MICROBIAL GENOMICS

RESEARCH ARTICLE

Abudahab *et al.*, *Microbial Genomics* 2018;4

DOI 10.1099/mgen.0.000220

MICROBIOLOGY SOCIETY

OPEN DATA OPEN MICROBIOLOGY

PANINI: Pangenome Neighbour Identification for Bacterial Populations

Khalil Abudahab,[1] Joaquín M. Prada,[2] Zhirong Yang,[3] Stephen D. Bentley,[4] Nicholas J. Croucher,[5] Jukka Corander[3,6,*†] and David M. Aanensen[1,7,*†]

Abstract

The standard workhorse for genomic analysis of the evolution of bacterial populations is phylogenetic modelling of mutations in the core genome. However, a notable amount of information about evolutionary and transmission processes in diverse populations can be lost unless the accessory genome is also taken into consideration. Here, we introduce PANINI (Pangenome Neighbour Identification for Bacterial Populations), a computationally scalable method for identifying the neighbours for each isolate in a data set using unsupervised machine learning with stochastic neighbour embedding based on the t-SNE (t-distributed stochastic neighbour embedding) algorithm. PANINI is browser-based and integrates with the Microreact platform for rapid online visualization and exploration of both core and accessory genome evolutionary signals, together with relevant epidemiological, geographical, temporal and other metadata. Several case studies with single- and multi-clone pneumococcal populations are presented to demonstrate the ability to identify biologically important signals from gene content data. PANINI is available at http://panini.pathogen.watch and code at http://gitlab.com/cgps/panini.

DATA SUMMARY

1. PANINI is accessible at http://panini.pathogen.watch.

2. All example data utilized within the manuscript are available at https://gitlab.com/cgps/panini/datasets.

3. Code for the PANINI web application is available at http://gitlab.com/cgps/panini.

4. A video walkthrough is available at https://vimeo.com/230416235.

INTRODUCTION

In less than a decade, bacterial population genomics has progressed from the sequencing of dozens to thousands of strains [1–4]. The biological insights enabled by population genomics are particularly important in evolutionary epidemiology, as the genome sequences provide high-resolution data for the estimation of transmission and evolutionary dynamics, including the horizontal transfer of virulence and resistance elements. Phylogenetic trees are the main tool utilized for visualization and exploration of population genomic data, both in terms of the level of relatedness of strains and for mapping relevant metadata such as geographical locations and host characteristics [5]. While trees are very useful, they are in general estimated using only core-genome variation (i.e. those regions of the genome common to all members of a sample), which may represent only a fraction of the relevant differences present in genomes across the study population. Several recent studies highlight the importance of considering variation in gene content when investigating the ecological and evolutionary processes leading to the observed data [6, 7].

The rapidly increasing size of population genomic datasets calls for efficient visualization methods to explore patterns of relatedness based on core-genomic polymorphisms, accessory gene content, epidemiological, geographical and other metadata. Here, we introduce a framework that integrates within the web application Microreact [5], by

Received 6 April 2018; Accepted 26 August 2018

Author affiliations: [1]Centre for Genomic Pathogen Surveillance, Wellcome Genome Campus, Hinxton, UK; [2]School of Veterinary Medicine, University of Surrey, Guildford, UK; [3]Department of Mathematics and Statistics, Helsinki Institute of Information Technology, University of Helsinki, FI-00014 Helsinki, Finland; [4]Pathogen Genomics, Wellcome Trust Sanger Institute, Hinxton, UK; [5]Department of Infectious Disease Epidemiology, Imperial College London, London, UK; [6]Department of Biostatistics, Institute of Basic Medical Sciences, University of Oslo, N-0317 Oslo, Norway; [7]Big Data Institute, Li Ka Shing Centre for Health Informatics, University of Oxford, Oxford, UK.

***Correspondence:** Jukka Corander, jukka.corander@medisin.uio.no; David M. Aanensen, david.aanensen@sanger.ac.uk

Keywords: pangenome; microbial population genomics; machine learning; web application.

Abbreviations: 2D, two-dimensional; PANINI, Pangenome Neighbour Identification for Bacterial Populations; PCV7, seven-valent polysaccharide conjugate vaccine; PRCI, phage-related chromosomal island ; t-SNE, t-distributed stochastic neighbour embedding.

†These authors contributed equally to this work.

Data statement: All supporting data, code and protocols have been provided within the article or through supplementary data files.

000220 © 2018 The Authors

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 work is properly cited.

Downloaded from www.microbiologyresearch.org by

IP: 155.198.8.192

On: Mon, 03 Dec 2018 14:31:56

utilizing a popular unsupervised machine-learning technique for big data to infer neighbours of bacterial strains from accessory gene content data and to efficiently visualize the resulting relationships. The machine-learning method, called t-SNE (t-distributed stochastic neighbour embedding), has already gained widespread popularity for exploring image, video and textual data [8, 9], but has to our knowledge not yet been widely utilized for bacterial population genomics.

Since gene content may in general be rapidly altered in bacteria, it provides a high-resolution evolutionary marker of relatedness that can extend far beyond core-genome mutations [7]. Different processes driving horizontal movement of DNA, such as homologous recombination, conjugative transfer of plasmids and phage infections, all affect the gene content within and outside of a chromosome. By contrasting core and non-core gene content, one can investigate and draw conclusions about genome dynamics across a sample collection. Here, we demonstrate the biological utility of such an approach by application to multiple population data sets.

IMPACT STATEMENT

The assessment of similarity in both the core and non-core regions of genomic datasets can shed light on the evolutionary, population and epidemiological dynamics of microbial populations. Common workflows tend to focus on clustering core variation and representing this as a tree, to which other parameters are added to make sense of the data and system under investigation, potentially missing additional information about evolutionary and transmission processes in the pangenome. Increasingly, with ever-growing population scale datasets, the importance and dynamics of the non-core (e.g. movement of phage, plasmids and other mobile elements) needs to be assessed as a matter of routine. We demonstrate the utility of a novel machine-learning method and its ease of use via a web application for visualization. Such methods, enabling the rapid and easy identification of similarity in the non-core, and the subsequent relation of this information to core phylogenies and other epidemiological and relevant system-level metadata, will aid our understanding of the dynamics within the pangenome, and shed further light on, for example, host, niche and pathogenic adaptation.

METHODS AND RESULTS

t-SNE is a machine-learning algorithm that is widely used for data visualization [8]. It is suitable for embedding a set of high-dimensional data items in a two-dimensional (2D) or three-dimensional space. The embedding approximately preserves the pairwise similarities between the data items.

The t-SNE algorithm consists of two main steps. First, it calculates the similarities between the data items in the high-dimensional space, which is typically based on normal distribution around each data item. The similarities are then normalized to be probabilities (i.e. they sum to one). Similarities in the low-dimensional space are analogously defined and normalized except that Student's *t*-distribution replaces the Gaussians. Second, t-SNE minimizes Kullback–Leibler divergence between the two probability matrices over the embedding coordinates. Finally, the 2D t-SNE result can be visualized as a scatter plot where each dot indicates a data item.

t-SNE as an unsupervised method is particularly useful for exploratory data analysis. It has a wide range of applications in music analysis, cancer research, computer security research, bioinformatics and biomedical signal processing. In many cases, t-SNE is able to identify meaningful data structures such as clusters even without feature engineering or structural assumptions, e.g. about number of clusters underlying the data, even in cases where Principal Component Analysis (PCA) has been demonstrated to fail. Here, we use the latest version of the t-SNE projection method, adopting the Barnes–Hut algorithm for accelerating the divergence minimization [9]. To demonstrate utility within population genomics, firstly, we explore how the method performs in a simulated setting, where the relationship between all sequences is known; and then we extend our analysis to published bacterial population data sets, allowing us to uncover previously unseen relationships between data and to address important biological questions.

Simulated data

To validate the methodology, we assessed how well it identified neighbours and clusters for simulated genetic sequences. Firstly, we randomly generated multiple synthetic datasets of related isolates, with each defined as a sequence of present/absent genes. Each dataset was generated using the following parameters. (1) There were 20 clusters as underlying subpopulations. (2) The number of isolates belonging to a cluster was drawn from a Poisson distribution with mean 15. (3) Each cluster was defined by a number of core genes, which ranged uniformly from 1 to 100. (4) Each isolate had a probability between 80 and 99 % of independently carrying each of the core genes of the cluster it belonged to. (5) Conversely, each isolate had a probability (PN) to independently carry each of the non-core genes of its cluster. Non-core genes were composed of core genes of other clusters and 'noise' genes that were not defining characteristics of any cluster (in total 300 genes).

Each generated dataset had on average 300 isolates with a gene content of 1300 genes present/absent on average. For each dataset, we estimated the genetic pairwise Hamming distance (d_H) and the distance using the t-SNE algorithm (d_t). The Hamming distance here was simply the number of differences between two binary sequences, where each element was an indicator for whether a particular gene was present in an isolate or not. The implementation of the t-SNE algorithm that we used yields a coordinate in a 2D

Downloaded from www.microbiologyresearch.org by
IP: 155.198.8.192
On: Mon, 03 Dec 2018 14:31:56

plane for each isolate, and we calculated the distance d_t simply as the Euclidean distance for each pair of isolates.

If a cluster is sufficiently differentiable in terms of its gene content, we expect the Hamming distance within the cluster to be smaller than to any other isolate not belonging to it. For the t-SNE algorithm to be considered valid, it should be able to project the isolates from the same cluster on the 2D plane sufficiently close together so that the Euclidean distance within the cluster is smaller than to any other isolate. Given the conditions that were used to generate the synthetic datasets, not all clusters were necessarily differentiable in terms of their gene content; therefore, we classified the t-SNE algorithm as performing erroneously only when a pair of isolates belonging to a different cluster were not identified as such by the algorithm but were correctly identified using the Hamming distance. For high levels of noise, i.e. a large value of probability (PN), differentiating the clusters using their gene content becomes increasingly difficult as the isolates may lack a sufficiently stable signal of relatedness.

We analysed the performance of the t-SNE algorithm for three levels of noise PN, 0.001, 0.005 and 0.01, which measured the mean proportion of non-core genes in each isolate. We performed 100 repeats for each noise value, which for each repeat involved generating on average 300 sequences and comparing almost 45 000 pairs of isolates. The mean error for the three noise values was 0.5, 1 and 4 %, respectively, with a small error representing a particular isolate mis-allocated (i.e. very close to a different cluster) and a large error representing two clusters that were not appropriately differentiated by the t-SNE algorithm, illustrated in Fig. 1. The error of the t-SNE algorithm increases with the noise, as shown in Fig. 1(iii), and with the total number of clusters (not shown).

Web application – https://panini.pathogen.watch/

The t-SNE algorithm implemented in C++ (https://github.com/lvdmaaten/bhtsne) was wrapped as a Node.js native module and embedded within a web application. The application was written in JavaScript and utilizes React (https://reactjs.org/) for front-end and the Vis.js library (http://visjs.org) for network visualization. (1) Data are uploaded as a gene presence/absence matrix –PANINI (Pangenome Neighbour Identification for Bacterial Populations) expects data in the .RTab format (the output from Roary: the pan genome pipeline [10]; https://sanger-pathogens.github.io/Roary). However, this is simply a data file containing gene rows and isolate columns with '1' or '0' indicating presence/absence of a particular gene for a particular isolate. (2) Genes present in all isolates are ignored (i.e. core genome) and non-core genes are clustered using t-SNE with default parameters (auto perplexity and theta=0.5 – parameters can be changed by users). (3) The results [*x*, *y* coordinates, a '.dot' format file containing graph layout, csv and JSON (JavaScript object notation)] are made available for download and reuse. Results are also visualized directly within the PANINI web application as a graph layout.

To interpret the data in an epidemiological, phylogeographical and geographical context, the estimated network can also be uploaded directly to the Microreact platform allowing a user to add other forms of data to relate to the resulting neighbour embedding, typically a phylogenetic tree, geographical locations of the isolates and temporal data (further information and instructions are available at https://microreact.org).

Utility with existing published datasets

To demonstrate the utility of t-SNE clustering, we applied the method to four published datasets that used whole-genome sequencing to study the evolution of the bacterium *Streptococcus pneumoniae*. The first, a population-level dataset, detailed population-wide diversity of pneumococci within Massachusetts, USA, pre- and post-vaccine introduction [2], while the second and third detailed international collections of globally disseminated multidrug-resistant lineages of *Streptococcus pneumoniae* [11, 12]. The fourth data set comprised 115 *Salmonella enterica* serovar Weltevreden isolates mostly from the tropics, representing an emerging agent of diarrheal disease [13]. Additional biological insights made possible with PANINI are described, and links to the projects within Microreact for further exploration of the associated metadata and download of raw data formats are provided.

Analysis of a diverse pneumococcal population

Data visualization and download

Data are available at: https://microreact.org/project/panini-sparc?ui=nt.

Source data and .RTab file

Source data and the .RTab file are available at: https://gitlab.com/cgps/panini/datasets/tree/master/SPARC.

Video walkthrough for PANINI and Microreact creation/use

A walkthrough video for PANINI is available at: https://vimeo.com/230416235.

Pneumococcal population analysis

PANINI was applied to a collection of 616 systematically sampled pneumococcal isolates from a vaccine and antimicrobial-resistance surveillance project in Massachusetts, USA [14]. The original analysis of the gene content in this collection identified 5442 'clusters of orthologous genes' (COGs) [2], the core set of which was used to define 15 'sequence clusters' with BAPS (http://www.helsinki.fi/bsg/software/BAPS) [15]. For most of the sequence clusters, the correspondence between a group in the PANINI output and the original sequence clusters was exact (Fig. 2a), reflecting their similarity both in terms of the core and accessory genomes [16]. These sets of isolates, therefore, represent well-defined distinct lineages. However, SC1, SC6, SC10 and SC12 all exhibited distinct substructuring in the PANINI output. This corresponded well with the diverse core genome observed in these clusters (Fig. 2b), and in each case, these groups were consistent with clades within the sequence clusters. These sequence clusters are, therefore, likely to represent

Downloaded from www.microbiologyresearch.org by
IP: 155.198.8.192
On: Mon, 03 Dec 2018 14:31:56

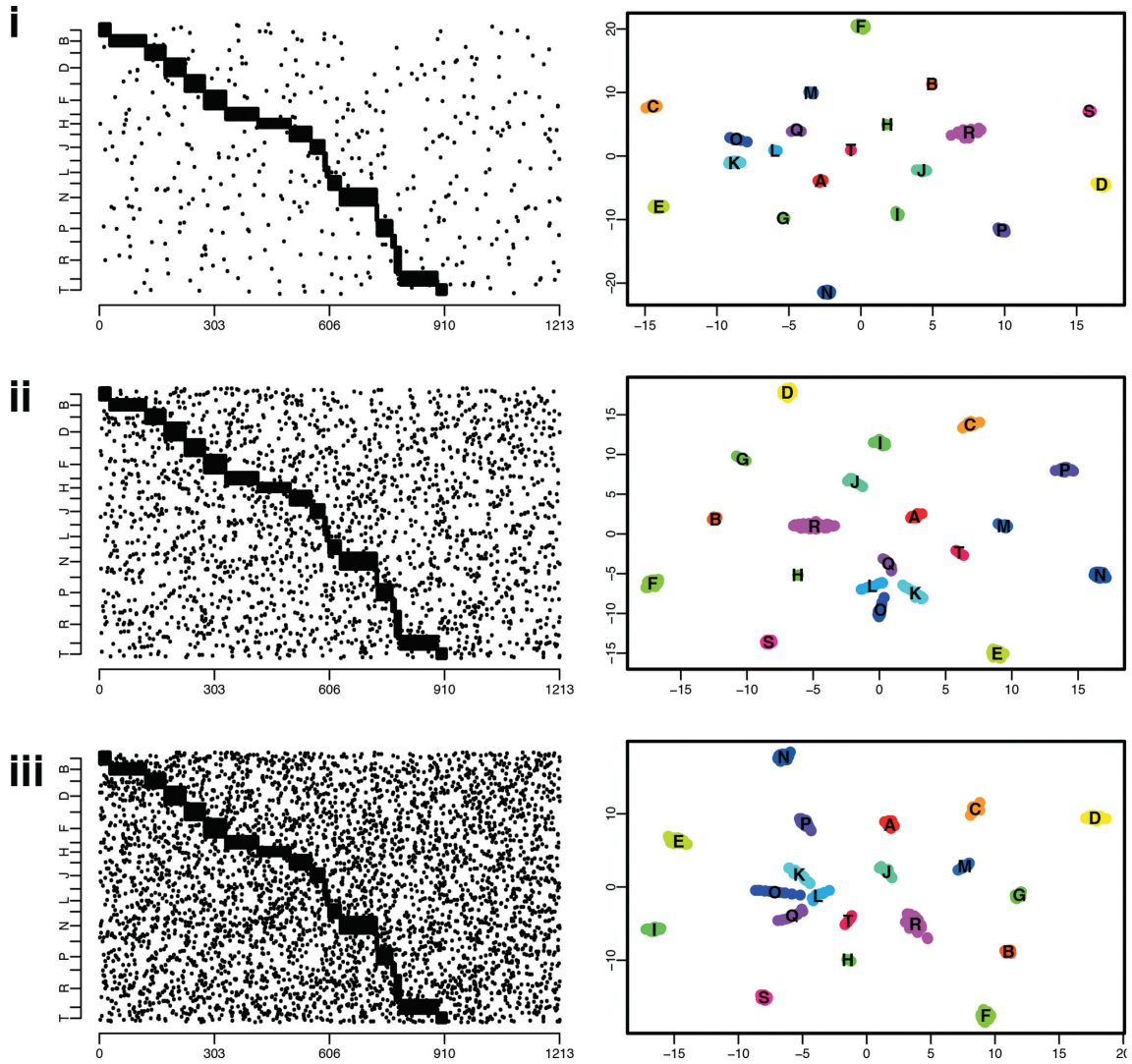


Fig. 1. Illustration of a simulated dataset, with the isolates' gene content (left), black dots indicate the presence of a gene, the *x*-axis represents all the considered genes (a total of 1213 genes in this simulation). The right panels show the embedded locations in the 2D plane as estimated by the t-SNE algorithm, with each colour representing a cluster in the underlying simulation model. Clusters are named using the alphabet (A, B, C…). From top to bottom, plots indicate simulations generated with 0.1 % (i), 0.5 % (ii) and 1 % (iii) noise, respectively.

amalgams of genotypes that should be subdivided into multiple clusters. Conversely, PANINI revealed clear substructuring within the previously unclustered SC16, which was also consistent with the core-genome phylogeny. Hence, PANINI can easily facilitate the division of a diverse population into discrete genotypes that are coherent in their accessory- and core-genome content.

Extensive prophage variation in a multidrug-resistant lineage

Data visualization and download

Data are available at: https://microreact.org/project/panini-pmen2?ui=nt.

Source data and .RTab file

Source data and the .RTab file are available at: https://gitlab.com/cgps/panini/datasets/tree/master/PMEN2.

Prophage variation

PANINI was applied to an analysis of orthologous genes across a global collection of 190 isolates from the multidrug-resistant *Streptococcus pneumoniae* clone PMEN2 [11], which caused a large outbreak of disease in Iceland starting in the late 1980s (Fig. 3a). Multiple distinct clusters were again evident in the output (Fig. 3b). In some cases, these were consistent with the phylogeny. The original analysis identified two independent entries of the lineage into Iceland, clades IC1 and IC2, the latter of which contained

Downloaded from www.microbiologyresearch.org by
IP: 155.198.8.192
On: Mon, 03 Dec 2018 14:31:56

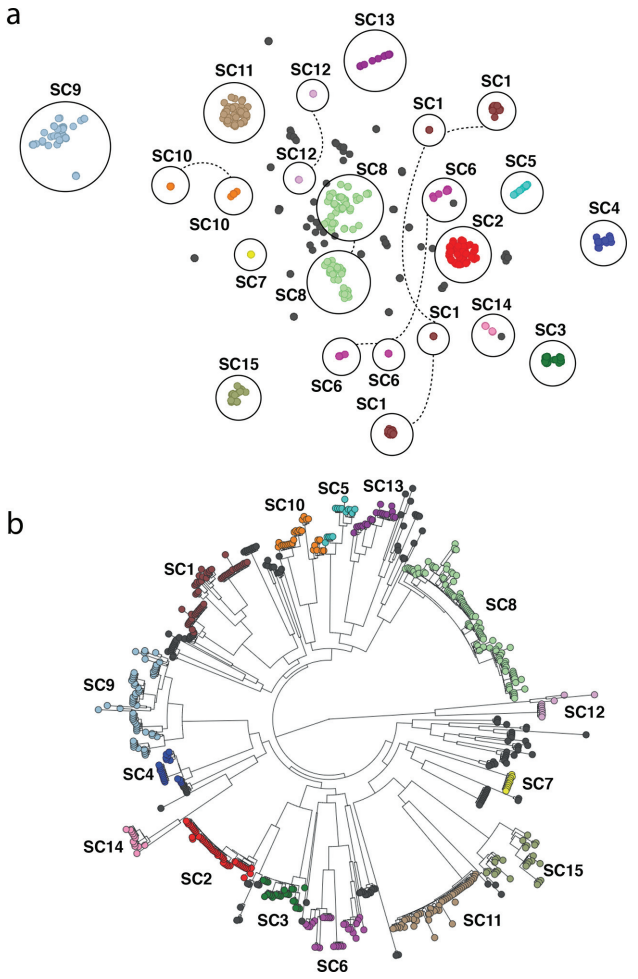


Fig. 2. (a) Annotated output of the PANINI algorithm applied to 616 *Streptococcus pneumoniae* isolates from a diverse population in Massachusetts, USA. Each node represents an isolate, each of which is coloured according to its sequence cluster, as defined using the core genome. Clusters of isolates belonging to the same sequence cluster are circled and annotated. Where sequence clusters are divided into multiple groups in the PANINI network, the circles are joined by dashed lines. (b) Core-genome phylogeny based on comparison of conserved clusters of orthologous genes (COGs) adapted from [2] and displayed within Microreact. Sequence clusters are annotated for comparison with non-core clustering.

many fewer isolates and was clustered as IcA in the annotated output. By contrast, IC1 was distributed across four clusters IcB–IcE, which did not correspond with clear clades in the phylogeny. The difference between IcB and IcC is technical, rather than biological: all IcB isolates were sequenced early in the project with 54 nt reads, whereas most IcC isolates were sequenced with 75 nt reads. Unusually for pneumococci, the isolates in both these groups were trilysogenic, carrying prophage similar to ϕ670-6B.1 and ϕ670-6B.2, found in the *Streptococcus pneumoniae* 670-6B genome inserted between *dnaN* and *pth* (att_{670}), and within the *comYC* gene (att_{comYC}), respectively; and a prophage isolated from 0211+13275, inserted at SPN23F15280 – SPN23F15810 (att_{MM1}) [16]. The apparent rapid acquisition, and stable maintenance, of multiple viral loci may relate to the abrogation of these bacteria's competence system by the insertion of prophage ϕIC1 into *comYC* [11, 17]. Group IcD, interspersed with IcB and IcC within clade IC1 in the phylogeny, differs in the absence of prophage similar to ϕ670-6B.2. IcE, also polyphyletic within clade IC1, differed in having lost the region of pneumococcal pathogenicity island 1 (PPI-1) that encodes the *pia* iron-transport operon, which plays a role in pneumococcal pathogenesis in animal models [18]. Hence, it is not surprising to find these isolates were only recovered from sputum, otitis media samples or nasopharyngeal swabs.

Multiple distinct clusters of non-Icelandic isolates were also observed. These all represented cases where t-SNE grouped isolates that were disparate in terms of their country and year of isolation, as well as having a polyphyletic distribution across the whole genome phylogeny. These groupings represented cases of convergent evolution through parallel acquisition very similar prophage. Group IntA lacked any prophage similar to those shown in Fig. 3(c); group IntB had prophage with some similarity to both prophage in the reference genome; group IntC only had a prophage with similarity to ϕ0211+13275; whereas group IntD had prophage similar to ϕ0211+13 275 and ϕ670-6B.1 as well. Hence, the rapid movement of prophage sequences within lineages [16] clearly substantially contributes to the changes in gene content observed over short timescales. PANINI facilitates rapid analysis of these diverse elements, and their complex relationship with bacterial population structure.

Mobile element and serotype variation in a vaccine-escape lineage

Data visualization and download

Data are available at: https://microreact.org/project/panini-pmen14?ui=nt.

Source data and .RTab file

Source data and the .RTab file are available at: https://gitlab.com/cgps/panini/datasets/tree/master/PMEN14.

Mobile element and serotype variation

PANINI was similarly applied to 176 isolates of the multidrug-resistant *Streptococcus pneumoniae* PMEN14 lineage [11]. Although the sequences came from many countries, the collection was strongly enriched for bacteria from the Maela refugee camp in Thailand [11], which fell into five clades (ML1–ML5), of which ML2 was the largest. The groups identified by PANINI were again polyphyletic (Fig. 4a), with ML2 split up in a similar manner to the PMEN2 clade IC1. This was again driven by the distribution of prophage sequence: group 1 isolates were free of prophage, whereas group 2 isolates were infected with a 'group 2-type' prophage, and group 3 isolates were infected with a similar, but distinct, 'group 3-type' prophage (Fig. 4b). Clade ML2 isolates in group 4 were distinguished by variation in another mobile genetic element, a phage-related

Downloaded from www.microbiologyresearch.org by
IP: 155.198.8.192
On: Mon, 03 Dec 2018 14:31:56

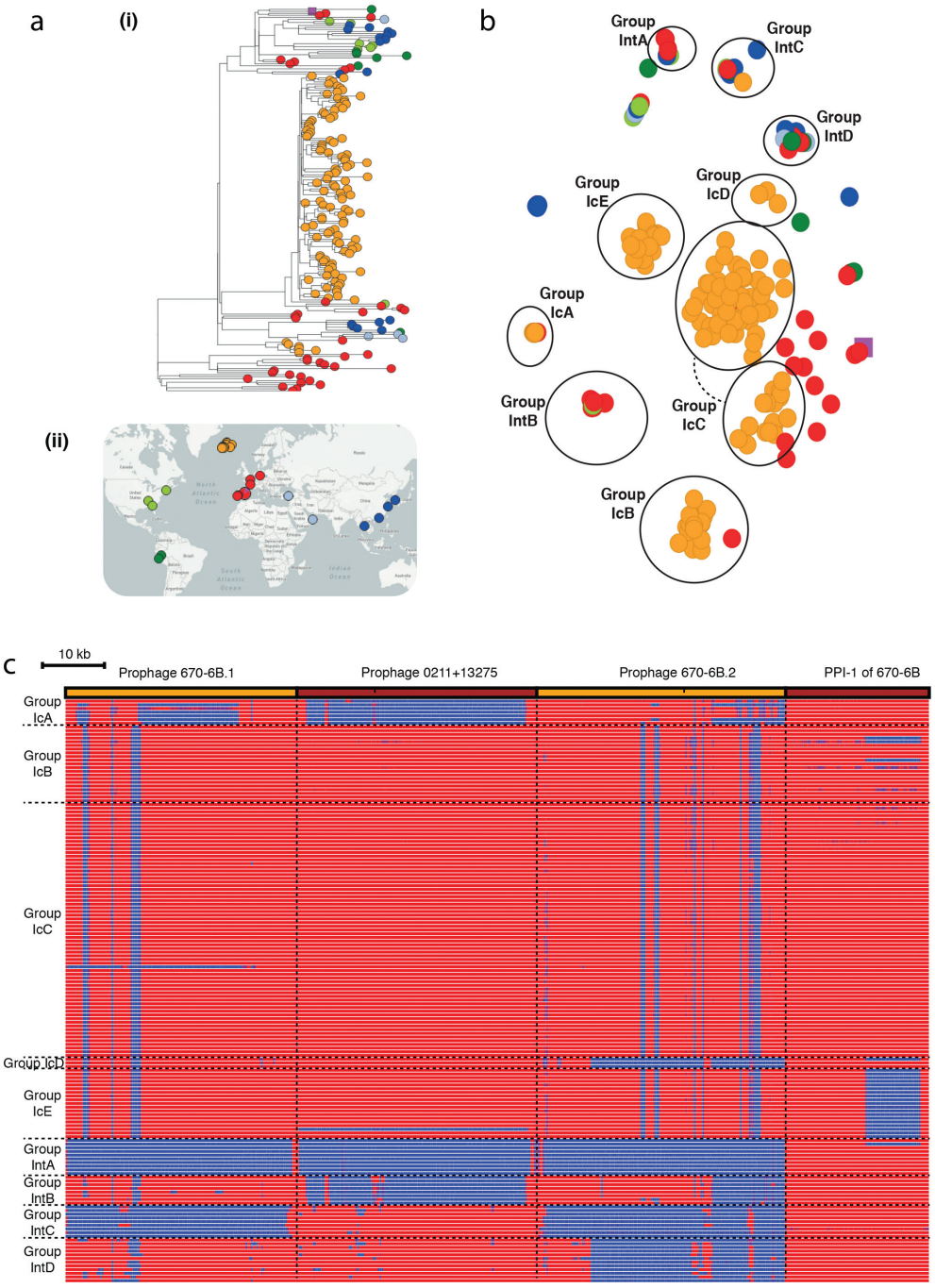


Fig. 3. Analysis of the *Streptococcus pneumoniae* PMEN2 lineage. (a) (i) Core-genome phylogeny with tree leaves coloured by country of origin and (ii) geographical origin of isolates. (b) Annotated output of the PANINI algorithm applied to 189 isolates from an international collection of representatives of the *Streptococcus pneumoniae* PMEN2 lineage. Each point is coloured according to its region of origin. Groups defined by the structure of the PANINI output are circled and annotated. Clusters containing primarily Icelandic isolates (coloured orange) are labelled with 'Ic' prefixes, whereas those containing isolates from multiple countries are labelled with 'Int' prefixes. (c) Variation in accessory loci associated with differential classification of isolates into groups. The orange and brown bands across the top of the figure indicate the extent of the three prophage and pneumococcal pathogenicity island 1 (PPI-1) sequences, against which the short-read data from the isolates were mapped. The heatmap below includes one row per isolate, which were ordered according to their grouping in (a). The heatmap is coloured blue where mapping coverage was low, indicating a locus is absent, and red were mapping coverage was high, indicating a sequence was present. Horizontal dashed lines indicate the boundaries between the groups of isolates, vertical dashed lines indicate the boundaries between loci.

Downloaded from www.microbiologyresearch.org by
IP: 155.198.8.192
On: Mon, 03 Dec 2018 14:31:56

chromosomal island (PRCI), shared by most of the isolates. This PRCI was absent from these assemblies, either because at least part of the element had been lost through deletion, due to replacement with a related sequence (isolate 6259_1-15) or due to the acquisition of a second, highly similar PRCI that prevented effective assembly of either (isolates 6237_8-12, 6237_8-13 and 6237_8-18). In this latter case, mapping to the element was still evident.

A fifth group, which did not include any Maela isolates, corresponded to the antibiotic-susceptible outgroup isolates. These differed through the absence of a third type of mobile element, the Tn*916* integrative and conjugative element, an antibiotic resistance-encoding genomic island that was absent from these 'outgroup' isolates. Additionally, these bacteria shared two smaller genomic islands, encoding putative lantibiotic biosynthesis and restriction-modification operons, which were absent from the multidrug-resistant isolates. Variation in other non-mobile element islands was also detectable. The group 1-19A subcluster contained isolates of serotype 19A, produced through two independent serotype switching recombinations at the capsule polysaccharide synthesis (*cps*) locus that resulted in genotypes '19A ST320' and '19A ST236'. These changes were responsible for allowing isolates to evade the seven-valent polysaccharide conjugate vaccine (PCV7), which targeted the lineage's ancestral serotype 19F, expressed by almost all the rest of the collection [12]. A smaller serotype switching recombination, which did not replace the entire serotype-determining *cps* locus, generated the '19A ST271' isolates [12]. The smaller associated change in gene content meant this isolate was not clearly distinguished from the rest of group 1 (Fig. 4a).

Phylogeographical structure of *Salmonella enterica* serovar Weltevreden

Data visualization and download

Data are available at: https://microreact.org/project/panini-salmonella.

Source data and .RTab file

Source data and the .RTab file are available at: https://gitlab.com/cgps/panini/datasets/tree/master/Salmonella.

Phylogeographical structure

Makendi *et al.* [13] notified that the *Salmonella enterica* isolate collection harboured substantial variation in gene content, such that in total 7923 putative coding sequences (CDSs) were detected in the accessory genomes of the 115 isolates. The authors noted that each isolate had numerous prophage elements and that *Salmonella enterica* serovar Weltevreden appeared to undergo rapid gains and losses of genetic material. PANINI analysis shows that accessory clusters follow largely the clade structure of the core-genome tree (Fig. 5). However, there are some notable counter examples where isolates clustering closely together in the PANINI output are very distant from each other in the core-genome tree (Fig. 5). Conversely, there are also examples of isolates that are neighbours in the core-genome tree, but cluster clearly separately in the PANINI network. The interactive features of Microreact enable a rapid exploration of such cases (for example timeline and interactive zoom/select), which can then be followed by a more thorough analysis of the gene content differences responsible for the detected discrepancies between the two types of genetic relatedness.

DISCUSSION

The rapid increase in sampling density of bacterial populations for epidemiological and evolutionary studies highlights the need of combining traditional genomic markers, such as single nucleotide polymorphism (SNP) loci and small insertions or deletions in coding regions, with measures of difference in terms of gene content. As many bacteria have varied accessory genomes, changes in the gene content can offer a way to identify epidemiologically or evolutionarily important clues about the evolutionary processes affecting a pathogen's spread. As we have illustrated here, such information is most useful when clustering is combined within a phylogeographical approach, and visualized jointly in a seamless fashion enabling the rapid interpretation of core and non-core clustering in the context of where and when data were collected.

The t-SNE algorithm is a very efficient approach to cluster isolates based on their gene content. In the simulated scenarios considering synthetic data, the errors in clustering always remained small, either representing an isolate allocated to a wrong cluster or two clusters that were merged. However, this only occurred in simulations with the 'noise' level much higher than expected in nature. In general, what we defined as 'core' genes in a cluster rarely appear in isolates not belonging to the cluster, and if they do, it is typically at much lower frequencies than those we considered. Furthermore, in our synthetic datasets, we formed clusters defined by as few as a single core gene. These clusters with a limited number of core genes, combined with relatively high levels of 'noise', are in practice almost completely indistinguishable from others, as illustrated in Fig. 1 (iii – clusters K, L, O and Q). Overall, our simulated datasets are conservative, as the gene absence and presence variation is higher than expected in natural populations, and therefore indicate that the t-SNE is a promising approach for rapidly and accurately clustering bacteria based on gene content. Nevertheless, it is important to be aware that issues with gene calling may in some cases influence the accuracy of PANINI due to for example different alleles of a gene being assigned as separate genes in a Roary analysis. Consequently, unless using expert-curated pangenomes, we advise a user to test multiple Roary thresholds to see whether the PANINI results with the corresponding different Roary outputs are robust with respect to small changes in the threshold.

When applied to a population-wide genomic dataset, the algorithm was clearly able to identify distinct lineages within a diverse collection. This analysis could highlight which clusters, defined using the core genome, could be

Downloaded from www.microbiologyresearch.org by
IP: 155.198.8.192
On: Mon, 03 Dec 2018 14:31:56

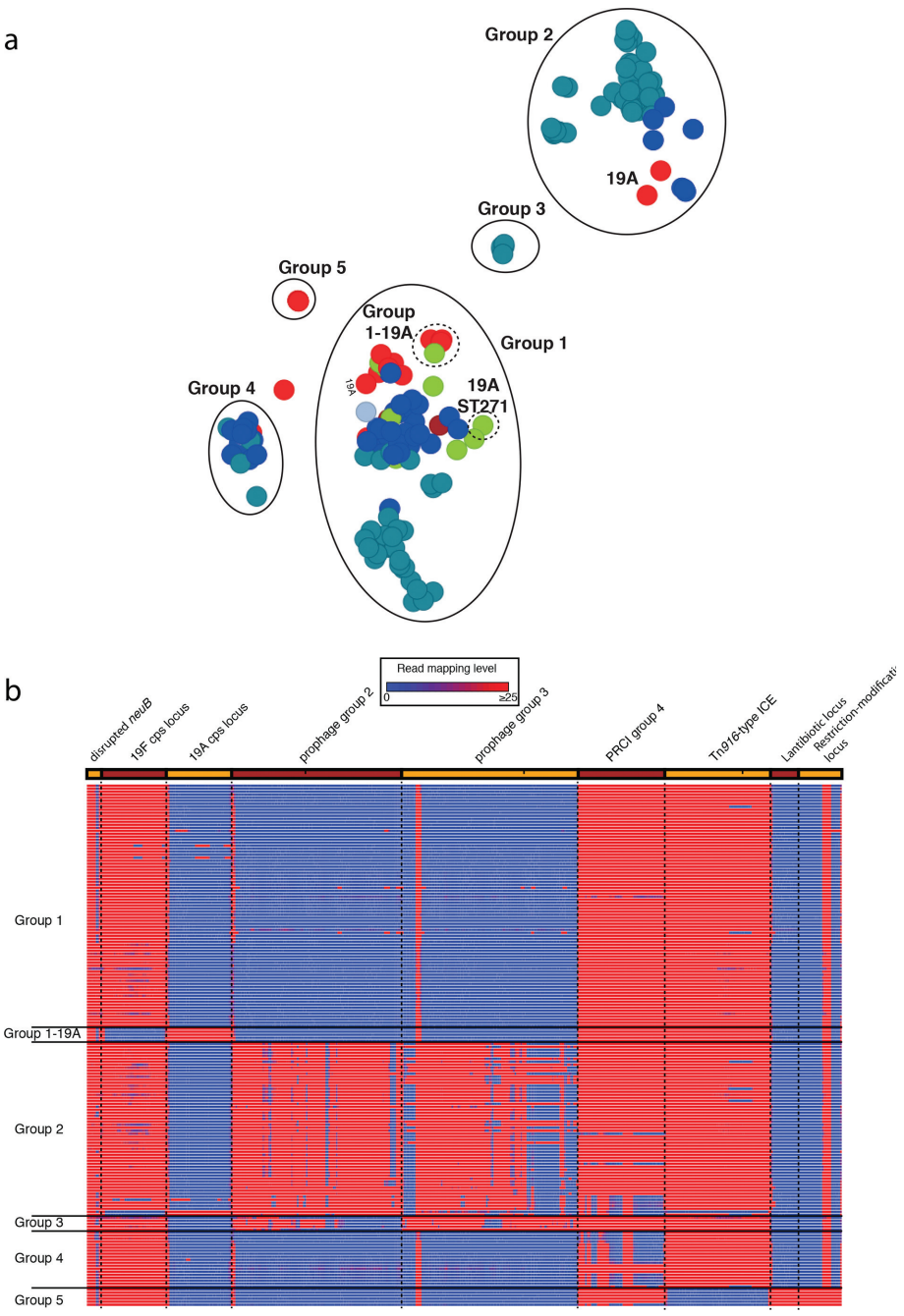


Fig. 4. Analysis of the *Streptococcus pneumoniae* PMEN14 lineage. (a) Annotated output of the PANINI algorithm applied to 176 isolates from an international collection of representatives of the *Streptococcus pneumoniae* PMEN14 lineage. The main groups 1–5 are circled with solid lines and named; the subgroups within group 1 are circled by dashed lines. (b) Variation in accessory loci associated with differential classification of isolates into groups. This heatmap is displayed as in Fig. 3. In this case, the sequence loci across the top are more functionally diverse. The first is the *neuB* coding sequence with an IS*Spn*8 element inserted into it. The lack of mapping to the middle of this column indicates the absence of this insertion sequence anywhere in the chromosome. The next loci are alternative alleles of the capsule polysaccharide synthesis locus, one encoding for the biosynthesis of the PCV7 type 19F polysaccharide, the other for the non-PCV7 type 19A polysaccharide. These are followed by two similar prophage, one associated with group 2 isolates, the other with group 3 isolates; the similarity between these two viruses means there is extensive mapping to both, even when an isolate only contains one of them. The PRCI absent from the assemblies of group 4 isolates is next; mapping suggests this is actually present in some, but PANINI nevertheless included them in this group because the acquisition of a further, related PRCI prevented either assembling accurately. This is followed by the Tn*916* conjugative element, absent from the group 5 isolates, which possess genomic islands encoding for the biosynthesis of a lantibiotic and a restriction-modification system, included at the right-hand end of the panel.

Downloaded from www.microbiologyresearch.org by
IP: 155.198.8.192
On: Mon, 03 Dec 2018 14:31:56



Fig. 5. Analysis of the *Salmonella enterica* serovar Weltevreden as displayed within Microreact (https://microreact.org/project/panini-salmonella). (a) Core-genome tree of 115 *Salmonella enterica* serovar Weltevreden isolates, colour coded by the country of isolation. (b) Output of the PANINI algorithm with isolates colour coded similar to (a). (c) Timeline indicating date of sampling to aid interpretation and interactivity.

sensibly subdivided, and which small groups within a diverse set of strains could be justifiably regarded as new clusters. Within lineages, the same congruence between core and accessory genomes across clades was not observed. Instead, clusters were distinguished by rapidly occurring, homoplasic alterations, such as phage infection. In this context, PANINI provides an intuitive way in which to understand the distribution of rapidly evolving aspects of the genome, which are difficult to analyse in a conventional phylogenetic framework. PANINI is, therefore, a promising platform through which biologically important changes in bacterial gene content can be uncovered at all levels of evolutionary, ecological and epidemiological analyses. To quantify properties of the inferred neighbourhood structure as a function of different underlying biological processes in closer detail by simulation will be an interesting topic for future research.

Funding information
The development of PANINI was funded by the Centre for Genomic Pathogen Surveillance, Wellcome Trust grant no. 099202 and MRC grant no. MR/N019296/1. J. C. was supported by ERC grant no. 742158. Z. Y. was supported by the COIN Centre of Excellence. J. M. P. gratefully acknowledges funding of the NTD Modelling Consortium by the Bill & Melinda Gates Foundation, in partnership with the Task Force for Global Health. The views, opinions, assumptions or any other information set out in this report should not be attributed to the Bill & Melinda Gates Foundation, nor the Task Force for Global Health, nor any person connected with them. N. J. C. is funded by a Sir Henry Dale Fellowship, jointly funded by the Wellcome Trust and the Royal Society (grant no. 104169/Z/14/Z). D. M. A and K. A. are supported by The NIHR Global Health Research Unit on Genomic Surveillance of Antimicrobial Resistance.

Conflicts of interest
The authors declare that there are no conflicts of interest.

References

1. Harris SR, Feil EJ, Holden MT, Quail MA, Nickerson EK *et al.* Evolution of MRSA during hospital transmission and intercontinental spread. *Science* 2010;327:469–474.
2. Croucher NJ, Finkelstein JA, Pelton SI, Mitchell PK, Lee GM *et al.* Population genomics of post-vaccine changes in pneumococcal epidemiology. *Nat Genet* 2013;45:656–663.
3. Chewapreecha C, Harris SR, Croucher NJ, Turner C, Marttinen P *et al.* Dense genomic sampling identifies highways of pneumococcal recombination. *Nat Genet* 2014;46:305–309.
4. Aanensen DM, Feil EJ, Holden MT, Dordel J, Yeats CA *et al.* Whole-genome sequencing for routine pathogen surveillance in public health: a population snapshot of invasive *Staphylococcus aureus* in Europe. *MBio* 2016;7:e00444-16.
5. Argimón S, Abudahab K, Goater RJ, Fedosejev A, Bhai J *et al.* Microreact: visualizing and sharing data for genomic epidemiology and phylogeography. *Microb Genom* 2016;2:e000093.
6. Marttinen P, Croucher NJ, Gutmann MU, Corander J, Hanage WP. Recombination produces coherent bacterial species clusters in both core and accessory genomes. *Microb Genom* 2015;1:e000038.
7. McNally A, Oren Y, Kelly D, Pascoe B, Dunn S *et al.* Combined analysis of variation in core, accessory and regulatory genome

Downloaded from www.microbiologyresearch.org by
IP: 155.198.8.192
On: Mon, 03 Dec 2018 14:31:56

regions provides a super-resolution view into the evolution of bacterial populations. *PLoS Genet* 2016;12:e1006280.

8. **van der Maaten L, Hinton G.** Visualizing data using t-SNE. *JMLR* 2008:2579–2605.
9. **van der Maaten L.** Accelerating t-SNE using tree-based algorithms. *JMLR* 2014:3221–3245.
10. **Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S *et al.*** Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* 2015;31:3691–3693.
11. **Croucher NJ, Hanage WP, Harris SR, McGee L, van der Linden M *et al.*** Variable recombination dynamics during the emergence, transmission and 'disarming' of a multidrug-resistant pneumococcal clone. *BMC Biol* 2014;12:49.
12. **Croucher NJ, Chewapreecha C, Hanage WP, Harris SR, McGee L *et al.*** Evidence for soft selective sweeps in the evolution of pneumococcal multidrug resistance and vaccine escape. *Genome Biol Evol* 2014;6:1589–1602.
13. **Makendi C, Page AJ, Wren BW, Le Thi Phuong T, Clare S *et al.*** A phylogenetic and phenotypic analysis of *Salmonella enterica* serovar Weltevreden, an emerging agent of diarrheal disease in tropical regions. *PLoS Negl Trop Dis* 2016;10:e0004446.
14. **Finkelstein JA, Huang SS, Daniel J, Rifas-Shiman SL, Kleinman K *et al.*** Antibiotic-resistant *Streptococcus pneumoniae* in the heptavalent pneumococcal conjugate vaccine era: predictors of carriage in a multicommunity sample. *Pediatrics* 2003; 112:862–869.
15. **Corander J, Marttinen P, Sirén J, Tang J.** Enhanced Bayesian modelling in BAPS software for learning genetic structures of populations. *BMC Bioinformatics* 2008;9:539.
16. **Croucher NJ, Coupland PG, Stevenson AE, Callendrello A, Bentley SD *et al.*** Diversification of bacterial genome content through distinct mechanisms over different timescales. *Nat Commun* 2014;5: 5471.
17. **Croucher NJ, Mostowy R, Wymant C, Turner P, Bentley SD *et al.*** Horizontal DNA transfer mechanisms of bacteria as weapons of intragenomic conflict. *PLoS Biol* 2016;14:e1002394.
18. **Brown JS, Gilliland SM, Ruiz-Albert J, Holden DW.** Characterization of pit, a *Streptococcus pneumoniae* iron uptake ABC transporter. *Infect Immun* 2002;70:4389–4398.

Five reasons to publish your next article with a Microbiology Society journal

1. The Microbiology Society is a not-for-profit organization.
2. We offer fast and rigorous peer review – average time to first decision is 4–6 weeks.
3. Our journals have a global readership with subscriptions held in research institutions around the world.
4. 80% of our authors rate our submission process as 'excellent' or 'very good'.
5. Your article will be published on an interactive journal platform with advanced metrics.

Find out more and submit your article at microbiologyresearch.org.

Downloaded from www.microbiologyresearch.org by
IP: 155.198.8.192
On: Mon, 03 Dec 2018 14:31:56