained from *S. pyogenes* strains and one from *S. agalactiae*. Details of all strains are available as Additional file 2.

The sequences of the seven MLSA loci were also obtained from the complete genome sequences of *S. mutans* UA159 (GenBank:AE014133), *S. uberis* 0140J (Sanger Institute website; http://www.sanger.ac.uk), *S. equi* 4047 (Sanger Institute website) and *S. suis* P1/7 (Sanger Institute website). The positions of the latter species on the reference tree are not shown in this paper as we have not examined whether all seven MLSA genes can be amplified from these species with the viridans group primer set, but there is an option to include these strains on the reference tree generated at the eMLSA.net website.

Assignment of strains to viridans species using standard taxonomic methods

The strains from the laboratory of MK were initially identified over a period of 20 years according to the current taxonomic status by a combination of extensive phenotypic analysis, sequencing of 16S rRNA and sequencing of selected house-keeping genes [16,18,25]. Serotypable *S. pneumoniae* strains were identified by standard methods; non-typable presumptive pneumococci were assigned to either *S. pneumoniae* or *S. pseudopneumoniae* as described previously [11].

Multilocus sequence analysis of viridans group streptococci

Suitable loci for the discrimination of species clusters must be present in all strains of the species under study, and should be sufficiently conserved that a fragment of each locus can readily be amplified from each species using a single set of primers, but sufficiently diverse that they are useful in resolving species clusters. Two of the loci used in the new MLSA scheme (*rpoB* and *sodA*) were included in the study of Hoshino et al [16]. To select additional house-keeping loci, we used the multi-genome homology comparison tool available through the Comprehensive Microbial Resource (CMR; http://cmr.jcvi.org/) to search for proteins present in all available streptococcal genomes that had at least 80% amino acid sequence identity. This retrieved 138 proteins and, from those that had house-keeping functions, the corresponding genes were selected at random and examined for their suitability as candidate MLSA loci. An alignment was produced of the sequences of each candidate locus from *S. agalactiae* NEM3 (GenBank:AL732656), *S. mitis* NCTC 12261 (available at J Craig Venter Institute website; http://www.jcvi.org), *S. mutans* UA159 (GenBank:AE014133), *S. pneumoniae* TIGR4 (GenBank:AE005672), *S. pyogenes* MGAS10394 (GenBank:CP000003), *S. thermophilus* LMD-9 (GenBank:CP000421) and *Lactococcus lactis* II1403 (GenBank:AE005176). Degenerate primers were designed for each selected locus, based on stretches of six amino acids that were conserved in all these species, which would allow an internal gene fragment greater than about 350 bp to be fully sequenced on both strands, and in which no indels were found in any of these genomes.

The pairs of PCR primers were each checked for their ability to amplify the correct gene fragment from a diverse subset of the collection of viridans group streptococci. Those primers that failed to amplify the correct fragment from all of the subset of strains were replaced with primers for amplifying a fragment of another of the conserved house-keeping genes, until a set of eight genes and PCR primers were obtained that allowed amplification of an internal gene fragment of each gene from the subset of strains. On characterizing the full strain set it was found that the primers for *guaA* (and two alternative pairs of primers) failed to amplify the fragment from a small number of strains and this gene was dropped from the final MLSA scheme. The genes, gene products, primer sequences, length of the sequences used in the MLSA scheme, and the annealing temperatures for amplification are shown in Table 1.

The gene fragments were amplified by PCR (30 cycles) using, for each gene fragment, a single annealing temperature for all strains (Table 1), and were sequenced on both strands, using the primers for the initial amplification, with an ABI 3700 or ABI 3730xl DNA analyzer. The sequences were aligned and trimmed to defined start and end positions using MEGA version 4 [47]. The trimmed sequences of the seven gene fragments from strains of each species can be found at http://viridans.eMLSA.net. The sequences of the seven gene fragments from each strain were joined in-frame, in the order *map-pfl-ppaC-pyk-rpoB-sodA-tuf*, to generate a single 3063 bp concatenated sequence. Unrooted individual gene trees, and trees obtained using the concatenated sequences, were generated by neighbour-joining or minimum evolution from the proportions of sequence differences between all strains using MEGA version 4 [47]. The robustness of the nodes was evaluated by bootstrapping (1000 replicates). All of the sequences (including the concatenated sequences and the sequences of each locus) from the 417 strains can be downloaded from http://viridans.eMLSA.net.

Electronic taxonomy website

A generic web application for electronic taxonomy was developed at http://www.eMLSA.net. This open-access site serves three main functions. It provides a database of

all characterized strains of a group of similar species (in this case, all of the 420 streptococcal strains characterized by MLSA), to which new strains can be added by the user community. Secondly, it allows the species assignment of query strains and, thirdly, it allows the identification of alleles that appear to have been introduced into the query strain from a closely related species.

The database (available to download at http://viridans.eMLSA.net) consists of a table containing strain information including, for each entry, the strain name; the species name assigned by MLSA and by standard taxonomic methods (a comment field provides the basis for these species assignments; another field allows the entry of any unusual features of the strain compared with typical strains of the assigned species); the country of origin, date of isolation and source of the strain (nasopharynx, blood, etc); whether it is a type strain; the names and addresses of the laboratories or culture collections that hold the strain; the contact details of the person who submitted the strain data; and any publication relating to the properties or taxonomic status of the strain. Further database tables hold the sequences that correspond to the alleles at each locus.

On-line species assignments

To observe the patterns of clustering of the strains in the database, and to show the species cluster into which a user's strain falls, the seven edited and trimmed sequences of a query strain are cut and pasted into individual text boxes. Sequences are checked for the correct length for the locus and the absence of any characters other than A, C, G and T; the submitter is alerted if any entered sequence is of non-standard length or contains unexpected characters. If the sequences are all of the expected length, they are concatenated in-frame, and the query sequence is added to the concatenated sequences of all strains in the MLSA database. To construct the tree, a module of MEGA was used (kindly provided by Dr Sudhir Kumar, Arizona State University), which takes the concatenated sequences and generates in Newick format either a neighbour-joining or minimum evolution tree, incorporating options for a range of genetic distance corrections and for bootstrap resampling. If the sequences have insertions or deletions (indels), a modification of the software is used that generates an alignment of the concatenated sequences, which is used by the MEGA module to produce the tree. A module of Phylowidget (http://www.phylowidget.org; [48]) developed using the visual programming language Processing http://www.processing.org, provides easy communication between the tree diagram and the underlying database. Phylowidget is embedded within the website and is passed the Newick file generated by the MEGA module for display as a radial tree (the tree can also be displayed in the standard rectangular format or in circular format). The position of the query strain is highlighted on the tree and, using the intuitive click-drag-zoom interface, the tree can be explored further. Individual gene trees can also be constructed by the same method. A standard database search facility is available allowing details of any strain or set of strains to be viewed, and a search box on the reference tree allows the position of any strain to be highlighted on the tree.

A species assignment can be proposed if the concatenated sequence of the strain falls within a known species cluster, and it can be assigned as an unknown species if it falls within a cluster with no assigned species name, or as a unique divergent strain if it clusters apart from any of the sequence clusters. In addition, a list of the species names assigned to the five concatenated sequences in the database that are most similar to the concatenated query sequence is shown, together with the % sequence similarity values to these strains (an alert shows if there are less than five strains of this species in the database).

The tree produced on-line must be the same for all users of the website and only a single tree-building method is employed within eMLSA.net and, as the tree is to discern clustering patterns rather than phylogeny, an unrooted tree is generated and is displayed by default in radiation format, although other tree display formats can be selected by the user. The default settings for the trees produced for a particular MLSA scheme (for example, minimum evolution or neighbour-joining), can be changed by the curator of the scheme to any of those provided by the MEGA module, and one of a number of different genetic distance corrections can be applied. All of the concatenated sequences can be downloaded from the website in MEGA format to allow other types of tree to be constructed off-line, using MEGA4 or alternative tree-building packages.

On-line assignment of source of individual alleles

To assign the alleles at each locus in a strain as 'resident', 'resident compatible' or 'foreign', the sequence at each locus of a query strain that has been assigned to species A is compared with all other sequences at that locus, and the species assigned to the strains with the five most similar sequences are displayed (an alert shows if there are less than five strains of the latter species in the database). If the most similar sequences are all from strains of species A, the sequence at that locus is defined as a resident allele. However, if the best matches are sequences from strains assigned to another species, the allele is considered to be foreign, as long as species A and the species cluster assigned for that allele are well resolved on the individual gene tree. Resident compatible is used where the allele sequence is within a cluster of two or more species (including species A) that are not resolved on the gene

tree. Resolution and lack of resolution between clusters are guided by the bootstrap values for the nodes on the individual gene trees separating the relevant species. A cluster is defined as resolved if the bootstrap value of the node separating it from neighbouring clusters is ≥ 80%.

To ensure that the assignment of alleles as foreign is dependent on good resolution on the gene tree, a look-up table was generated for each locus, which lists those species that could not be resolved on the gene tree for that locus. For example, only two of the individual gene trees (*map* and *pyk*) could resolve *S. mitis* and *S. pseudopneumoniae* strains (see Additional file 1). Thus, if the concatenated sequence of a particular strain was assigned to the *S. mitis* cluster, but it had its five closest matches at *rpoB* with *S. pseudopneumoniae* strains (or a mixture of *S. mitis* and *S. pseudopneumoniae* strains), this would not be scored as a foreign *rpoB* allele within the *S. mitis* strain, as the look-up table indicates that the *rpoB* tree cannot resolve *S. mitis* and *S. pseudopneumoniae*. This result is therefore consistent with the *S. mitis rpoB* allele being a resident allele (resident compatible). Where the best matches are not from species A, but are from two species that are both well resolved from species A (but not from each other), the allele at that locus is assigned as foreign, and the source is shown to be one or other of the latter two species. The website then returns the predicted resident, resident compatible or foreign status of each of the sequences at the seven loci for the query strain (or for any strain in the database).

For each locus, the website also shows the % sequence divergence of the query sequence from the best-matching sequences. These values identify sequences that are highly divergent from other alleles of the same species, as one or more divergent alleles could result in anomalous clustering of the concatenated sequence of the strain.

Strain submission to eMLSA.net

Submission to the eMLSA.net curator of the details of new strains, with their sequences at the seven loci and the original sequencer trace files, and proposed species assignment, is strongly encouraged to build up the viridans group electronic taxonomy database. After checking the sequences for quality and correct trimming, and the proposed species assignment, each new strain and the strain details are added to the database. Submission of viridans group strains that do not fall within the existing sequence clusters are particularly encouraged to allow the identification of potential new species within this group.

Authors' contributions

CJB carried out the experimental work, DMA developed the eMLSA website, GEJ produced a version of Phylowidget for the eMLSA website, and MK provided strains and expertise in streptococcal taxonomy. WPH and BGS conceived the project. All authors contributed to the writing of the paper. All authors have read and approved the final manuscript.

Additional material

Additional file 1

Supplementary figure. *Individual gene trees. (A-C) Neighbour-joining trees were produced for each MLSA gene using the different sequences at each locus present among the 420 strains. The colour codes for species clusters are as in Figure 2.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1741-7007-7-3-S1.ppt]

Additional file 2

Supplementary table. *Database of strain*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1741-7007-7-3-S2.xls]

Acknowledgements

We thank Dr Sudhir Kumar for kindly providing the module of MEGA. This work was funded by the Wellcome Trust (grant 030662) and the Royal Society. WPH is supported by a Royal Society University Research Fellowship and BGS by a Wellcome Trust Principal Research Fellowship.

References

1. Cohan FM: **Towards a conceptual and operational union of bacterial systematics, ecology, and evolution.** *Phil Trans R Soc B* 2006, **361:**1985-1996.
2. Gevers D, Cohan F, Lawrence J, Spratt BG, Coenye T, Feil EJ, Stackebrandt E, Manfio G, Peer Y Van de, Nesme X, Thompson F, Swings J: **Re-evaluating bacterial species.** *Nature Microbiol Rev* 2005, **3:**733-739.
3. Gevers D, Dawyndt P, Vandamme P, Willems A, Vancanneyt M, Swings J, de Vos P: **Stepping stones to a new prokaryotic taxonomy.** *Phil Trans Roy Soc B* 2006, **361:**1911-1916.
4. Staley JT: **The bacterial species dilemma and the genomic-phylogenetic species concept.** *Phil Trans R Soc B* 2006, **361:**1899-1909.
5. Achtman M, Wagner M: **Microbial diversity and the genetic nature of microbial species.** *Nat Rev Microbiol* 2008, **6:**431-440.
6. Cohan FM: **What are bacterial species?** *Ann Rev Microbiol* 2002, **56:**457-487.
7. Fraser C, Hanage WP, Spratt BG: **Recombination and the nature of bacterial speciation.** *Science* 2007, **315:**476-480.
8. Vandamme P, Pot B, Gillis M, de Vos P, Kersters K, Swings J: **Polyphasic taxonomy, a consensus approach to bacterial systematics.** *Microbiol Rev* 1996, **60:**407-438.
9. Wayne LG, Brenner DJ, Colwell RR, Grimont PAD, Kandler O, Krichevsky MI, Moore LH, Moore WEC, Murray RGE, Stackebrandt E, Starr MP, Trüper HG: **Report of the ad hoc committee on reconciliation of approaches to bacterial systematics.** *Int J System Bact* 1987, **37:**463-464.
10. Stackebrandt E, Frederiksen W, Garrity GM, Grimont PAD, Kämpfer P, Maiden MCJ, Nesme X, Rosselló-Mora R, Swings J, Trüper HG, Vauterin L, Wards AC, Whitman WB: **Report of the ad hoc committee for the re-evaluation of the species definition in bacteriology.** *Int J Sys Evol Microbiol* 2002, **52:**1043-1047.
11. Hanage WP, Fraser C, Spratt BG: **Sequences, sequence clusters and bacterial species.** *Phil Trans Roy Soc B* 2006, **361:**1917-1927.
12. Nørskov-Lauritsen N, Bruun B, Kilian M: **Multilocus sequence phylogenetic study of the genus Haemophilus with descrip-**

tion of Haemophilus pittmaniae sp. nov. *Int J Syst Evol Microbiol* 2005, **55:**449-456.
13. Godoy D, Randle G, Simpson AJ, Aanensen DM, Pitt TL, Kinoshita R, Spratt BG: **Multilocus sequence typing and evolutionary relationships among the causative agents of melioidosis and glanders, *Burkholderia pseudomallei* and *Burkholderia mallei*.** *J Clin Microbiol* 2003, **41:**2068-2079.
14. Hanage WP, Fraser C, Spratt BG: **Fuzzy species in recombinogenic bacteria.** *BMC Biol* 2005, **3:**6.
15. Hanage WP, Kaijalainen T, Herva E, Saukkoriipi A, Syrjänen R, Spratt BG: **Using multilocus sequence data to define the pneumococcus.** *J Bacteriol* 2005, **187:**6223-6230.
16. Hoshino T, Fujiwara T, Kilian M: **Use of phylogenetic and phenotypic analyses to identify nonhemolytic streptococci isolated from bacteremic patients.** *J Clin Microbiol* 2005, **43:**6073-6085.
17. Bennett JS, Jolley KA, Sparling PF, Saunders NJ, Hart CA, Feavers IM, Maiden MCJ: **Species status of *Neisseria gonorrhoeae*: evolutionary and epidemiological inferences from multilocus sequence typing.** *BMC Biol* 2007, **5:**35.
18. Kilian M, Poulsen K, Blomqvist T, Håvarstein LS, Bek-Thomsen M, Tettelin H, Sørensen UB: **Evolution of *Streptococcus pneumoniae* and its close commensal relatives.** *PLoS ONE* 2008, **3:**e2683.
19. Martens M, Dawyndt P, Coopman R, Gillis M, De Vos P, Willems A: **Advantages of multilocus sequence analysis for taxonomic studies: a case study using 10 housekeeping genes in the genus *Ensifer* (including former *Sinorhizobium*).** *Int J Syst Evol Microbiol* 2008, **58:**200-214.
20. Sawabe T, Kita-Tsukamoto K, Thompson FL: **Inferring the evolutionary history of vibrios by means of multilocus sequence analysis.** *J Bacteriol* 2007, **189:**7932-7936.
21. Postic D, Garnier M, Baranton G: **Multilocus sequence analysis of atypical *Borrelia burgdorferi* sensu lato isolates – description of *Borrelia californiensis* sp. nov., and genomospecies 1 and 2.** *Int J Med Microbiol* 2007, **297:**263-271.
22. Thompson FL, Gomez-Gil B, Vasconcelos AT, Sawabe T: **Multilocus sequence analysis reveals that *Vibrio harveyi* and *V. campbellii* are distinct species.** *Appl Environ Microbiol* 2007, **73:**4279-4285.
23. Maiden MCJ, Bygraves JA, Feil E, Morelli G, Russell JE, Urwin R, Zhang Q, Zhou J, Zurth K, Caugant DA, Feavers IM, Achtman M, Spratt BG: **Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms.** *Proc Natl Acad Sci USA* 1998, **95:**3140-3145.
24. Kilian M: ***Streptococcus and Lactobacillus*.** In *Topley and Wilson's Microbiology and Microbial Infections* Edited by: Borriello P, Murray PR, Funke G. London: Hodder Arnold; 2005:833-881.
25. Kilian M, Mikkelsen L, Henrichsen J: **Taxonomic study of viridans streptococci: description of *Streptococcus gordonii* sp. Nov. and emended descriptions of *Streptococcus sanguis* (White and Niven, 1946), *Streptococcus oralis* (Bridge and Sneath, 1982), and *Streptococcus mitis* (Andrewes and Horder, 1906).** *Int J Syst Bacteriol* 1989, **39:**471-484.
26. Whiley RA, Fraser HY, Douglas CW, Hardie JM, Williams AM, Collins MD: ***Streptococcus parasanguis sp. nov.*, an atypical viridans *Streptococcus* from human clinical specimens.** *FEMS Microbiol Lett* 1990, **56:**115-121.
27. Whiley RA, Beighton D: **Emended descriptions and recognition of *Streptococcus constellatus*, *Streptococcus intermedius*, and *Streptococcus anginosus* as distinct species.** *Int J Syst Bacteriol* 1991, **41:**1-5.
28. Whiley RA, Duke B, Hardie JM, Hall LM: **Heterogeneity among 16S-23S rRNA intergenic spacers of species within the '*Streptococcus milleri* group'.** *Microbiology* 1995, **141:**1461-1467.
29. Whiley RA, Hall LM, Hardie JM, Beighton D: **A study of small-colony, beta-haemolytic, Lancefield group C streptococci within the anginosus group: description of *Streptococcus constellatus* subsp. pharyngis subsp. nov., associated with the human throat and pharyngitis.** *Int J Syst Bacteriol* 1999, **49:**1443-1449.
30. Kawamura Y, Hou X-G, Todome Y, Sultana F, Hirose K, Shu SE, Ezaki T, Ohkuni H: ***Streptococcus peroris* sp. nov. and *Streptococcus infantis* sp. nov., new members of the *Streptococcus mitis* group, isolated from human clinical specimens.** *Int J Syst Bacteriol* 1998, **48:**921-927.
31. Kawamura Y, Whiley RA, Shu SE, Ezaki T, Hardie JM: **Genetic approaches to the identification of the mitis group within the genus *Streptococcus*.** *Microbiology* 1999, **145:**2605-2613.
32. Kawamura Y, Whiley RA, Zhao L, Ezaki T, Hardie JM: **Taxonomic study of "tufted mitior" strains of streptococci (*Streptococcus sanguinis* biotype 11); recognition of a new genospecies.** *Syst Appl Microbiol* 2000, **23:**245-250.
33. Whatmore AM, Efstratiou A, Pickerill AP, Broughton K, Woodard G, Sturgeon D, George R, Dowson CG: **Genetic relationships between clinical isolates of *Streptococcus pneumoniae*, *Streptococcus oralis*, and *Streptococcus mitis*: characterization of "Atypical" pneumococci and organisms allied to *S. mitis* harboring *S. pneumoniae* virulence factor-encoding genes.** *Infect Immun* 2000, **68:**1374-1382.
34. Dowson CG, Hutchison A, Brannigan JA, George RC, Hansman D, Liñares J, Tomasz A, Smith JM, Spratt BG: **Horizontal transfer of penicillin-binding protein genes in penicillin-resistant clinical isolates of *Streptococcus pneumoniae*.** *Proc Natl Acad Sci USA* 1989, **86:**8842-8846.
35. Laible G, Spratt BG, Hakenbeck R: **Interspecies recombinational events during the evolution of altered PBP 2× genes in penicillin-resistant clinical isolates of *Streptococcus pneumoniae*.** *Mol Microbiol* 1991, **5:**1993-2002.
36. Hakenbeck R, Balmelle N, Weber B, Gardès C, Keck W, de Saizieu A: **Mosaic genes and mosaic chromosomes: intra- and interspecies genomic variation of *Streptococcus pneumoniae*.** *Infect Immun* 2001, **69:**2477-2486.
37. Dowson CG, Coffey TJ, Kell C, Whiley RA: **Evolution of penicillin resistance in *Streptococcus pneumoniae*; the role of *Streptococcus mitis* in the formation of a low affinity PBP2B in *S. pneumoniae*.** *Mol Microbiol* 1993, **9:**635-643.
38. Stanhope MJ, Walsh SL, Becker JA, Italia MJ, Ingraham KA, Gwynn MN, Mathie T, Poupard JA, Miller LA, Brown JR, Amrine-Madsen H: **Molecular evolution perspectives on intraspecific lateral DNA transfer of topoisomerase and gyrase loci in *Streptococcus pneumoniae*, with implications for fluoroquinolone resistance development and spread.** *Antimicrob Agents Chemother* 2005, **49:**4315-4326.
39. Enright MC, Spratt BG: **A multilocus sequence typing scheme for *Streptococcus pneumoniae*: identification of clones associated with serious invasive disease.** *Microbiology* 1998, **144:**3049-3060.
40. Hohwy J, Kilian M: **Clonal diversity of the *Streptococcus mitis* biovar 1 population in the human oral cavity and pharynx.** *Oral Microbiol Immunol* 1995, **10:**19-25.
41. Tong H, Gao X, Dong X: ***Streptococcus oligofermentans* sp. nov., a novel oral isolate from caries-free humans.** *Int J Syst Evol Microbiol* 2003, **53:**1101-1104.
42. Majewski J, Zawadzki P, Pickerill P, Cohan FM, Dowson CG: **Barriers to genetic exchange between bacterial species: *Streptococcus pneumoniae* transformation.** *J Bacteriol* 2000, **182:**1016-1023.
43. Falush D, Torpdahl M, Didelot X, Conrad DF, Wilson DJ, Achtman M: **Mismatch induced speciation in Salmonella: model and data.** *Phil Trans Roy Soc B* 2006, **361:**2045-2053.
44. Konstantinidis KT, Tiedje JM: **Genomic insights that advance the species definition for prokaryotes.** *Proc Natl Acad Sci USA* 2005, **102:**2567-2572.
45. Arbique JC, Poyart C, Trieu-Cuot P, Quesne G, Carvalho MS, Steigerwalt AG, Morey RE, Jackson D, Davidson RJ, Facklam RR: **Accuracy of phenotypic and genotypic testing for identification of *Streptococcus pneumoniae* and description of *Streptococcus pseudopneumoniae* sp. nov.** *J Clin Microbiol* 2004, **42:**4686-4696.
46. Hanage WP, Cohen J: **Stimulation of cytokine release and adhesion molecule expression by products of Viridans streptococci.** *J Infect Dis* 2002, **185:**357-367.
47. Tamura K, Dudley J, Nei M, Kumar S: **MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0.** *Mol Biol Evol* 2007, **24:**1596-1599.
48. Jordan GE, Piel WH: **PhyloWidget: web-based visualizations for the tree of life.** *Bioinformatics* 2008, **24:**1641-1642.

OPEN ACCESS Freely available online

PLoS one

EpiCollect: Linking Smartphones to Web Applications for Epidemiology, Ecology and Community Data Collection

David M. Aanensen[1]*, Derek M. Huntley[2], Edward J. Feil[3], Fada'a al-Own[3], Brian G. Spratt[1]

1 Department of Infectious Disease Epidemiology, Imperial College London, London, United Kingdom, **2** Centre for Bioinformatics, Imperial College London, London, United Kingdom, **3** Department of Biology and Biochemistry, University of Bath, Bath, United Kingdom

Abstract

Background: Epidemiologists and ecologists often collect data in the field and, on returning to their laboratory, enter their data into a database for further analysis. The recent introduction of mobile phones that utilise the open source Android operating system, and which include (among other features) both GPS and Google Maps, provide new opportunities for developing mobile phone applications, which in conjunction with web applications, allow two-way communication between field workers and their project databases.

Methodology: Here we describe a generic framework, consisting of mobile phone software, EpiCollect, and a web application located within www.spatialepidemiology.net. Data collected by multiple field workers can be submitted by phone, together with GPS data, to a common web database and can be displayed and analysed, along with previously collected data, using Google Maps (or Google Earth). Similarly, data from the web database can be requested and displayed on the mobile phone, again using Google Maps. Data filtering options allow the display of data submitted by the individual field workers or, for example, those data within certain values of a measured variable or a time period.

Conclusions: Data collection frameworks utilising mobile phones with data submission to and from central databases are widely applicable and can give a field worker similar display and analysis tools on their mobile phone that they would have if viewing the data in their laboratory via the web. We demonstrate their utility for epidemiological data collection and display, and briefly discuss their application in ecological and community data collection. Furthermore, such frameworks offer great potential for recruiting 'citizen scientists' to contribute data easily to central databases through their mobile phone.

Citation: Aanensen DM, Huntley DM, Feil EJ, al-Own F, Spratt BG (2009) EpiCollect: Linking Smartphones to Web Applications for Epidemiology, Ecology and Community Data Collection. PLoS ONE 4(9): e6968. doi:10.1371/journal.pone.0006968

Editor: Simon I. Hay, University of Oxford, United Kingdom

Received June 23, 2009; **Accepted** August 6, 2009; **Published** September 16, 2009

Copyright: © 2009 Aanensen et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This work was supported by the Wellcome Trust. BGS is a Wellcome Trust Principal Research Fellow. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: d.aanensen@imperial.ac.uk

Introduction

Increasingly, the web is used in biology as a method to collate and analyse data collected by field workers (for example, ecologists or epidemiologists) from multiple locations. A single central database, accessed through a website, can provide the tools for the submission, visualisation and analysis of data collected by many users from many different locations. A good example is provided by the molecular epidemiological databases that have been developed for a number of bacterial and fungal pathogens, where communities of researchers and public health laboratories submit data on strains of individual pathogens, including basic epidemiological information and molecular typing data (for example, http://www.mlst.net) [1,2]. By linking these databases to Google Maps, the geographic distribution of pathogen genotypes can be displayed and the databases can be explored and analysed conveniently (http://maps.mlst.net).

Central web databases for the collation of data submitted by field workers are also of utility to ecologists, and for mapping distributions of endangered species, and a single web portal can significantly decrease the time between collation and analysis of data collected by multiple users. However, transferring data to these databases (e.g. from field notebooks) can be tedious and in some cases will be subject to transcription errors. Alternatives to paper collection, such as personal digital assistants (PDAs) or standard mobile phones, with subsequent synchronization with a user's laptop or desktop computer, offer attractive opportunities for remote data collection in many areas. For example, EpiSurveyor (www.episurveyor.org) allows the use of standard Nokia mobile phones for the collection of text-based data and has been utilised in resource-poor areas of Africa for many kinds of data collection.

Recently, the availability of a new generation of mobile phones, bracketed under the term 'smartphones', offers further novel approaches by the development of software that allows submission and two-way retrieval of data from the field to a central database by mobile phone [3,4].

Smartphones offer PC-like functionality and web connectivity far superior to traditional mobile phones. Built-in GPS receivers provide the detailed location of the phone, accelerometers can recognise changes in movement, and cameras provide the ability to record static images as well as video. Data networks allow

in-built software to access the internet providing web browsing, email and mapping (such as Google Maps) and office documents can be viewed/edited using touch screen keyboards (or hardware keyboards) for textual input.

The versatility of these phones is greatly enhanced by the ability for software developers to produce and deploy their own applications directly on the phones. Google, in conjunction with the Open Handset Alliance (a group of more than 45 technology and mobile phone companies), has recently released the Android operating system for mobile devices (http://www.android.com). Android is a completely open source operating system for, but not restricted to, mobile phones. Android gives software developers the freedom to access all aspects of a smartphone's functionality and software is delivered to phones through the 'Android Market', an in-built application that provides a user with a 'click and install' method based on browsing a list of all developer-produced software. The first handsets running Android were released in October 2008 and currently four variants are commercially available, the T-Mobile G1 and G2, the Vodafone HTC Magic and the Orange HTC Hero with other manufacturers releasing handsets in the near future.

The increased sophistication of smartphones and developments within web technology, loosely grouped under the heading of 'Web 2.0', allow web developers to produce 'mashups' of data from different sources, the most well known examples of which involve displaying data in its geographical context using Google Maps [5]. The availability of 'geographically aware' mobile phones which can interact with central web databases, visualising data within the geographical context that they were collected, provides the opportunity to develop new data collection, visualisation and analysis strategies for many kinds of biological data, particularly for field workers and epidemiologists.

Here we present a framework for such data collection, and for the two-way communication between field workers and their project databases, which takes advantage of the open source nature of Android (to develop mobile phone software) and open development tools for web applications (Google Maps and Google Earth). A key aspect of this approach is that, as well as providing a way of submitting data by phone to a project database, all data or selected data can be retrieved from the database and viewed and analysed using Google Maps, both via the web and on the mobile phone. In this way the field worker has essentially the same ability to display and analyse data from the project database on their mobile phone as they would have if they were back in their laboratory accessing their database via the web. Our initial examples focus on the use of this framework for the purposes of epidemiological data collection, but EpiCollect has been designed to be generic, and the same principles are applicable for a wide range of projects, some of which are discussed.

Results and Discussion

The framework

The framework consists of two major components. Firstly, we have developed generic software for Android, EpiCollect, which allows multiple data records to be entered and stored on a mobile phone (text variables, GPS position, photo etc.) and sent to a central web database. Secondly, we have developed a web application, located within www.spatialepidemiology.net that, for a specific project, allows the mapping, visualisation and analysis of data submitted to the central database. A schematic of the two-way connectivity between mobile phone and project database obtained using EpiCollect and spatialepidemiology.net is shown in Figure 1.

For a new project, data variables to be collected for each data record (sample) are firstly defined. These variables are then used in three ways. 1) To define the fields in the project-specific central database and the web forms for standard data entry via the project website. 2) To define the interface for text input for each data record within EpiCollect loaded onto a mobile phone. 3) To define the data values or ranges for filtering the data that are displayed at the project website or on the mobile phone.

For each project, the central registration of phones loaded with EpiCollect is undertaken using a curator's interface (Figure 1A), allowing multiple users to submit data centrally to the project database (Figure 1B). The project-specific website provides a public interface that displays all data using Google Maps or Google Earth (Figure 1C). Points on the map can be coloured either by the user (or group of users) who submitted the data, or alternatively, based on a variable submitted to the database. Furthermore, the display of data can be filtered based on the variables collected. For example, filtering allows display of only those samples obtained within a specific range of a variable or within a specific time period, and graphical summaries based on the variables collected are displayed (Figure 1D). Each time a new data record is sent from EpiCollect to the central database a new point appears on the map and the variables become immediately available for filtering. This allows the progress of the project to be monitored in real-time and, furthermore, a central project curator can, while fieldwork is underway, communicate with individual users, or all users, via instant messaging provided by Google Talk.

EpiCollect

When EpiCollect is launched on the mobile phone three options are initially available, 'New Entry', 'List Entries' and 'Display Map' (Figure 2A). Selecting 'New Entry' creates a new data record within the phone's on-board database and assigns a unique ID to the record. The latitude, longitude and altitude of the current position of the user is returned from the GPS unit of the phone (Figure 2B). Three new options are now available, 'Photo', 'Data' and 'Store'. Selecting photo allows an image to be taken using the phone's camera, which is assigned to the record (Figure 2C). Selecting 'Data' displays the data entry screen (Figure 2D), which can contain any standard form field (text fields, list boxes, check boxes etc.). These fields correspond to those created when defining the project database structure. Data can be entered via an on-screen keyboard (using the touch-screen) or via the hardware keyboard should this be part of the phone (e.g. T-Mobile G1). Following data entry, the 'Confirm' button returns the view to the Entry screen (Figure 2C). Selecting 'Store' saves the current record to the phone's database along with the date and time that the record was created. Subsequently, this process is repeated to create further records.

The 'List entries' option allows all records stored within the phone's database to be viewed and amended if necessary. From this screen, data can be synchronised with the central database with confirmation of successful data transfer (Figure 2E). The 'Display Map' option allows all records stored locally on the phone to be displayed on a map using the in-built Google Maps application (Figure 2F). Furthermore, records based on a particular variable (or all records) can be retrieved from the central database and displayed on the phone using Google Maps in conjunction with the locally stored records (Figure 2G). The standard click-drag-zoom interface to the map is available on the phone allowing map navigation and can, using the GPS, instantly zoom to the phone's current location (Figure 2H).

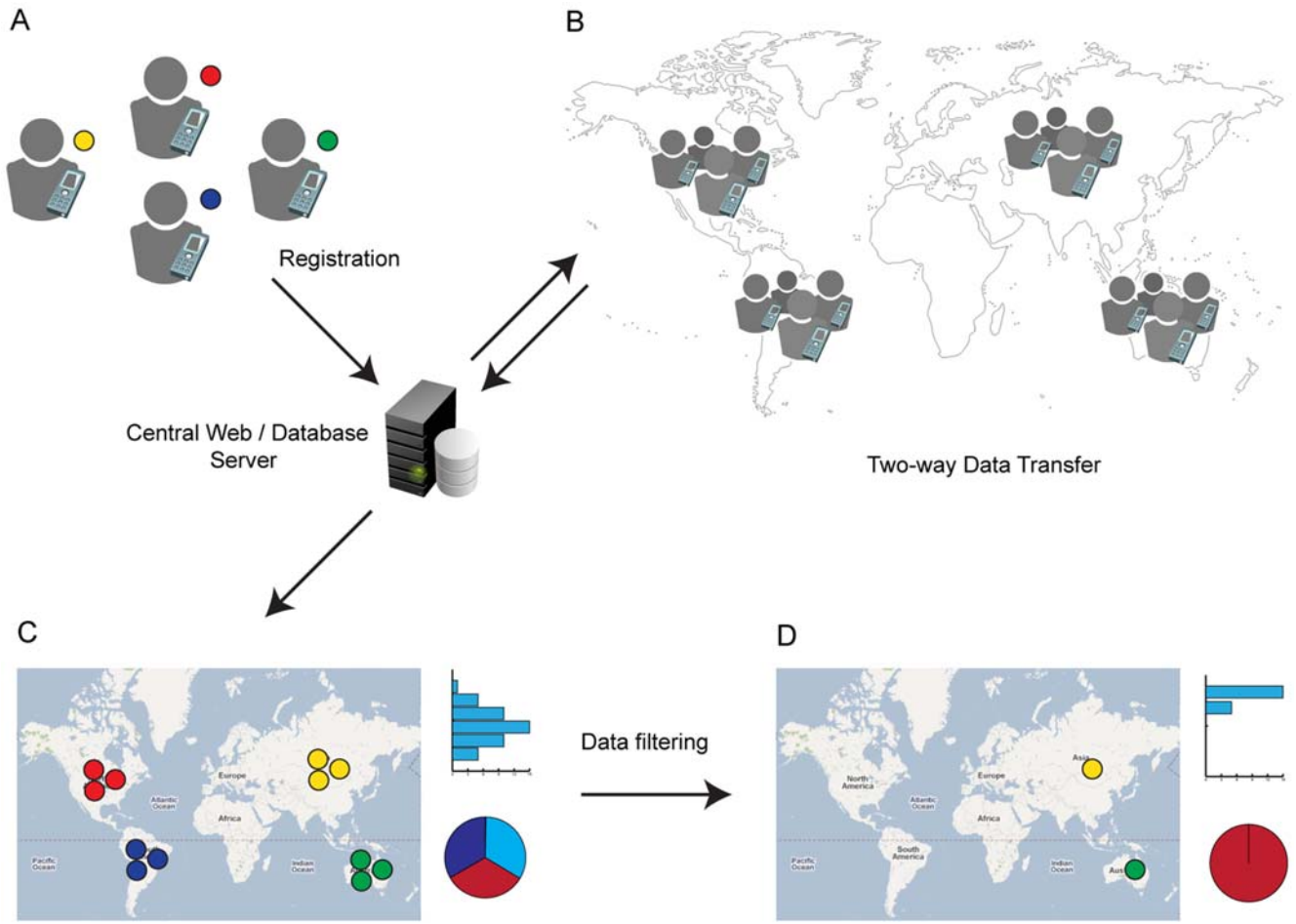


Figure 1. Data collection, collation and visualisation framework using EpiCollect and www.spatialepidemiology.net. See text for details.

Field Testing

A simple field test of EpiCollect was carried out in which soil samples were taken across Southern England by a single field worker, to obtain isolates of a soil bacterium (*Xenorhabdus nematophila*) that is pathogenic to insect larvae. At each location, a sample of soil was taken, along with a photograph of the area. Data variables were then recorded and entered into EpiCollect including soil pH, temperature and moisture. Data from each sampling location were then submitted via EpiCollect from the phone to the central database located at http://www.spatialepidemiology.net/epicollect/xenorhabdus.

Figure 3A shows the project-specific web interface and each of the data points on the map indicate a data record (sampling site) submitted via EpiCollect. All samples submitted are listed to the left of the map (Figure 3B) and, using the click-drag-zoom interface, exploration of the data can be undertaken. Clicking one of the sample points displays an 'info window' containing the variables collected (soil pH, temp, moisture, GPS position, date etc) and any photograph associated with the record (Figure 3C). Graphical summaries of the variables collected are displayed to the right of the map (Figure 3D) and a number of options are provided for filtering which of the samples points are displayed (Fig 3E). Time sliders allow filtering of data so that only samples collected within a chosen time window are displayed (Figure 3G).

As mentioned, the project-specific curator can communicate with the field workers directly from the web interface using the instant messaging tool, Google Talk. Figure 3F illustrates the chat interface seen by the curator (this is not available in the public view of the website). Chat requires that each submitter has a Gmail account and is sent from the website and automatically received on the phone as seen in Figure 3I. Message transcripts are automatically stored within the user's Gmail account for future reference. Figure 3H illustrates the view using Google Maps of data requested via the phone from the central database and Figure 3J shows the same data viewed at the project website using Google Earth.

Subsequently, the isolates of the bacterial pathogen recovered from the soil samples at each sampling point will be genotyped (by multilocus sequence typing; MLST [1]) and these molecular typing data will be added to the database, allowing further display options, such as the geographic distribution of particular *X. nematophila* genotypes.

Illustrative project involving multiple field workers

The simple field test illustrated the collection and submission of data from one field worker and the ability to view data retrieved from the central database on the phone using Google Maps. However, EpiCollect is designed for situations where multiple field workers submit data via mobile phone to a central database. For example, multiple field workers may submit data on animal deaths (e.g. surveillance of living and dead amphibians), or animal infections (e.g. using simple field diagnostic tests to distinguish

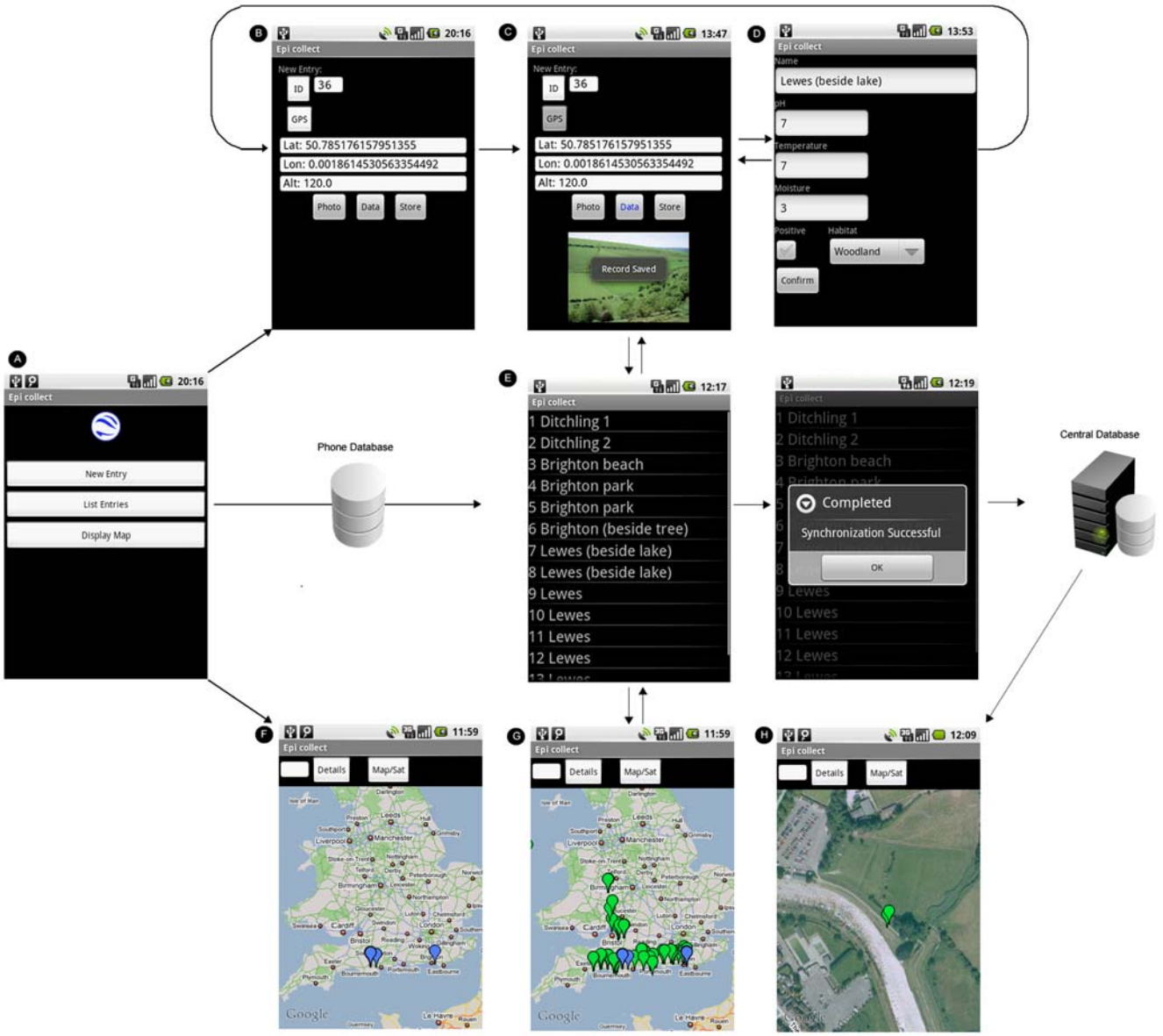


Figure 2. The EpiCollect workflow. See text for details.
doi:10.1371/journal.pone.0006968.g002

infected and non-infected animals), and the web application allows colour-coding (or symbol-based coding) of the data points submitted by phone to the project database by different field workers.

To demonstrate this further, Figure 4 shows the web interface for an illustrative dataset where data from each sampling point have been submitted via mobile phone to the project database by multiple field workers across Europe using EpiCollect. Field tests have been used to identify those animals infected with a specific bacterial pathogen and, as in many real life examples, samples from the infected animals are taken back to a laboratory for strain characterisation (in this case determining the serogroup of the strain and its susceptibility or resistance to an antibiotic). The illustrative data can be explored at http://www.spatialepidemiology.net/epicollect/demo.

Figure 4A shows the overall data summary with each point on the map representing sampling locations across Europe. The points are colour-coded based on whether the pathogen recovered from each of these locations is antibiotic-susceptible (green) or antibiotic-resistant (red). All data points are listed at the top left of the interface and the data variables are graphically represented to the right of the map. The distribution in the total sample of each serogroup, and the extent of resistance to the antibiotic are represented as pie charts, and the estimated ages of the infected animals are represented as a bar chart (<1 year, 1–5 years, >5 years). In Figure 4B the sampling points have been colour-coded based on the individuals who have submitted data from their phones. Using the filtering options shown in Figure 4A we can delimit the data displayed on the map to show only those samples where the animals were infected with serogroup 1 of the pathogen. As seen in Figure 4C, delimiting the data in this way also updates the graphical summaries and shows that all strains of serogroup 1 are antibiotic-resistant and, furthermore, that they are almost always isolated from young animals. In infectious disease epidemiology pathogen isolates are generally characterised at a much finer level than provided by serogrouping using discriminatory molecular

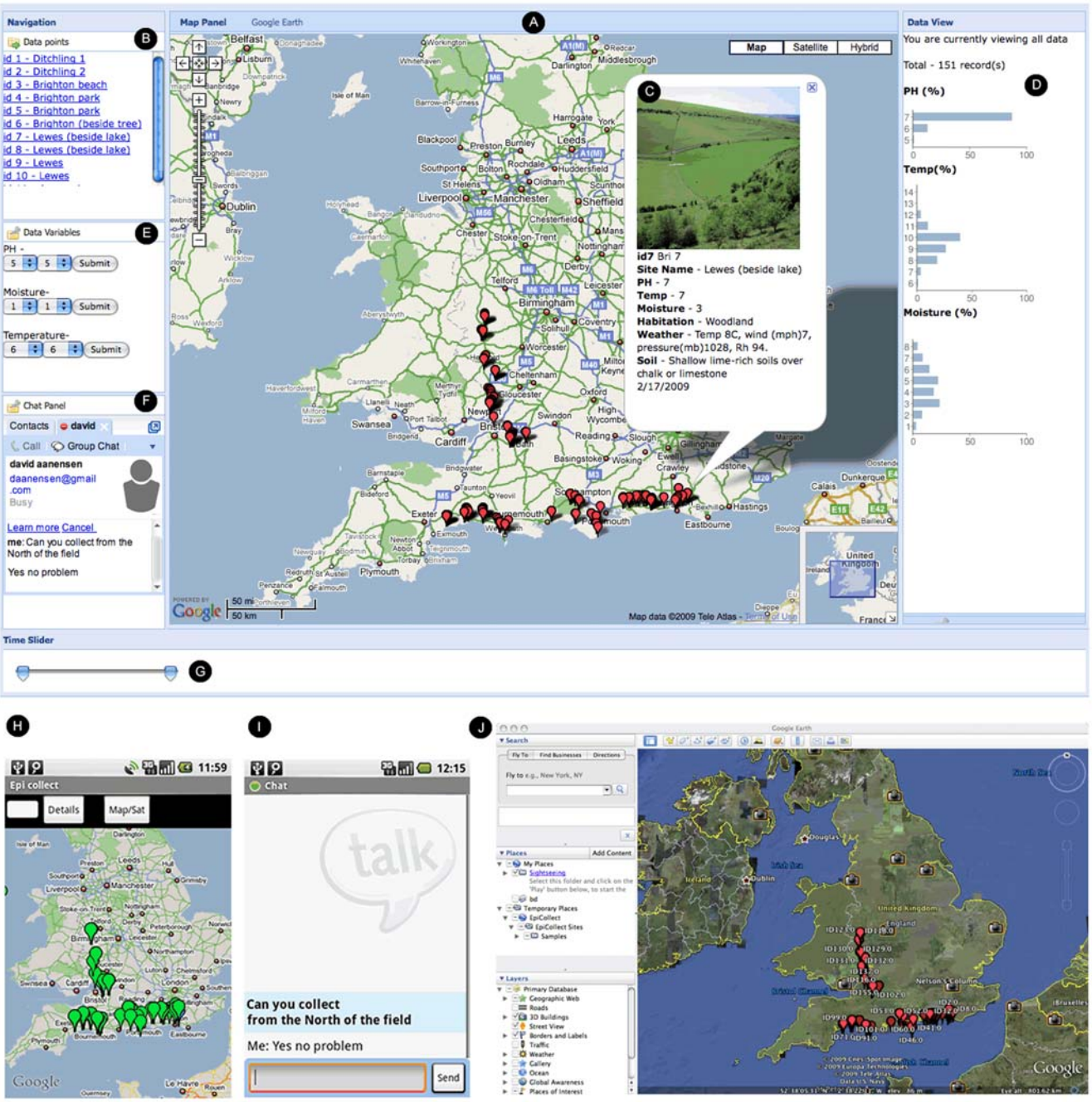


Figure 3. Simple field test of EpiCollect with data submitted to a project-specific website. See text for details.
doi:10.1371/journal.pone.0006968.g003

typing techniques and, after this has been performed, the geographic distribution and extent of antibiotic resistance of individual genotypes could be displayed on the map. Importantly, these maps can be displayed both on the project website and, by using the filtering options provided by EpiCollect, the same data can be requested and viewed as Google Maps on the mobile phone.

Community data collection, citizen scientists and further uses of EpiCollect

As the name implies, EpiCollect has been designed to allow the collection of epidemiological data. However, there are many other situations where this kind of two-way communication between mobile phones and a central database could be utilised; for example, biodiversity research, where multiple researchers record details (and photograph) the presence and distribution of a particular species. This approach is exemplified by the Zoological Society of London's EDGE of existence programme which is the only global initiative to focus specifically on threatened species (http://www.edgeofexistence.org).

Increasingly, the potential for 'citizen scientists' to collect data for community projects within the biological sciences is being exploited. For example, The Open Air Laboratories (OPAL) network (http://www.opalexplorenature.org) encourages anyone with an interest in nature to explore, study, enjoy and protect their local environment,

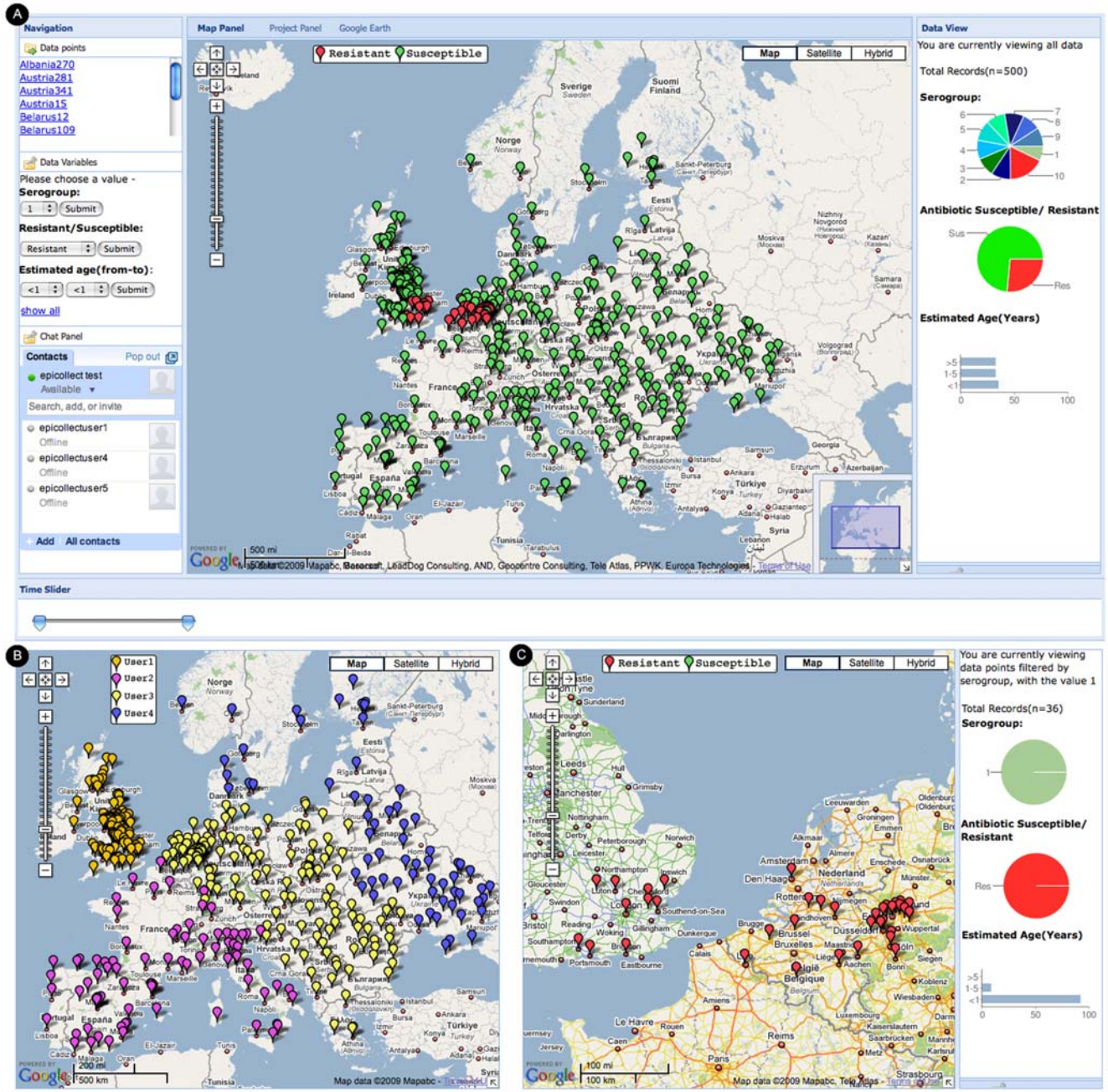


Figure 4. Illustrative dataset showing use of EpiCollect by multiple users. See text for details.
doi:10.1371/journal.pone.0006968.g004

by undertaking animal and plant surveys (such as earthworms found in gardens) with subsequent web-based data submission. Further examples are provided by FrogLife, which encourages the public submission of back-garden amphibian distributions to aid studies in species decline (http://www.froglife.org), the Evolution MegaLab, which encourages the recording of banding patterns of snails so that members of the public can see 'evolution in their own back-yards' (http://www.evolutionmegalab.org) and the RSPB's annual 'Big Garden Birdwatch' which requests members of the public to record all species of birdlife seen in back-gardens or parks and submit the data for central collation (http://www.rspb.org.uk/birdwatch). These initiatives involve people recording data in notebooks and then either faxing the data or entering them at the website used for the community project. The use of EpiCollect would allow individuals to directly submit their data to the project database from their mobile phone and could significantly increase the collection and collation of data for these types of community projects, especially amongst the young where the mobile phone is the natural way of transferring information.

We are currently exploring the use of EpiCollect for data submission to molecular epidemiological databases (http://www.mlst.net) and for submissions by field workers to a central database being used to catalogue the world-wide prevalence of the disease caused by the emerging amphibian fungal pathogen, *Batrachochytrium dendrobatidis*, for which we have developed the web application (http://www.spatialepidemiology.net/bd-maps).

Potential Issues and Further Developments

EpiCollect has been designed for projects that range from epidemiology and ecology to community data collection in developed and middle income countries, but a major potential use is for field surveys in resource-limited settings that collect key information for economic and public health decisions and resource allocation. There are a number of issues centred on cost and mobile data network availability that currently could limit the use of systems such as EpiCollect in these settings – for example in parts of sub-Saharan Africa. GPS satellites, in theory, cover the entire globe and obtaining map co-ordinates for data points entered into a phone using EpiCollect should be possible even in remote areas of resource-limited countries. However, the submission of data from EpiCollect to a central database relies on connectivity with mobile data networks. Mobile phone network coverage in Africa is good in some countries but mobile data networks are much less widely available. The global coverage of such networks can be easily accessed (http://www.gsmworld.com/) and is likely to increase considerably in the next few years. Importantly, data collection and storage within EpiCollect is not reliant on data network connectivity, as data records are stored in the phone's database. Data stored in the phone's database while in remote areas can then be synchronised with the central database after returning to an area that does have data connectivity. However, the major limiting factor for current use of EpiCollect and similar applications in many resource-limited countries is adequate access to data networks and the availability and cost of suitable smartphones. Further progress in the penetration of mobile phone technology may remove these limitations in such areas.

A further general issue is maintaining battery power for mobile phones in remote areas but this can now easily be solved by using solar-powered mobile phone chargers, which are now widely available at low cost.

We chose to develop EpiCollect for Android due to its open-source nature and the provision of suitable multiple phones by different network providers (e.g. T-Mobile, Vodafone and Orange). However, we envision that, similarly to Desktop software that is provided for Windows, Macintosh and Linux, versions will eventually be provided for other smartphone operating systems such as Windows mobile and Palm webOS and an iPhone version of EpiCollect will soon be available.

The generic nature of EpiCollect allows it to be used to submit data to any online database, not just those we host at www.spatialepidemiology.net. We provide EpiCollect as free software and will be further developing the tool to allow on-line project definition of variables for the immediate download of versions of the software tailored to specific needs. We encourage those interested in using EpiCollect within their own projects to contact us, so that we can provide software (and an associated website if required) adapted for use in their projects.

Materials and Methods

EpiCollect was developed using the Android Operating System SDK. Testing was undertaken using both the developer and commercially available T-Mobile G1 phone running both Android version 1.0 and 1.5. For the project-specific websites located within www.spatialepidemiology.net, interface development utilised the ExtJS, Prototype, Script.aculo.us and Mapstraction JavaScript libraries. We used the Google Maps API for 2D mapping, the KML specification when producing output for Google Earth and the Google Charts API for rendering graphical summaries. Server-side scripting was written in PHP and all data were stored in MS SQL Server.

Author Contributions

Conceived and designed the experiments: DMA DMH BGS. Wrote the paper: DMA BGS. Software and web interface development: DMA DMH. Field testing of software: EJF FaO.

References

1. Maiden CJ, Bygraves JA, Feil E, Morelli G, Russell JE, et al. (1998) Multilocus sequence typing: A portable approach to the identification of clones within populations of pathogenic microorganisms. Proc Natl Acad Sci U S A 95: 3140–3145.
2. Aanensen DM, Spratt BG (2005) The multilocus sequence typing network:mlst.net Nuc Acids Res 33: W728–33.
3. Kwok R (2009) Personal technology: Phoning in data. Nature 458(7241): 959–61.
4. Morris K (2009) Mobile phones connecting efforts to tackle infectious disease. Lancet Infect Dis 9(5): 274.
5. Butler D (2006) Mashups mix data into global service. Nature 439(7072): 6–7.