Speciation of the neisserial component of the infant upper respiratory tract microbiome by nonculture methods: impact of early life events and vaccination

Roberto Fernandez Crespo

A thesis presented for the degree of Doctor of Philosophy of Imperial College London

Section of Paediatrics

Division of Infectious Diseases

Department of Medicine

Faculty of Medicine

Imperial College London

United Kingdom

October 2018

Abstract

Background: *Neisseria meningitidis* is a major causative agent of meningitis in infants. The introduction of conjugate vaccines has led to a decrease of incidence of disease caused by serogroups targeted by vaccines. Recently, a new vaccine, Bexsero®, was developed to target serogroup B isolates. Due to differences in composition with previous meningococcal vaccines, it was unclear if this vaccine could have an impact on *N. meningitidis* carriage, if it could impact the carriage of other *Neisseria* species, or other bacteria present in the upper respiratory tract. However, very little is known about what *Neisseria* species and in what proportion they appear in infants.

Methods: Two sample cohorts were analysed. One was comprised of unvaccinated infants from which samples were collected at 6 weeks, 3, 4.5, 6, 9, 12, 18, 24 and 36 months of life. The second, was a group of samples from infants vaccinated with Bexsero® in two different vaccination schedules. Microbiome data was assessed by sequencing fragments of the 16S rRNA gene using the 454 GS FLX platform. *Neisseria rplF* gene fragments were sequenced using a novel approach using the MiSeq platform.

Results and conclusions: Vaccination was associated with a change of 3% in the upper respiratory tract microbiome by the one-year time point. When considering *Neisseria* species data, less than 1% of the changes in abundance of *Neisseria* species populations were correlated with Bexsero® vaccination. Both of these results point towards Bexsero® not having a large impact on bacterial populations in the upper respiratory tract in infants. *N. subflava* and *N. cinerea* were the two most abundant species throughout the first three years of life and were accompanied by several other species in varying proportions. Multiple early life events, such as being exposed to cigarette smoke, were associated with changes in abundance of *Neisseria* species

Declaration

The work presented in this thesis is my own, and no part of it has been submitted previously or has been approved for the award of a degree by this or any other university. Data presented in this document that is not my own has been cited and attributed to their owners.

London,

October 2018

Roberto Fernandez Crespo

Copyright Declaration

The copyright of this thesis rests with the author and is made available under a Creative Commons Attribution Non-Commercial No Derivatives licence. Researchers are free to copy, distribute or transmit the thesis on the condition that they attribute it, that they do not use it for commercial purposes and that they do not alter, transform or build upon it. For any reuse or redistribution, researchers must make clear to others the licence terms of this work.

Acknowledgements

I would like to thank my supervisor, Prof. Simon Kroll, for giving me the opportunity to be part of this project. I would also like to thank my co-supervisor, Prof. Paul Langford, for all his support, unsurmountable amount patience, constant encouragement, kindness, and great early morning conversations, since the beginning of my MSc until the end of my PhD.

I am very grateful to all the people working in the Department of Paediatrics, for making it a fantastic place to work in. Also, to Prof. Miriam Moffatt's and Prof. William Cookson's team: Phil, Mike, Colin, Elena and Leah. Thank you for teaching me everything I know about MiSeq sequencing, welcoming me into your lab and for all your patience and help you have given me.

I will always be grateful to all the people working in the 'Pooh Team': Alex, Liz, Kathy, Emma, Ming-Shi, Batia, Salina, Becky and Zoe. It has been a true privilege sharing the last four years of my life with these wonderful people and all the ups and (mostly) downs of my project. Also, a huge thanks to my fantastic office and lab buddies throughout the years: Stephanie, Meg and Sara. You have all made everything a bit easier every day, one cup of coffee and one chatty afternoon at a time. I would also like to thank Fadil Bidmos and Leon Leanse for their support and for the long conversations in the lab. Thank you, Janine Bosse, for encouraging me take on this PhD; and for all the trust and help you have given me all these years.

The biggest thank you goes to my family: to my mom, my dad, my sister, my brother-in-law, and my grandparents, who have supported and believed in me, always. If I have made it this far because of you, and I will never be able to be sufficiently thankful for it. Finally, to my dearest John Canning, who has been by my side every day since I started this journey and has kept me sane through the tough days, celebrated all my achievements, listened to my many rants, and encouraged me to the best I can.

Thanks to all of you.

Abbreviations

Abbreviation	Meaning
ANI	Average nucleotide identity
BHI	Brain Heart infusion
BLAST	Basic local alignment search tool
bp	Base-pair
BSA	Bovine serum albumin
DMSO	Dimethyl sulfoxide
DNA	Deoxyribonucleic acid
dsDNA	Double-stranded deoxyribonucleic acid
EDTA	Ethylenediaminetetraacetic acid
F	Forward
FBS	Foetal bovine serum
GNA	Genome-derived *Neisseria* antigen
ID	Identification number
kb	Kilobase
LME	Lising matrix E
M	Molar
MATS	Meningococcal Typing Antigen System
MED	Minimum entropy decomposition
ml	Millilitre
MLST	Multi locus sequence typing
mm	Millimetre
NGS	Next Generation Sequencing
NMDS	Nonmetric multidimensional scaling
OMV	Outer membrane vesicle
OTU	Operational taxonomic unit
p	Probability
PBS	Primer binding sequence
PCoA	Principal component analysis
PCR	Polymerase chain reaction
pM	Picomolar
Q score	Quality score
QC	Quality control
QIIME	Quantitative insights into microbial ecology
qPCR	Quantitative polymerase chain reaction
R	Reverse
RBH	Reciprocal best hits
rMLST	Ribosomal multilocus sequence typing

RNA	Ribonucleic acid
rRNA	Ribosomal ribonucleic acid
spp.	Species
SV	Sequence variant
TE buffer	Tris-EDTA buffer
Tm	Melting temperature
Tris base	Tris(hydroxymethyl)aminomethane
UK	United Kingdom
URT	Upper respiratory tract
URTome	Upper respiratory tract microbiome
UV	Ultraviolet
V	Version
vs.	Versus
WGS	Whole Genome Sequencing
µl	Microlitre

Table of Contents

ABSTRACT ..... **III**

DECLARATION ..... **V**

COPYRIGHT DECLARATION ..... **VII**

ACKNOWLEDGEMENTS ..... **IX**

ABBREVIATIONS ..... **XI**

TABLE OF CONTENTS ..... **XIV**

LIST OF FIGURES ..... **XXII**

LIST OF TABLES ..... **XL**

1. CHAPTER 1: NEISSERIA SPECIES AND DISEASE **46**

1.1. NEISSERIA 46

1.2. NEISSERIA SPECIES IDENTIFICATION 47

1.3. USE OF NGS TECHNOLOGIES TO STUDY THE COMPOSITION OF BACTERIAL COMUNITIES 49

1.3.1. Sequencing of rplF as a Neisseria identification tool *50*

1.4. NON-PATHOGENIC NEISSERIA SPECIES 51

1.5. N. MENINGITIDIS 53

1.5.1. Meningococcal vaccines *55*

1.5.1.1. Meningococcus B vaccines: Bexsero® 56

1.5.1.1.1. Development of the vaccine 56

1.5.1.1.2. Bexsero® composition 57

1.5.1.1.3. Expected impact and coverage of Bexsero® antigens 58

1.6. MOTIVATIONS BEHIND THE ANALYSIS 60

1.7. HYPOTHESES 62

1.8. AIMS 63

2. CHAPTER 2: MATERIALS AND METHODS 65

2.1. LIST OF REAGENTS USED 65

2.2. LIST OF EQUIPMENT USED 67

2.3. REFERENCE SEQUENCES FOR RPLF SPECIES IDENTIFICATION 68

2.4. REFERENCE SEQUENCES USED FOR 16S RRNA GENE TAXONOMY ASSIGNMENT 69

2.5. SAMPLES USED 69

2.5.1. Adult swabs collected from healthy volunteers 70

2.5.2. Bacterial isolates used as reference sequences 71

2.5.3. Artificial DNA mixture samples 71

2.6. DNA EXTRACTION PROTOCOL 71

2.7. AMPLIFICATION OF THE RPLF GENE 72

2.8. CHROMATOGRAM SEQUENCE ANALYSIS 73

2.9. SEQUENCING LIBRARY PREPARATION FOR 16S RRNA SEQUENCING 74

2.9.1. PCR amplification 74

2.9.2. Gel extraction and purification 75

2.9.3. Quantification and pooling 75

2.10. SEQUENCING LIBRARY PREPARATION FOR RPLF GENE SEQUENCING 76

2.10.1. Protocol version 1 76

2.10.1.1. PCR amplification 76

2.10.1.2. Size selection and replicate pooling 77

2.10.1.3. Gel purification 78

2.10.1.4. Library quality control 78

2.10.1.5. Preparation of library and PhiX spike 79

2.10.2. Protocol version 2 79

2.11. BIOINFORMATICS AND STATISTICS 80

2.11.1. Sequencing file processing for Roche 454 SFF files (16S rRNA gene amplicons) 81

2.11.2. Sequencing file processing for Illumina MiSeq fastQ files (rplF amplicons) 81

2.11.3. Statistical analysis 82

2.12. REAL-TIME QUANTITATIVE PCR FOR THE RPLF GENE 82

3. CHAPTER 3: VALIDATION OF RPLF AS A NEISSERIA SPECIATION TOOL 84

3.1. INTRODUCTION 84

3.2. AIMS 85

3.3. SUMMARY OF METHODS 85

3.3.1. Samples used and DNA extraction protocol 85

3.3.1.1. Adult oropharyngeal throat swabs 85

3.3.1.2. Bacterial isolates 85

3.3.1.3. DNA extraction protocol 86

3.3.2. PCR amplification of the rplF gene 86

3.3.3. Analysis of Sanger sequencing chromatograms 86

3.3.4. Bioinformatic analysis of genomic sequences 87

3.4. RESULTS 88

3.4.1. Reproducibility of rplF sequencing, and testing using bacterial genomic DNA and that extracted from oropharyngeal throat swabs 88

3.4.2. Description of the rplF amplicon 92

3.4.3. Identification of Neisseria isolates from pure cultures and human oropharyngeal swabs. ... 93

3.4.4. Primer specificity 95

3.4.4.1. Polymorphisms in primer binding sites 96

3.4.5. Similarity of rplF sequences across upper respiratory tract commensal and other Neisseriaceae. 99

3.4.5.1. Is Morococcus cerebrosus related to N. mucosa? 102

3.5. DISCUSSION 104

4. CHAPTER 4: ADAPTATION OF THE RPLF SEQUENCING METHOD TO THE ILLUMINA MISEQ PLATFORM 111

4.1. INTRODUCTION 111

4.2. AIM 112

4.3. SUMMARY OF METHODS 112

4.4. NEXT GENERATION SEQUENCING: ILLUMINA MISEQ SEQUENCING METHOD 113

4.5. PRIMER DESIGN FOR CUSTOM AMPLICON SEQUENCING IN THE MISEQ PLATFORM 115
4.5.1. rplFV1 primer set development 117
4.5.1.1. Results obtained from sequencing libraries prepared using the rplfV1 primers 119
4.5.2. rplFV2 primer set development 124
4.5.2.1.1. Results for the rplFV2 sequencing run 1 126
4.5.2.1.2. Results from sequencing run 3 using the rplFV2 primer set 128
4.5.2.1.3. Results for sequencing run 5 130
4.5.2.1.4. Results from library 6 and 7 133
4.5.2.2. Summary of results for the rplFV2 primer set 134
4.5.3. Preparation of the rplFV3 primer set 139
4.5.3.1. Results from libraries generated using the rplFV3 primer set 140
4.6. DISCUSSION 145
5. CHAPTER 5: DEVELOPMENT OF A PIPELINE FOR THE ANALYSIS OF RPLF FASTQ FILES AND PRE-PROCESSING OF 454 SEQUENCING READS. 149
5.1. INTRODUCTION 149
5.2. AIMS 150
5.3. SUMMARY OF METHODS 150
5.4. RESULTS 151
5.4.1. Proposed read processing pipeline 151
5.4.1.1. Data pre-processing using QIIME 151
5.4.1.1.1. Tests using synthetic samples 155
5.4.1.2. Different approaches tested for rplF sequence identification 158
5.4.1.2.1. Use of the MED algorithm to cluster rplF sequences 159
5.4.1.3. Data processing in the R environment: Custom noise-filtering steps 168
5.4.2. Data pre-processing for 454 sequencing reads 172
5.5. DISCUSSION 175

6. CHAPTER 6: EFFECTS OF BEXSERO VACCINATION ON THE UPPER RESPIRATORY TRACT MICROBIOME 180

6.1. INTRODUCTION 180

6.2. AIM 181

6.3. SUMMARY OF METHODS 181

6.3.1. DNA extraction method 181

6.3.2. Library preparation for 16S rRNA gene libraries 181

6.3.3. Data processing and statistical tests 182

6.3.4. Sample summary 182

6.4. RESULTS 184

6.4.1. Rarefraction of OTU table 184

6.4.2. Validation work: comparison of repeated samples 187

6.4.3. Validation work: effect of different 454 GS FLX flow pattern in the data 188

6.4.4. Alpha diversity value comparisons 190

6.4.5. Comparison of the microbiome of the Dormice and Oxford cohorts 192

6.4.6. Assessment of Neisseria carriage and abundance using 16S rRNA reads 195

6.4.7. Analysis of the microbiome data using ordination methods 197

6.4.7.1. NMDS analysis 199

6.4.7.2. Constrained ordination and PERMANOVA analysis 203

6.4.8. Differential abundance comparisons 209

6.5. DISCUSSION 217

7. CHAPTER 7: EFFECT OF BEXSERO® VACCINATION IN THE CARRIAGE OF NEISSERIA SPECIES 222

7.1. INTRODUCTION 222

7.2. AIM 223

7.3. SUMMARY OF METHODS 223

7.3.1. Samples used 223

7.3.2. Data processing and statistics 223

7.3.3. Quantification of neisserial burden in samples 224

7.4. RESULTS 225
7.4.1. Sequencing data rarefraction 225
7.4.2. Validation of potential biases due to Forward primer binding 228
7.4.3. Validation of inter-experiment variation 229
7.4.4. Comparison of α-diversity values. 230
7.4.5. Summary of Neisseria species identified per time point 233
7.4.6. Effects of Bexsero® vaccination of carriage of Neisseria species 236
7.4.6.1. NMDS ordinations 237
7.4.6.2. PCoA ordinations 240
7.4.6.3. Constrained ordinations 243
..... 245
7.4.6.4. Species-based ordination 245
7.4.6.5. Differential abundance testing 246
7.4.6.5.1. Differential abundance testing using metagenomeSeq 247
7.4.6.5.1. Differential abundance testing using LEfSe 248
7.4.6.5.2. Changes observed within sample pairs 251
7.4.6.6. Quantification of the neisserial burden in infants 253
7.5. DISCUSSION 255
8. CHAPTER 8: EFFECT OF EARLY LIFE EVENTS ON NEISSERIA SPECIES 261
8.1. INTRODUCTION 261
8.2. AIMS 262
8.3. METHODS 262
8.3.1. Samples used 262
8.3.2. Statistical analysis 262
8.4. RESULTS 263
8.4.1. Neisseria population in infants 263
8.4.2. Effect of early life events in neisserial populations 269
8.4.2.1. Effects of crowding in Neisseria species carriage 272
8.4.2.2. Effect of the use of a pacifier 275

8.4.2.3. Effect of living with a smoker in the household 278
8.4.2.4. Effect of pets in household 282
8.4.2.5. Effects of feeding in neisserial populations 286
8.5. Discussion 290
9. CHAPTER 9: DISCUSSION 297
9.1. rplF as a Neisseria speciation tool 297
9.1.1. Problems arising from lack of sequence diversity 297
9.1.2. Implication of the rplF primer mismatches 300
9.2. Development of the rplF library preparation and data processing pipeline 301
9.3. Effect of Bexsero® on the infant upper respiratory tract microbiome 304
9.4. Effect of Bexsero® on the infant upper respiratory tract Neisseria population 305
9.5. Early life events and Neisseria populations 306
9.6. Future work 308
REFERENCES 311
APPENDIX 1: R ENVIRONMENT 329
APPENDIX 2: RPLFV3 PRIMER SEQUENCES 331
APPENDIX 3: LIST OF ISOLATE SEQUENCES USED IN CHAPTER 2 333
Isolates used for Neisseriaceae rplF sequence comparison 333
List of isolates used for upper respiratory tact commensals rplF sequence comparison 337
List of Neisseria isolates used in both analyses 350

List of figures

Figure 1 Number of reported cases of invasive meningococcal disease per serogroup in the European Union and in the United Kingdom (Y axis) from 1999 to 2016 (X axis). Data obtained from the European Centre for Disease Control. 54

Figure 2 Age standardized rate of invasive meningococcal disease in European countries in 2016. Darker colors represent a higher incidence. Figure obtained from the European Centre for Disease Control atlas. 55

Figure 3 Vaccination strategy and sample collection for the Oxford study samples. Each column represents the month of life a sample was taken, or a vaccine given. The labels inside the boxes represent the codes given to each group of samples collected. Yellow sections represent a time point in which the vaccine was administered and green secion when a oropharyngeal throat swab sample was collected. Sections with both colors indicate time points where the vaccine was administered the same day a oropharyngeal swab was collected. 70

Figure 4 Dendrogram illustrating the differences among rplF reference sequences. Each species is colour-coded as shown in the legend, and the black bar at the bottom of the Figure shows a distance equivalent to a 1% difference among sequences (4.13 bp difference). The values on the tree branches indicate the consensus support of each branch. 89

Figure 5 Alignment of rplF amplicon sequences. The alignment was constructed using MAFFT. Eight rplF reference sequences (ids: 1,121, 111, 16, 13, 10, 9 and 83) were used as representative sequences. The coloured bar above the sequence alignment indicates the level of conservation of each base, with higher values represented by taller green segments, less conserved bases in yellow, and the most variable bases by short red segments. The numbering of the bases is shown above the coloured bars. 90

Figure 6 Graphical representation of the primer binding sequences to the rplF gene. The green segment represents the rplF gene from N. meningitidis MC58. The blue bar the rplF amplicon described by Bennett et al., the yellow bar the amplicon proposed in this manuscript and the red segments the binding sites of the rplF forward (rplF_F) and reverse (prlF_R) primers..... 92

Figure 7 Dendrogram constructed using a subset of rplF reference and sequences obtained from bacterial isolates and oropharyngeal human throat swabs. The coloured blocks represent clusters of the samples being tested, each one representing one species (yellow: N. polysaccharea, orange: N. meningitidis, blue: N. subflava, grey: N. cinerea, red: N. lactamica, green: N. mucosa, light blue: N. elongata. Grey circles indicate that the sample was a bacterial isolate, while the orange triangles indicate a human sample. The bar underneath the graph represents a similarity of 5% in the sequence (20 bp). The values on the tree branches indicate the consensus support of each branch. 94

Figure 8 Comparison of chromatograms obtained from four each of human oropharyngeal samples, and bacterial genomic DNA. Each coloured line represents the fluorescent signal produced by each base as follows: T = green; A =red; C = blue; and G = yellow. The light-blue section behind the traces represents the quality for base sequenced, a higher bar representing a higher quality. The base corresponding to the most prominent peak for each position in shown below the traces. The numbers at the top of the figure represents the base of the amplicon being shown. 95

Figure 9 Heatmap showing the pairwise total number of differences among Neisseria rplF primer binding sequences. Brighter colours indicate a higher similarity of sequences. 97

Figure 10 Heatmaps showing the pairwise dissimilarities among *rplF* amplicons, and rplF primer binding sequences in the Neisseriaceae and other upper respiratory tract commensals. Figures 9A and 9C show the percentage of identity of rplF amplicons among Neisseriaceae and upper respiratory tract commensals, respectively. Figures 9B and 9D show the number of dissimilarities (mismatches), in the same order as Figures 9A and 9C. All four images have a colour bar on the top and right margins, depicting the area of the picture occupied by each genus, with the legend at the right side of the Figure

relating each genus to each colour. The percentage of identity is shown in green and the number of mismatches in pink, according to the legend at the bottom right corner. 101

Figure 11 Heatmap showing the pairwise ANI scores comparing Neisseria and Morococcus species. Each tile is coloured according to the legend, wth higher ANI scores shown in brighter colours. 103

Figure 12 Comparison of Protocol version 1 and 2 used to generate MiSeq-compatible libraries. Each block represents one of the steps of the library preparation protocol. The standard library preparation, also refered to as single-step library protocol, is shown in orange, and the two-step library protocol in green. 113

Figure 13 Bridge amplification of DNA molecules present in in a sequencing library. P5 and P7 binding sequences are shown in purple, P5 and P7 sequences are shown in green, index sequences in yellow, pad sequences in orange, and primer sequences in red, amplicon sequences are shown in grey. 114

Figure 14 Description of the four different sequencing steps taking place during a MiSeq sequencing experiment. P5 and P7 binding sequences are shown in purple, P5 and P7 sequences are shown in green, index sequences in yellow, pad sequences in orange, and primer sequences in red, amplicon sequences are shown in grey. Blue segments and arrows indicate primer bindings and the direction of the amplidication of the DNA, respectively. 115

Figure 15 PCR and primer setup for the rplFV1 primer set. P5 and P7 sequences are shown in green, index sequences in yellow, pad sequences in orange, and primer sequences in red, amplicon sequences are shown in grey. 117

Figure 16 Summary of the quality scores obtained for the forward (A) and reverse (B) reads for the third rplF library sequenced. The X axis shows the base, or group of bases, sequenced in each direction, and the Y axis represents the quality of the scores. The red, yellow and green background segments indicate quality scores in between 0 and 20,20 and 20 and above 30, respectively. The sequencing quality scores of each base, or group of bases is represented with a boxplot (shown in yellow), with a red line indicating the median score. The blue line indicates the average quality value. 120

Figure 17 Estimated ratio of appearance of each base in the rplF amplicon, image obtained from the PubMLST Neisseria database. The appearance of each bease is represented with colours according to the color key on the top left corner. 122

Figure 18 PCR and primer set up for the rplFV2 primer set. P5 and P7 sequences are shown in green, index sequences in yellow, pad sequences in orange, link sequences in purple and primer sequences in red, amplicon sequences are shown in grey. 124

Figure 19 Quality per base obtained for all the bases sequenced for the first sequencing run for the rplFV2 primer set. The X axis indicates the bases sequenced during the sequencing steps in a MiSeq sequencing experiment, and the Y axis the mean sequence quality score. Segments are colored according the percentage of reads that shared that same quality scores for each base. 127

Figure 20 Quality per base obtained for all the bases sequenced for the third sequencing run for the rplFV2 primer set. The X axis indicates the bases sequenced during the sequencing steps in a MiSeq sequencing experiment, and the Y axis the mean sequence quality score. Segments are colored according the percentage of reads that shared that same quality scores for each base. 128

Figure 21 Quality scores per base reverser read for the third rplFV2 library sequenced. The X axis shows the base, or group of bases, sequenced in each direction, and the Y axis represents the quality of the scores. The red, yellow and green background segments indicate quality scores in between 0 and 20,20 and 20 and above 30, respectively. The sequencing quality scores of each base, or group of bases is represented with a boxplot (shown in yellow), with a red line indicating the median score. The blue line indicates the average quality value. 129

Figure 22 Boxplots showing the cluster counts for the Forward read (blue) and the Reverse (green). The Y axis indicates the density of clustering at a scale of 1000 clusters per mm^2. 131

Figure 23 Quality score per base obtained for the fifth sequencing run using the rplFV2 primers. The X axis indicates the bases sequenced during the sequencing steps in a MiSeq sequencing experiment, and the Y axis the mean sequence quality score. Segments are colored according the percentage of reads that shared that same quality scores for each base. 132

Figure 24 Bioanalyzer traces obtained for run 1. The X axis indictates the size of DNA fragments, measured in base pairs, and the Y axis the fluorescence units (FU) measured. A higher fluorescence intensity represents a higher DNA concentration. The peaks representing 15 and 1500 bp are the internal controls used in the experiment. 133

Figure 25 Quality scores per base obtained during sequencing using the V3 kit (A) and a V2 kit (B). The X axis indicates the bases sequenced during the sequencing steps in a MiSeq sequencing experiment, and the Y axis the mean sequence quality score. Segments are colored according the percentage of reads that shared that same quality scores for each base. 134

Figure 26 Diagram showing the two-step PCR approach for MiSeq library preparation. P5 and P7 sequences are shown in green, index sequences in yellow, pad sequences in orange, link sequences in purple and primer sequences in red, amplicon sequences are shown in grey. 136

Figure 27 Schematic representation of the two-step PCR approach including a spacer segment. P5 and P7 sequences are shown in green, index sequences in yellow, pad sequences in orange, spacer sequences in blue,link sequences in purple and primer sequences in red, amplicon sequences are shown in grey. 137

Figure 28 Quality scores per sequencing cycle for the rpFV3 Library1. The X axis indicates the bases sequenced during the sequencing steps in a MiSeq sequencing experiment, and the Y axis the mean sequence quality score. Segments are colored according the percentage of reads that shared that same quality scores for each base. 142

Figure 29 Quality scores per sequencing cycle for the rpFV3 Library 3. The X axis indicates the bases sequenced during the sequencing steps in a MiSeq sequencing experiment, and the Y axis the mean sequence quality score. Segments are colored according the percentage of reads that shared that same quality scores for each base. 143

Figure 30 Summary of scripts used to process sequencing data files. The scripts used are shown in colour depending on the package or language used (QIIME = yellow, Cutadapt = dark green, FlowClus = red, Bash = grey, Oligotyping = purple, BLAST = light green, and R = blue). 150

Figure 31 Results of 16 randomly selected samples obtained after processing the reads using different clustering thresholds. The X axis indicates the results for each of the UCLUST clustering thresholds used, and the Y axis the number of reads assigned to the different species, which are coloured according tto the legend on the left. Reads without a definite species assignment are classified as 0. 154

Figure 32 relative proportion of species identified in the synthetic samples created and processed with UCLUST. The X axis indicates the results for each of the UCLUST clustering thresholds used, and the Y axis the number of reads assigned to the different species, which are coloured according tto the legend on the left. Reads without a definite species assignment are classified as 0. 158

Figure 33 Shannon entropy values calculated for each base of the rplF amplicon using rplF reference sequences. Each base of the alignment on the X axis, and the entropy of each base on the Y axis. 159

Figure 34 Relative proportion of species identified in the synthetic samples created and processed with MED for 3 sets of randomly selected rplF sequence combinations. The X axis indicates the results for each of the different cluster size filtering steps tested, and the Y axis the number of reads assigned to the different species, which are coloured according tto the legend on the left. 162

Figure 35 Relative abundances of species identified when testing different variable combinations in the MED algorithm. The X axis indicates the results for each of the different cluster size filtering steps tested, and the Y axis the number of reads assigned to the different species, which are coloured according to the legend on the left. The names of the different variables on the X axis are omitted to favour visualization. 164

Figure 36 Entropy per base calculated for the analysis of 10 different MiSeq libraries. The red line shows the entropy threshold used to select bases considered to differentiate nodes), which is an indication the appropriateness of MED analyses for the rplF amplicon. 168

Figure 37 Identity values per node found in the libraries prepared with human samples. Nodes, shown in the X axis, are ordered by the identity percentage, displayed on the Y axis, of the most

prominent read within each node. The nodes accepted appear in blue, the rejected ones in read, and the green line marks the 97% threshold used. 169

Figure 38 Node contribution towards total amount of reads. the Y axis indicates the number of nodes identified that contributed towards a percentage of reads (X axis). The green horizontal line indicates the threshold selected for node filtering 171

Figure 39 Alpha diversity mean values (Y axis) for Dormice (red) and Oxford (blue) samples at different sampling depths (X axis). The results presented here represent 100 randomly selected samples. 185

Figure 40 Number of samples lost (Y axis) at different sampling depth (X axis). The three different coloured lines indicate the numbers of samples lost for Oxford samples (blue), Dormice samples (green) or control samples (red). The two dashed vertical lines highlight the sampling depth selected for further inspection. 186

Figure 41 Number of samples lost at (Y axis) sampling depth (X axis) in between 650 and 720 reads. The three different coloured lines indicate the numbers of samples lost for Oxford samples (blue), Dormice samples (green) or control samples (red). The two dashed vertical lines highlight the sampling depth selected for further inspection. 187

Figure 42 Read counts per genus and sample repeats identified for 10 repeated samples. Each column represents one sample, as per the label above, and each row represents the read counts per genus, showing one bar per repeat of a given sample. 188

Figure 43 NMDS ordination using the Weighted UniFrac distance. Sample are coloured according to the flow pattern used to sequence the library the samples were in. 189

Figure 44 Abundance per genus in samples sequenced using flow pattern A and B. Each column represents one sample, as per the label above, and each row represents the read counts per genus. All sample was sequenced using flow patterns A and B. 189

Figure 46 Alpha diversity values calculated per time sample collection group in the Oxford cohort. Each time point is represented in the X axis, while the Y axis indicates the alpha diversity values

calculated using the total number of OTUs (Observed), the Shannon index (Shannon) and the inverse Simpson metric (InvSimpson). 191

Figure 45 Boxplots describing the alpha diversity values calculated per time point of the Dormice study. Each time point is represented in the X axis, while the Y axis indicates the alpha diversity values calculated using the total number of OTUs (Observed), the Shannon index (Shannon) and the inverse Simpson metric (InvSimpson). 191

Figure 47 Alpha diversity values calculated for earlier time points (6mo for Dormice, V06 and V07 for Oxford. All shown in blue) and later time points (12mo for Dormice, V12-G1, V12-G2 for Oxford. All shown in red) in both cohorts. Each time point is represented in the X axis, while the Y axis indicates the alpha diversity values calculated using the total number of OTUs (Observed), the Shannon index (Shannon) and the inverse Simpson metric (InvSimpson). 192

Figure 48 Relative proportion of the average reads (Y axis) for the top 20 genera identified for each time point included in the Dormice and Oxford cohorts (X axis). Each of the 20 most abundant genera are coloured according to the legend on the right. All other readsbelonging to other genera were grouped in the 'Other' category. 193

Figure 49 Comparison of sample groups in earlier and later time points for Dormice and Oxford samples. Early (6mo, V06 and V07) and late (12mo, V12-G1, V12-G2) time points are shown in the X axis, and the average proportion of reads assigned to each genus on the Y axis. Each of the 20 most abundant genera are coloured according to the legend on the right. All other readsbelonging to other genera were grouped in the 'Other' category. 194

Figure 50 Average percentage of reads belonging to Neisseria per time point for Dormice (blue) and Oxford (red) samples. Time points are shown chronologically, with lighter colours indicating earlier time points. 195

Figure 51 Percentage of infants with Neisseria reads per time point. All time points from Dormice and Oxford cohorts are shown chronologically. Dormice time points are shown in blue, and Oxford time points in red, with lighter colours indicating earlier time points. 196

Figure 52 Number of total reads per Neisseria OTU and time pont. The three most abundant OTUs (Neisseria_4, Neisseria_ 9 and Neisseria_27 appear) in yellow, blue and red, respectively. All other Neisseria OTUs were grouped into the 'Other' category. 196

Figure 53 Log2 values of the top three most abundan Neisseria OTUs per time point. For each OTU, the read counts are summarized using a boxplot per time point. 197

Figure 54 NMDS ordinations generated for the earlier time point samples (6mo, V06 and V07) using weighted UniFrac (A-D), UniFrac (E-H), Jaccard (H-K) and Bray-Curtis (M-P) metrics. Samples are coloured based on its time point, vaccination status, sex, and sample origin. Ellipses represent a 95% confidence interval for each group of samples, and the large circles marked in black represent the average position of each group. 201

Figure 55 NMDS ordinations generated for the later time point samples (12mo, V12-G1 and V12-G2) using weighted UniFrac (A-D), UniFrac (E-H), Jaccard (H-K) and Bray-Curtis (M-P) metrics. Samples are coloured based on its time point, vaccination status, sex, and sample origin. Ellipses represent a 95% confidence interval for each group of samples, and the large circles marked in black represent the average position of each group. 202

Figure 56 Constrained ordination plots obtained for the earlier time points using the Weighted UniFrac distance and the dbRDA method for the time point (A), Vaccination status (B), Sex (C) and Sample origin (D) variables. 208

Figure 57 Constraind ordination plots obtained for the later time points using the Weighted UniFrac distance and the dbRDA method for the time point (A), Vaccination status (B), Sex (C) and Sample origin (D) variables. 208

Figure 58 Cladogram generated using taxonomic levels in the OTU reads obtained. Coloured clades and nodes, according to the legend on the top left corner, highlight taxonomic levels where a significant difference was identified in a particular group at earlier (A) and later (B) time points. Taxonomic levels above genus, were marked with a letter, as shown in the legend in the right. A two-letter code was used to designate each taxonomic level (Fa: family, Or: order, Cl: class). 211

Figure 59 Bar charts showing LDA scores calculated for the OTUs associated with the time point (A), vaccination status (B) and sample origin (C) variables in earlier time points. Larger LDA score represent stronger associations in between an OTU and a group. For each plot, bars are coloured according to the legend at the top. 212

Figure 60 Bar charts showing LDA scores calculated for the OTUs associated with the time point (A) and vaccination status (B) variables in later time points. The sample origin variable is not shown as it is identical to the vaccination status at this point. Larger LDA score represent stronger associations in between an OTU and a group. For each plot, bars are coloured according to the legend at the top. 213

Figure 61 Alpha diversity mean values (Y axis) for Dormice (red) and Oxford (blue) samples at different sampling depths (X axis). The results presented here represent 100 randomly selected samples. 225

Figure 62 Samples lost (Y axis) per rarefraction depth (X axis) for Dormice and Oxford samples. Each coloured lines indicate the numbers of samples lost for Oxford samples (blue) or Dormice samples (red). The vertical line indicates the sampling depth used for rarefraction. 226

Figure 63 Number of samples lost due after rarefaction (Y axis) per time point of Dormice and Oxford samples (X axis). 227

Figure 64 Relative abundance of species identified when sequencing the libraries prepared using the modified and unmodified Forward primer (Modified and Unmodified rows, respectively). A reference of the expected values is shown in the row with the same name. Each column shows the species used to create the mock community groups (Nlac = N. lactamica, Nmen = N. meningitidis, Nsub = N. subflava, Npol = N. polysaccharea). 229

Figure 65 Comparison of 12 randomly selected sample pairs to identify potential sources of variation among sequencing experiments. Each column represents one sample, as per the label above, and each row represents the read counts per genus, showing one bar per repeat of a given sample. 230

Figure 66 Boxplots describing the alpha diversity values calculated per time point of the Dormice study. Each time point is represented in the X axis, while the Y axis indicates the alpha diversity values calculated using the total number of OTUs (Observed), the Shannon index (Shannon) and the inverse Simpson metric (InvSimpson). 231

Figure 68 Alpha diversity values calculated per time sample collection group in the Oxford cohort. Each time point is represented in the X axis, while the Y axis indicates the alpha diversity values calculated using the total number of OTUs (Observed), the Shannon index (Shannon) and the inverse Simpson metric (InvSimpson). 233

Figure 67 Alpha diversity values calculated for earlier time points (6mo for Dormice, V06 and V07 for Oxford. All shown in green) and later time points (12mo for Dormice, V12-G1, V12-G2 for Oxford. All shown in blue) in both cohorts. Each time point is represented in the X axis, while the Y axis indicates the alpha diversity values calculated using the total number of OTUs (Observed), the Shannon index (Shannon) and the inverse Simpson metric (InvSimpson). 233

Figure 69 Proportion of Neisseria species (Y axis) counts identified per time point (X axis). Time points are displayed chronologically, including Dormice and Oxford samples. 234

Figure 70 Relative abundances of Neisseria species identified in earlier (6mo, V06 and V07) and later (12mo, V12-G1, V12-G2) time points in the Dormice and Oxford cohorts. 235

Figure 71 Proportion of infants colonized by each Neisseria species identified per time point. The number of samples per time point is indicated underneath each time point. Both N. elongata subspecies were merged with the remaining reads for that species. Species were sorted by the total sum of carriers. The number of samples per time point is shown underneath the time point labels. 236

Figure 72 Shepard plots showing the correlation in between the Observed dissimilarities (X) and distances shown in the ordination (Y) for the weighted UniFrac (left) and UniFrac (right) distances. The linear and non-metric R^2 values are displayed on the top left corner of each plot. Blue dots represent the values obtained in each iteration and the red line the mean value obtained. 237

Figure 73 NMDS ordinations using earlier samples and created using the weighted UniFrac (A-D) and UniFrac (E-H) distances. Samples were coloured according to the sample's time point (A, E), vaccination status (B, F) sex (C,G) and origin (D,H). 239

Figure 74 NMDS ordinations using later samples and created using the weighted UniFrac (A-D) and UniFrac (E-H) distances. Samples were coloured according to the sample's time point (A, E), vaccination status (B, F) sex (C, G) and origin (D, H)..... 239

Figure 75 Proportion of the total variation (eigenvalues) explained by each axis in the PCoA ordinations calculated using weighted UniFrac (A), UniFrac (B), Bray-Curtis (C), and Jaccard (D) metrics. 241

Figure 76 PCoA ordinations using earlier samples and created using the weighted UniFrac (A-D) and UniFrac (E-H) distances. Samples were coloured according to the sample's time point (A, E), vaccination status (B, F) sex (C,G) and origin (D,H). 242

Figure 77 PCoA ordinations using earlier samples and created using the weighted UniFrac (A-D) and UniFrac (E-H) distances. Samples were coloured according to the sample's time point (A, E), vaccination status (B, F) sex (C,G) and origin (D,H). 242

Figure 78 Constrained ordinations created using the dbRDA algorithm using the weighted UniFrac metric when comparing time points (A), vaccination status (B), Sample origin (C) and sex (D) variables. 245

Figure 79 PCoA ordination obtained using 6-month (A, B) and later (C, D) samples, using the weighted UniFrac distance Plots A and C are sample-based plots while B and D focus on species.. 246

Figure 80 Carriage of N. meningitidis in later time point samples (Dormice: 12mo, Oxford: V12-G1, V12-G2). Figure A shows the total number (count) of infants colonized per time point. Figure B shows boxplots describing number reads identified for this bacterium, while C focuses on those samples with fewer than than 150 reads. 247

Figure 81 Results obtained when using the LEfSe algorithm when using samples collected at the 12-month time point. The results show the comparison, at the species level, of Dormice samples and

the two Oxford sample time points (A, B), samples compared by vaccination status. Larger LDA score values represent a stronger association of the species with a group, according the legend. For each plot, bars are coloured according to the legend at the top..... 250

Figure 82 Variation within sample pairs observed for Dormice samples (Not vaccinated) and the two Oxford sample groups (Group 1, Group 2) when comparing N. cinerea (A), N polysaccharea (B), N. subflava (C), and N. lactamica (D) abundances. For each plot, dots represent the abundance in each sample, and samples from the same infant are connected by a coloured line (red if decreases, blue If it increases)..... 252

Figure 83 Total neisserial burden measured per time point using a qPCR approach described using bloxplots. Neisserial burden is measured as the log10 of the estimated total number of copies of the rplF gene per swab. 253

Figure 84 Mean absolute species abundance per time point in earlier (6mo, V06, V07) and later (12mo, V12-G1, V12-G2) time points. The Y axis shows the mean abundance of all the samples in each time point in terms of number of Neisseria per swab. Reads were normalized to account for the different number of samples in each time point..... 254

Figure 85 Relative abundances of each Neisseria species identified for 70 randomly selected infants at each sample collection time point.All available time points are displayed for each set of samples from a given infant. Empty columns represent samples that either could not be collected for a particular time point, or a sample that was collected but either no neisserial DNA could be identified or it did not meet the rarefaction threshold of 2000 reads. 265

Figure 86 Odd ratios calculated for each species, marked in the X axis, when considering their abundance or presence, based on the presence the species marked in the Y axis labels. The reference group considered when calculating the odds was those colonized, hence, odd ratios above 1 indicate an increase in odds of having a higher abundance or presence when the species in the Y label is present. Comparisons of the effect of one species onto themselves were ignored for this analysis as they will always be correlated. The confidence interval bars are not shown for the N. mucosa and N.

mucosa/N. oralis odds when considering the presence of N. subflava as they were too varied and obfuscated the visualization. The upper and lower confidence intervals were 2.258 and 512 for N. mucosa, 2.146 and 486 for N. mucosa/N. oralis..... 267

Figure 87 Odd ratios calculated for each species, marked in the X axis, when considering the abundance of each of the species marked in the Y axis labels. The reference group considered when calculating the odds was those colonized, hence, odd ratios above 1 indicate an increase in odds of having a higher abundance or presence when the species in the Y label is present..... 269

Figure 88 Example of output values obtained from the fitTimeSeries data (A), and box plots generated per time point and level in the variable of interest. Both figures show normalized read counts in the Y axis, and the time variable showing the months of life for figure A, and each sample collection time point for figure B on the X axis. 271

Figure 89 Distribution of crowding index values (A), and the number of samples per level (B). Figure A shows the number of samples aggregated by crowding index. Figure B shows the total number of samples when divided by into a 'low' or 'High' crowding index. A low crowding index is defined as a ratio of people living in a household devided by the number of rooms (excluding kitchen and bathrooms) equal or lower than one. A High crowding index was defined as a ratio higher than one. 272

Figure 90 A) The modelled difference in N. lactamica abundance based on normalised read counts and B) the distribution of the normalized abundance data over time with regard to crowding at birth Both figures show normalized read counts in the Y axis, and the time variable showing the months of life for figure A, and each sample collection time point for figure B on the X axis..... 273

Figure 91 A) The modelled difference in N. subflava abundance based on normalised read counts and B) the distribution of the normalized abundance data over time with regard to crowding at birth. Both figures show normalized read counts in the Y axis, and the time variable showing the months of life for figure A, and each sample collection time point for figure B on the X axis..... 274

Figure 92 A) The modelled difference in N. oralis abundance based on normalised read counts and B) the distribution of the normalized abundance data over time with regard to crowding at birth. Both figures show normalized read counts in the Y axis, and the time variable showing the months of life for figure A, and each sample collection time point for figure B on the X axis. 274

Figure 93 Number of samples (counts) per level when considering the use of pacifier. 276

Figure 94 A) The modelled difference in N. lactamica abundance based on normalised read counts and B) the distribution of the normalized abundance data over time with regard s to the use of a pacifier. Both figures show normalized read counts in the Y axis, and the time variable showing the months of life for figure A, and each sample collection time point for figure B on the X axis. 277

Figure 95 A) modelled difference in N. meningitidis abundance based on normalised read counts and B) the distribution of the normalized abundance data over time with regard s to the use of a pacifier. Both figures show normalized read counts in the Y axis, and the time variable showing the months of life for figure A, and each sample collection time point for figure B on the X axis. 278

Figure 96 A) modelled difference in N. mucosa/N. oralis abundance based on normalised read counts and B) the distribution of the normalized abundance data over time with regard s to the use of a pacifier. Both figures show normalized read counts in the Y axis, and the time variable showing the months of life for figure A, and each sample collection time point for figure B on the X axis. 278

Figure 97 Number of samples (counts) per level when considering the presence of a smoker in the house. 279

Figure 98 A) modelled difference in N. bergeri abundance based on normalised read counts and B) the distribution of the normalized abundance data over time when considering infants living with or without a smoker in the house. Both figures show normalized read counts in the Y axis, and the time variable showing the months of life for figure A, and each sample collection time point for figure B on the X axis. 280

Figure 99 A) modelled difference in N. cinerea abundance based on normalised read counts and B) the distribution of the normalized abundance data over time when considering infants living with

or without a smoker in the house. Both figures show normalized read counts in the Y axis, and the time variable showing the months of life for figure A, and each sample collection time point for figure B on the X axis. 280

Figure 100 A) modelled difference in N. mucosa abundance based on normalised read counts and B) the distribution of the normalized abundance data over time when considering infants living with or without a smoker in the house. Both figures show normalized read counts in the Y axis, and the time variable showing the months of life for figure A, and each sample collection time point for figure B on the X axis. 281

Figure 101 A) modelled difference in N. subflava abundance based on normalised read counts and B) the distribution of the normalized abundance data over time when considering infants living with or without a smoker in the house. Both figures show normalized read counts in the Y axis, and the time variable showing the months of life for figure A, and each sample collection time point for figure B on the X axis. 281

Figure 102 A) modelled difference in N. oralis abundance based on normalised read counts and B) the distribution of the normalized abundance data over time when considering infants living with or without a smoker in the house. Both figures show normalized read counts in the Y axis, and the time variable showing the months of life for figure A, and each sample collection time point for figure B on the X axis..... 282

Figure 103 Number of samples collected (counts) in households without pets (red), and those were at least one pet was present (Cat = yellow, Dog = green, Fish = blue, Rabbit = purple). 283

Figure 104 A) modelled difference in N. elongata abundance based on normalised read counts and B) the distribution of the normalized abundance data over time when considering infants living in a household with a pet. Both figures show normalized read counts in the Y axis, and the time variable showing the months of life for figure A, and each sample collection time point for figure B on the X axis. 284

Figure 105 A) modelled difference in N. bergeri abundance based on normalised read counts and B) the distribution of the normalized abundance data over time when considering infants living in a household with a pet. Both figures show normalized read counts in the Y axis, and the time variable showing the months of life for figure A, and each sample collection time point for figure B on the X axis. 284

Figure 106 A) modelled difference in N. lactamica abundance based on normalised read counts and B) the distribution of the normalized abundance data over time when considering infants living in a household with a furry pet. Both figures show normalized read counts in the Y axis, and the time variable showing the months of life for figure A, and each sample collection time point for figure B on the X axis. 285

Figure 107 Number of samples (Count) from infants that were breastfed, formula fed, or had mix feeding at birth. Bars representing the number of samples per level are coloured according to the legend of the right. 287

Figure 108 Number of samples (Count) from infants that were breastfed, formula fed, or had mix feeding prior to sample collection. . Bars representing the number of samples per level are coloured according to the legend of the right. 287

Figure 109 Increase or decrease in abundance measured for each feeding routine. In each ellipsis, the inner darker circles show the results obtained from birth feeding data, while the others show the data for the feeding at the time of sample collection. The arrows show an increase (up) or a decrease (down) measured for each species. 289

List of Tables

Table 1 Protein components in Bexsero®. GNA denotes the genome-derived Neisseria antigen code given to each gene. GNA numbers denote the code given to genome-derived neisserial antigens. 58

Table 2 Reference rplF amplicon sequences obtained from the PubMLST Neisseria database. The first column shows the species included in the database, the ID numbers assigned to the sequences and the total number of sequences for each species. IDs marked with '*' refer to sequences without a definite species assignment, while '**' identifies sequences with more than one species assignment. 68

Table 3 Summary of the dissimilarities in between rplF amplicon sequences measured in bp differences. This Table excludes rplF reference sequences without a species assigned, or those labelled as Neisseria sp. The Table shows the average distance among sequences from the same species (intraSD), the average distance with the other species (interSD), the species with the closest distance (nearest group), and the average distance to the nearest group, which shows the average number of bp differences to the reference rplF sequences belonging to the closest species. 91

Table 4 Number of isolates with a perfect match to the published rplF Forward primer (Match), those with at least 1 mismatch (Mismatch) and the base change seen. 98

Table 5 Sequences used in the rplFV1 primer set 118

Table 6 Melting temperature of rplF primers and sequencing and indexing primers using in the rplFV1 primer set. For each sequence, the minimum calculated Tm (Min Tm), the average Tm (Avg Tm) and highest possible (Max Tm) is shown. The minimum and maximum Tm values are shown only for those sequences with ambiguous bases in its sequence. 119

Table 7 Comparison of sequences making up the rplFV1 and rplFV2, all sequences shown are in the 5' to 3' direction. 125

Table 8 Sequence and melting temperatures for the rplFV2 sequencing and indexing primers. All temperatures are shown in ºC. For each sequence, the minimum calculated Tm (Min Tm), the average Tm (Avg Tm) and highest possible (Max Tm) is shown. The minimum and maximum Tm values are shown only for those sequences with ambiguous bases in its sequence. 125

Table 9 Summary of MiSeq sequencing experiments for the rplFV2 runs. For each sequencing run, the library used, The MiSeq kit version, the percentage of PhiX spike in the final library, and the concentration, in pM, of of the PhiX spike and library used is shown. 126

Table 10 Sequences designed for the Reverse primer spacers for a two-step, staggered rplF library design. Proposed spacer sequences are shown in black, linker sequence in purple, and spacer sequences in red. Nucleotides, and the total count for each nucleotide is displayed per sequencing cycle. 138

Table 11 Sequences designed for the Forward primer spacers for a two-step, staggered rplF library design. Proposed spacer sequences are shown in black, linker sequence in purple, and spacer sequences in red. Nucleotides, and the total count for each nucleotide is displayed per sequencing cycle. 139

Table 12 Sequences used to construct the rplFV3 PCR primers. 140

Table 13 Run metrics for rplFV3 Libraries 1 and 2. For each library and sequencing direction, the cluster density (in 1000 clusters/mm^2), the number of clusters passing quality filtering (CPF), the values of phasing and prephasing, the percentage of reads above a mean quality score of 30 (Q>30), and the percentage of PhiX reads identified (PhiX reads (%)) is displayed. 141

Table 14 Sequencing statistics from 12 libraries prepared using human oropharyngeal throat swabs. For each library, the percentage PhiX spike, the library concentration used (pM), the cluster density (in 1000 clusters/mm^2) and number of clusters passing quality filtering (CPF), the values of phasing and prephasing, the percentage of reads above a mean quality score of 30 (Q>30), and the percentage of PhiX reads identified (PhiX reads (%)) and sequencing yield (in Gb) is displayed. 144

Table 15 Combinations of rplF reference sequences used to create synthetic samples. Diff indicates the number of differences (Diff), in base pairs, in between the two sequences. 156

Table 16 Most variable bases in the rplF amplicon. 160

Table 17 Parameter combinations used to test the MED algorithm 163

Table 18 Summary statistics obtained after analysing 32 different MED outputs. The first three columns show the proportion of nodes in which the species identified for all the reads within them matched that of the most abundant read in each node at the 98, 99 and 100% level (98%, 99% and 100% in the Table). This script also provides the final number of nodes identified (Nodes), the average ratio of nodes with the same species as the most abundant SV in a node (average species count, ASC), the average ratio of nodes with the same rplF ID as the main sequence (average profile count, APC), the percentage of nodes that had a BLAST result with an identity to rplF sequences below 95% (Identity<95%), the percentage of nodes that did not match anything in the rplF database (lost), and the final number of reads obtained (total reads). 167

Table 19 Summary statistics of a subset of discarded nodes for 15 selected nodes. For each node, the total number of reads, the percentage of identity, the rplF prifiles found within that node, and the best rplF and BLAST match are displayed. 170

Table 20 Number of samples present per variable analysed. Samples are devided according to the different levels in each variable. 183

Table 21 Stress values, linear R^2 and non-linear R^2 values calculated for the different NMDS ordinations generated, the values obtained when using the Weighted UniFrac metric are highlighted in red. 200

Table 22 R^2 and p-values obtained for the constrained ordination analyses performed using the Weighted UniFrac using the PERMANOVA, dbRDA and CCA methods. P-values lower than 0.05 are highlighted in red. 207

Table 23 OTUs with a significant difference in abundance identified when using the zero-inflated log-normal distribution model in metagenomeSeq for earlier time points. The table shows the number

of samples in each group that had reads for each OTU, the total read count in each group, the odds ratio, the log2 fold change plus the standard error, and the adjusted P-values for each OTU. 215

Table 24 OTUs with a significant difference in abundance identified when using the zero-inflated log-normal distribution model in metagenomeSeq for later time points. The table shows the number of samples in each group that had reads for each OTU, the total read count in each group, the odds ratio, the log2 fold change plus the standard error, and the adjusted P-values for each OTU. 216

Table 25 Number of samples per time point and cohort included in the analysis ("Samples included"), the proportion of samples obtained after sequencing, quality filtering of reads and rarefraction ("Samples after QC"), and the proportion of the total number of samples per time point. 228

Table 26 Pair-wise comparisons of alpha rarefaction values calculated for the number of observed taxa. Significant values are shown in red 232

Table 27 Pair-wise comparisons of alpha rarefaction values calculated for the Shannon (lower triangle) and Inverse Simpson (upper triangle) metrics. Significant values are shown in red..... 232

Table 28 P-values and R^2 values obtained for the PERMANOVA, dbRDA and CCA constrained ordinations using the weighted UniFrac distances. P-values below 0.05 are highlighted in red. 244

Table 29 Sequence variants identified as significantly different when compared to its corresponding group. The columns show, in order, the comparison in which the SV were found to be significant, the SV involved, the number of samples in each group carrying that SV, the total number of reads per group, the odds ratio, log2 fold change and standard error, and the adjusted p-values 248

Table 30 Number of infants per time point in each level considered. 273

Table 31 Odds ratios calculated based on abundance and presence of N. lactamica, N. subflava and N. oralis when considering infants in a high or low crowding index household. The odds ratio (Odds estimate) is presented with confidence intervals (Conf. int low and Conf. int High), the correspondent p-value, and the adjusted p-value using the Bonferroni correction. 275

Table 32 Number of infants per time point using (Yes) or not using (No) a pacifier..... 276

Table 33 Odds ratios calculated based on abundance and presence of N. lactamica, N. bergeri, N. meningitidis and N. mucosa/N. oralis when considering the use of a pacifier. The odds ratio (Odds estimate) is presented with confidence intervals (Conf. int low and Conf. int High), the correspondent p-value, and the adjusted p-value using the Bonferroni correction..... 277

Table 34 Number of infants per time point living with (Yes) or without (No) a smoker in their household..... 279

Table 35 Odds ratios calculated based on abundance and presence of N. bergeri, N. cinerea, N. mucosa, N. oralis and N. subflava when considering infants living with a smoker in the household. The odds ratio (Odds estimate) is presented with confidence intervals (Conf. int low and Conf. int High), the correspondent p-value, and the adjusted p-value using the Bonferroni correction..... 282

Table 36 Number of infants per time point not living with pets (No), those living with pets (Yes-All), and those living with furry pets in particular (Yes – Furry) 283

Table 37 Odds ratios calculated based on abundance for all Neisseria species when considering infants living with a pet (pets-All), or just furry pets (Pets-Furry). The odds ratio (Odds estimate) is presented with confidence intervals (Conf. int low and Conf. int High), the correspondent p-value, and the adjusted p-value using the Bonferroni correction 286

Table 38 Number of infants per time point and feeding routine considered for the final analysis. The data is show when considering the feeding at birth, and at each sample collection visit..... 288

Table 39 Odds ratios calculated based on abundance for N. lactamica, N. meningitidis, N. oralis and N. mucosa when considering breastfed infants at birth and at each sample collection visit. The odds ratio (Odds estimate) is presented with confidence intervals (Conf. int low and Conf. int High), the correspondent p-value, and the adjusted p-value using the Bonferroni correction..... 290

1. Chapter 1: Neisseria species and disease

1.1. Neisseria

The genus *Neisseria* is comprised of Gram-negative diplococci that belong to the class betaproteobacteria, and are usually considered commensal bacteria residing on various mucosal surfaces (1). Species such as *Neisseria meningitidis* and *Neisseria lactamica* colonize the oropharyngeal epithelium, whereas *Neisseria gonorrhoeae* inhabits the urogenital epithelium (2,3). It is believed that humans are the only reservoir known for several *Neisseria* species, such as *N. meningitidis* (1), although others, especially those not usually associated with disease, have been identified in animal hosts (4).

There is a persistent level of carriage of different neisserial species in the population, acting largely as commensals, but also as occasional "accidental" pathogens; *N. meningitidis* in particular in the paediatric population (5). Carriage studies have established that colonization with *N. meningitidis* starts during infancy, steadily increasing up to 15-19 year olds, when it can reach levels of 24-37%, followed by a decline in older age groups (6,7). Other commensals such as *N. lactamica* are also carried in the oropharynx, but contrary to *N. meningitidis*, carriage is maximal during infancy, then decreases (6). It is believed that *N. lactamica* carriage confers cross-reactive immunity against *N. meningitidis*, based on the shared similarity of their outer membrane proteins. (8), but it is also thought that it can prevent colonization of *N. meningitidis* (9,10).

The main factors contributing to *N. meningitidis* carriage are gender, age, and behaviours, such as smoking or frequenting crowded spaces (11,12). There is also an uneven distribution of serotypes and indicence rates across the word, the highest being in Africa (13) .Age is the only variable associated with carriage of *N. lactamica,* carriage being decreased with age (12). Several microbiome studies have reported correlations of *Neisseria* with different diseases such as cystic fibrosis (14), chronic kidney disease (15) or chronic rhinosinusitis (16).

1.2. Neisseria species identification

Identification of different *Neisseria* species has relied on the categorisation of certain characteristics of isolates such as acid production patterns from different carbohydrates, the ability to reduce nitrate, polysaccharide production, or their ability to grow on different types of media (1,17). However, identification of *Neisseria* using phenotypic methods are time consuming, and human errors in interpretation of the results obtained may lead to misidentification. Furthermore, certain isolates may not be identified due to their inability to grow under certain laboratory conditions (1).

Molecular tools have been developed to facilitate the identification of microbes at different levels of phylogenetic resolution. One of the most widely used gene markers to identify bacteria is the 16S rRNA gene. This gene is ubiquitous in bacteria and is comprised of conserved and variable regions. Sequencing of this gene has allowed the identification of a plethora of different new bacterial species and has also been used to identify clinical isolates. As not all bacteria can grow in laboratory

conditions, or do so very slowly, having an effective and fast method, such as sequencing the 16S rRNA gene, has proved to be an invaluable tool (18).

Other bacterial identification techniques, such as Multilocus Sequence Typing (MLST) of strains, extends the resolution obtained from the 16S rRNA gene, but requires sequencing seven conserved housekeeping genes. Although MLST was originally designed with the aim of identifying *N. meningitidis* isolates, it has also been applied to *N. gonorrhoeae* and *N. lactamica* (19). MLST is an allele-based method in which every sequence for each gene is assigned an allele number, and the combination of different alleles, called sequence types, can be used to identify different isolates. Sequence types can be then grouped based on their similarity, aggregating them in groups where at least six of the seven genes analysed are identical, these groups are referred to as clonal complexes. This method has been especially important for monitoring the epidemiology of *N. meningitidis* strains over several years and during outbreaks of disease (20-22).

Another method that provides a higher degree of identification than MLST is ribosomal MLST (rMLST), which allows greater resolution, but at the expense of increasing the number of loci needed to be analysed to 53 (23). By categorizing the changes in ribosomal protein genes present in an isolate, phylogeny can be resolved up to the strain level. This method is widely applicable to all bacteria, as the genes included in the rMLST scheme are present in the chromosome of all bacteria and provide sufficient diversity to distinguish isolates. The rise in whole genome sequencing (WGS) technologies has led to a reduced use of MLST and rMLST usage based on Sanger sequencing (24). However, these two methods have been adapted to extract information from whole genomes and are still used as epidemiological tools (25).

1.3. Use of NGS technologies to study the composition of bacterial comunities

16S rRNA gene sequencing approach combined with massively parallel sequencing platforms (also known as Next Generation Sequencing, NGS) has been used to study the bacterial composition of different ecological niches, also known as microbiomes (26). While it is not possible to sequence the entirety of the 16S rRNA gene, it is possible to sequence one or more of the variables regions in it, depending on the sequencing platform used, reaching phylogenetic resolutions up to the genus level. Sequencing of samples from different niches using this approach has revolutionized the knowledge and understanding of bacterial populations in relation to their habitats (27), and to human health and disease (28,29).

Changes in different bacterial populations have been linked to different disease outcomes. *Neisseria* in particular have been associated with asthma and allergic airway disease (30), chronic kidney disease (15), metabolic syndrome (31), and pneumonia (32). Decreases in *Neisseria* populations have also been related to caries formation during childhood (33). However, due to the limitation of sequencing length of NGS platforms, only fragments of the 16S rRNA gene can be sequenced, making it possible to know that the *Neisseria* genus can be correlated with certain diseases, but it is impossible to identify what species are responsible.

More recently the use of metagenomic approaches has been used to identify bacterial species without relying on the sequencing of the 16S rRNA gene. With this approach, all the bacterial genome content of a sample is sequenced, and each read obtained is assigned to a particular taxonomic level using different bioinformatic pipelines, such as Kraken (34). Such approaches provide greater

resolution than that obtained from sequencing 16S rRNA genes, but at the cost of increased experimental cost and complexity. For microbiome studies, samples can be multiplexed in a single library, which allows sequencing of several samples per experiment, at the cost of sequencing depth. Metagenomic experiments also suffer a similar but exaggerated trade-off, as they require a higher sequencing depth per sample to capture all the diversity present.

Metagenomic approaches have been used to characterize different *Neisseria* species present in human hosts. Lassalle *et al*. (35) used saliva samples from hunter-gatherers and traditional farmers from the Philippines and compared them to those of Western countries. They identified higher levels of *N. meningitidis* and *N. lactamica* in traditional farmers, while hunter-gatherer populations had higher levels of *Neisseria sicca* and other unspecified *Neisseria* species. Donati *et al.* (36) used patterns of single nucleotide polymorphisms to identify species present in samples collected and sequenced for the Human Microbiome Project (37). Using this approach, they suggested that different *Neisseria* species may appear predominantly in certain locations in the oral cavity. It was also suggested that this approach could be used to track specific strains in longitudinal analyses.

1.3.1. Sequencing of rplF as a Neisseria identification tool

New non-culture techniques have been developed to identify species of a given genus. Sequencing the *rpoB* gene together with flanking sequences has been used to identify species of *Acinetobacter* (38)*; dnaJ* has been proposed as a target for identifying species of alphaproteobacteria (39); and recently, *rplF* has been proposed as a target specifically for *Neisseria* (40).

Bennett *et al.* (40) developed the *rplF* amplicon sequencing method with the aim of creating a fast and reliable way of identifying different *Neisseria* species from isolated bacterial colonies. The *rplF* gene encodes the L6 protein that forms part of the 50S subunit of the ribsosome. It has been used to identify *Neisseria* species present in human oropharyngeal swabs cultured in Thayer-Martin agar plates, from which Gram-negative diplococci isolates were selected, boiled, and a PCR was performed using *rplF* primers to identify the colonies and assign a species to those matching a database of *Neisseria*-specific *rplF* sequences. This method has been used to identify commensal *Neisseria* species, and to identify *N. meningitidis* specifically from human samples (41,42). It has also been used in conjunction with other conserved genes and biochemical testing to differentiate two newly discovered *Neisseria* species that colonise sea lions (*Neisseria zalophi*, (43)) and artic geese (*Neisseria arctica*, (44)).

1.4. Non-pathogenic Neisseria species

Little is known about the prevalence of other neisserial species in relation to the upper respiratory tract microbiome (URTome). Two culture-based studies (1,3) have been carried out to estimate the carriage rate of non-pathogenic *Neisseria* commensals. Both studies concluded that 40% of the population may carry just *Neisseria perflava* (or *N. perflava*/*N. sicca* according to Nieto *et al* (3)), while around 90% carry the bacterium in combination with another *Neisseria* species. *N. mucosa, Neisseria flava* and *Neisseria cinerea* appear in 25-28% of the population. These species are not usually associated with disease, although there are isolated case reports of infections, usually in immunosuppressed patients or individuals that have undergone surgery (38,45,46).

The consequences, beneficial or adverse, of non-meningococcal neisserial carriage are undefined for the majority of species. On the one hand, the role in engendering cross-protective immunity to meningococcal infection proposed for *N. lactamica* may also apply to these other common commensals. Muzzi *et al.* (45) indicated that the three recombinant proteins included in Bexsero could be found in at least one other neisserial commensal. The fHbp antigen was present in *N. cinerea, N. gonorrhoeae* and *Neissseria polysaccharea.* NHBA homologues were found in *N. lactamica, N. polysaccharea, N. gonorrhoeae* and *Neisseria flavescens*, and NadA homologs were only identified in *N. cinerea* (45). However, due to the lack of non-pathogenic neisserial genomes sequenced, it may be possible for these antigens to be present in other neisserial species not identified by Muzzi *et al*. Regarding *por*A, although the gene expression is thought to be exclusive to *N. meningitidis,* it is possible that other shared outer membrane proteins may be included in the vesicles containing PorA (46). On the other hand, it has been speculated that commensal neisserial strains could act as reservoirs for transfer of antimicrobial resistance and virulence genes to other *Neisseria* (47). Recent studies have also identified similarities in iron acquisition system (48,49), and capsule genes in non-pathogenic *Neisseria* when compared to those in *N. meningitidis* (50), further suggesting the possibility of *N. meningitidis* isolates acquiring potential virulence factors from other *Neisseria* species.

Neisserial carriage studies so far have been, in their majority, culture based. These studies are very labour intensive, and usually fail to identify more than one strain of the same species. Most studies have focused on *N. meningitidis* and *N. lactamica*, and none have looked at other commensal neisserial species on a large scale. However, the use of *rplF* as a speciation tool has allowed the expansion of carriage studies to focus on multiple *Neisseria* species. One such study was published by Diallo *et al.* (51). This study used the *rplF* sequencing approach to identify *Neisseria* isolates obtained by culturing human oropharyngeal swabs from over 7000 participants across over seven African

countries. Their analysis showed that species such as *N. polysaccharea, N. subflava* and a potential new species, denominated as *Neisseria bergeri* (52), are part of the upper respiratory tract *Neisseria* population throughout life.

1.5.N. meningitidis

N. meningitidis is the causative agent of invasive meningococcal disease. In particular strains of serogroup B, are responsible for the majority of bacterial meningitis and septicaemia cases in Europe since the introduction of serogroup C meningococcal conjugate vaccines in 1999 (53). It is estimated that 500,000 cases occur every year worldwide; 10% of cases result in death, and 4-11% of the survivors are left with major long-term sequelae (53).

There are 12 identified serogroups, but the five most prevalent are A, B, C, W-135 (from here on referred to as W) and Y are responsible for the majority of the cases worldwide, while a new emerging serogroup, denoted X, has been identified predominately in Africa (54). The distribution of these serogroups changes geographically. Serogroups B and C are the most prevalent in Europe, Australia and South America. B, C and Y are the most prominent in North America. In Asian countries, serotype A and C are the most abundant, while in African countries A, C, X and W are the most prominent (55). However, the introduction of different vaccines has led to surge in other less common in countries, such as Y and W in European countries (56) (Figure 1).

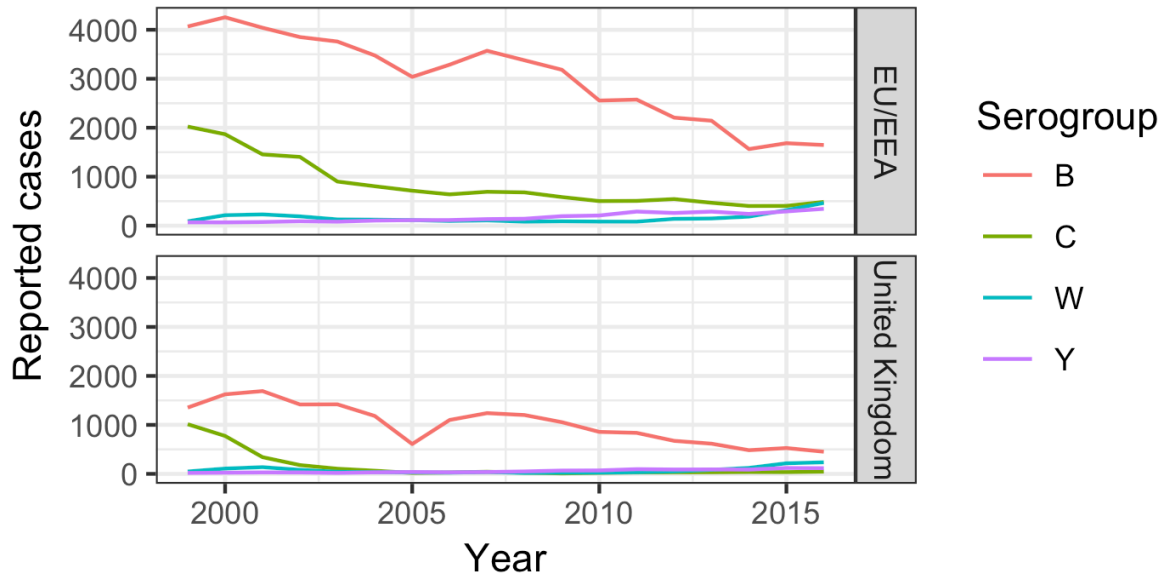


Figure 1 Number of reported cases of invasive meningococcal disease per serogroup in the European Union and in the United Kingdom (Y axis) from 1999 to 2016 (X axis). Data obtained from the European Centre for Disease Control.

The latest data collected form the European Centre for Disease Control (from 2016, as of July 2018), places the UK as the fifth country in numbers of cases of invasive meningococcal disease, with and age standardized rate of 1.25 per 100.000 people for confirmed cases (Figure 2). In the UK in 2016, there were 1647, 46, 116 and 236 reported cases of invasive meningococcal disease caused by serogroups B, C, Y and W, respectively.

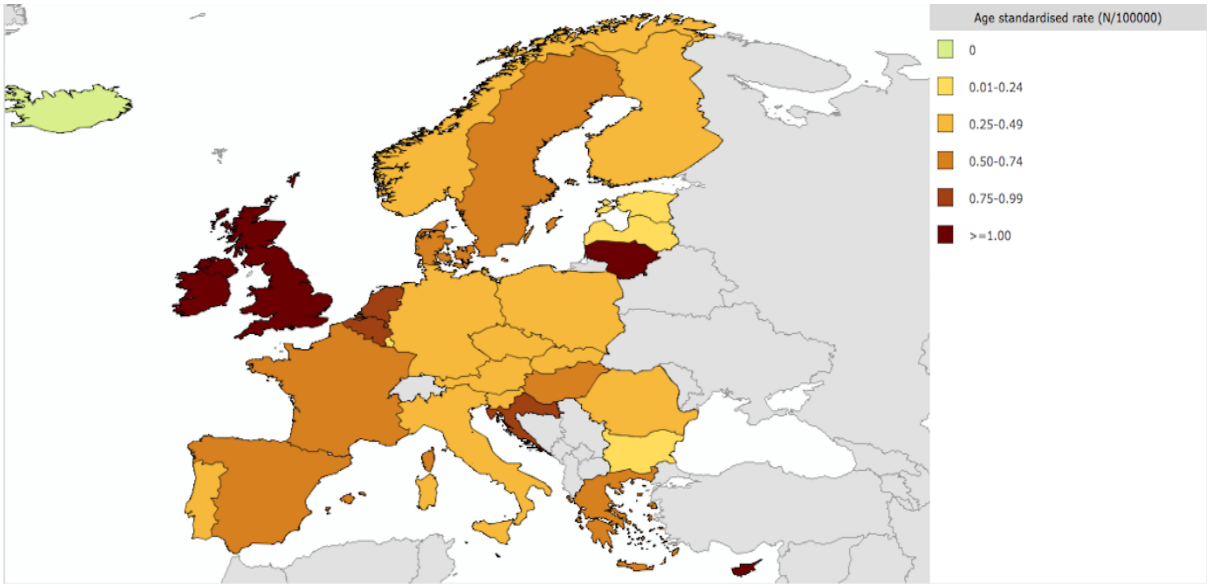


Figure 2 Age standardized rate of invasive meningococcal disease in European countries in 2016. Darker colors represent a higher incidence. Figure obtained from the European Centre for Disease Control atlas.

1.5.1. Meningococcal vaccines

Vaccines targeting the serogroup C meningococcus were introduced into the UK in 1999. The program initially targeted infants, administering three doses at 3, 4 and 12 months, with a catch-up program for those up to 5 years of age. Vaccines including serogroups A, C, W and Y polysaccharide conjugate are available to be administered to risk groups, and may be given routinely in countries where non-C serogroups pose a particular risk, such as those of serogroup A in the sub-Saharan meningitis belt in Africa (56,57).

Use of both the serogroup C and ACWY conjugate vaccines has led to a reduction in incidence of invasive meningococcal disease by the targeted serogroups, and both are cost-effective vaccines (57,58). High vaccination rates were followed by a decrease in incidence and transmission of *N.*

meningitidis not only among the vaccinated population, but also among those that remained unvaccinated, thanks to the herd immunity effect (59). However, the decrease in targeted serogroups has led to the rise in other, less abundant serogroups (Figure 1). With changing epidemiology, and the emergence of virulent serogroup W strains in the UK, a quadrivalent ACWY conjugate has been introduced into the adolescent vaccination schedule.

1.5.1.1. Meningococcus B vaccines: Bexsero®

Recently, a new vaccine, Bexsero®, also referred to as 4CMenB, was approved as a routine vaccine for infants in the UK, starting the summer of 2015. A vaccine targeting serogroup B has proven to more difficult, as the same strategy used for the development of vaccines for other serotypes (a conjugate vaccine using the capsular polysaccharide) could not be applied, due the structural similarity of *N. meningitidis* B capsule to polysialic acid present on human cells (60,61), making this molecule poorly immunogenic. It has also been hypothesised that a vaccine based on serogroup B capsule would lead to an autoimmune response (62) due to the aforementioned similarities.

1.5.1.1.1. Development of the vaccine

New approaches had to be considered to develop vaccines against serogroup B. One approach used was to exploit the natural release of outer membrane vesicles (OMVs) from Gram-negative bacteria. Neisserial OMVs contain a plethora of outer membrane proteins (63,64), among them PorA. One of the first uses of OMV-based vaccines occurred during an epidemic in New Zealand (1991-2007). While the vaccine used proved to be of great utility (65), it was not widely applicable to all *N.*

meningitidis serogroup B, isolates, as the antibodies generated by this vaccine were predominantly raised against the specific variant of PorA the isolates carried (66). Such vaccines have proved successful where a single strain (and PorA) can be targeted, *e.g*. Norway (67), Chile (68), and Cuba (69). However, typically, most countries have serogroup B strains circulating with many different PorA variants, and OMV vaccines would be of limited efficacy (70). Thus, the development of a broad-range meningococcal serogroup B vaccine arose as a necessity.

1.5.1.1.2. Bexsero® composition

Bexsero® was developed through the use of an approach coined reverse vaccinology (71). Through mining of the WGS of serogroup B strain MC58, potential outer membrane proteins were predicted, cloned, recombinant proteins produced, and their immunological functional activity carried out by performing serum bactericidal assays. Initially, 28 promising proteins were identified as potential vaccine candidates and finally, based on broad coverage, three were chosen for inclusion in a vaccine (72). These were fHbp, NadA and NHBA (Table 1). Two of these proteins, NHBA and fHbp, are fused to two other meningococcal proteins to improve their stability. The final formulation of the vaccine includes the PorA-containing OMVs used in the New Zealand outbreak, expected to provide additional coverage to PorA P1.4 strains as well as for adjuvant properties.

Table 1 Protein components in Bexsero®. GNA denotes the genome-derived *Neisseria* antigen code given to each gene. GNA numbers denote the code given to genome-derived neisserial antigens.

Antigen	Abbreviation	GNA number	Fused to other proteins
Heparin Binding Antigen	NHBA	GNA2132	Yes, GNA1030
Factor H Binding Protein	fHbp	GNA1879	Yes, GNA2091
Neisseria adhesin A	NadA	GNA1994	No

1.5.1.1.3. Expected impact and coverage of Bexsero® antigens

Bexsero® induces bactericidal antibodies, but several studies have identified a waning of antibody titres in infants (73,74) and adolescents (75) over time. One of the main points of contention when considering the implementation of this vaccine in the UK was the cost per dose. Due to its composition, prices per dose are higher than other capsule-based meningococcal vaccines. In the UK, the Joint Committee on Vaccination and Immunization released a statement in which it was recommended a vaccination schedule at 2, 4 and 12 months (76). In this document the importance of cost-effectiveness was highlighted, and a recommendation made that targeted carriage studies to further assess the impact of Bexsero® vaccination were required.

Despite the development of Bexsero® as a broad-range vaccine, it still remained necessary to prove its effectiveness. To identify predicted coverage of Bexsero® in a given country, the Meningococcal Antigen Typing System (MATS) was developed. This assay tests the efficacy of the Bexsero® antigens to create antibodies to target the NHBA, fHbp and NadA antigens of circulating strains by comparing their effect to those from reference stains in an ELISA assay. Additionally, the

PorA serosubtype of strains is determined and compared to that in the New Zealand OMV component, *i.e.* P1.4 (77) by Sanger sequencing. It is considered that as long as antibodies can target at least one of the antigens in the vaccine, it will be effective. This approach, although effective, has proven to be too conservative, and the results obtained are likely an underestimate of true coverage (78,79).

Parikh *et al*. (80) identified that 66% of isolates circulating between 2014 and 2015 in England, Wales and Northern Ireland were positive according to MATS, while 73% were positive during the 2007-2008 period. Abad *et al*. (81) published a similar study with isolates from Spain, collected between 2009 and 2010, finding 68.67% of isolates were MATS-positive. However, both publications state that the variation observed in coverage is generally attributed to changes in the circulating clonal complexes identified. Both of these publications highlight the importance of country-specific, and continuous antigen coverage assessment to assess the effectiveness of the vaccine.

Bexsero® was introduced in the UK vaccination schedule in the summer of 2015. The initial vaccination achieved a high coverage of eligible infants (99.5 and 88.6 % by the first and second dose, respectively), and it was estimated that the effect of the first two doses of the vaccine led to a decrease of half of the invasive meningococcal disease cases, when compared to values from previous years (82). Cost-effectiveness studies also estimated that vaccination with Bexsero®, and a catch-up program would be cost effective, but only as long as the vaccine could be supplied at low prices, and providing the vaccine could have an impact on meningococcal carriage (83,84).

1.6. Motivations behind the analysis

A reduction in carriage of *N. meningitidis* serogroup B after Bexsero® vaccination could lead to the herd immunity effect seen with other meningococcal vaccines, further increasing its cost-effectiveness. Although some studies have taken steps into measuring the potential effect of vaccination on carriage (85), there is no data on the potential reduction of meningococcal carriage in infants after receiving the vaccine.

It has been shown that Bexsero® can induce an antibody response against *N. meningitidis* serogroup B, but it has also been shown that the antigens present in Bexsero® are present in strains of other *Neisseria* commensals such as *N. lactamica, N. polysaccharea, N. cinerea* and *N. gonorrhoeae* (45). Thus, antibodies elicited by Bexsero® vaccination could potentially cross-react with other *Neisseria* commensals, altering their carriage, and possibly disrupting 'normal' neisserial carriage.

Bexsero® also includes the PorA P1.4 antigen which is imbedded in the OMV component. Proteomic analysis of the composition of neisserial OMVs found a plethora of proteins present in these vesicles (63,64,86), some also present in *N. lactamica* OMVs. Apart from a potential effect on *Neisseria* carriage, Bexsero® could also have an impact on the upper respiratory tract microbiome due to antibodies targeting the proteins in the OMVs.

Colonization by a particular species could be viewed as a prerequisite of infection, and understanding how certain life events, especially during infancy, may alter the likelihood of carriage can prove to be an invaluable tool to help decrease incidence of a disease. Several events have been

associated with an increase in *N. meningitidis* and *N. lactamica* carriage. Examples of this are living in crowded spaces (87), the use of pacifiers (88), or being exposed to cigarette smoke (89). However, very little is known about the effect of these, or other, effects on other *Neisseria* species.

1.7. Hypotheses

- The *rplF* sequencing method can be applied to a NGS sequencing platform to obtain high-quality sequence reads.

- Vaccination with Bexsero® has an effect on the infant upper respiratory tract microbiome.

- Vaccination with Bexsero® reduces the carriage of *Neisseria meningitidis*.

- Vaccination with Bexsero® alters the composition of *Neisseria* species colonizing the upper respiratory tract.

- Certain early life events can increase the carriage of particular *Neisseria* species.

1.8.Aims

- Validate the use of *rplF* to distinguish *Neisseria* species using bacterial isolates and human samples.

- Implement the *rplF* sequencing method in the Illumina MiSeq platform and develop reliable library preparation protocol for *rplF* sequences.

- Develop a bioinformatic pipeline to process *rplF* sequencing reads.

- Identify any potential differences in the upper respiratory tract microbiome caused by vaccination with Bexsero®

- Identify potential changes in *Neisseria* species populations after Bexsero® vaccination.

- Characterize changes in *Neisseria* species caused by early life events.

2. Chapter 2: Materials and Methods

2.1. List of reagents used

ITEM	SUPPLIER
DNA 1000 kit	Agilent
DNA high-sensitivity kit	Agilent
Adhesive sealing sheets	Applied Biosystems
Microamp optical 96-well reaction plates	Applied Biosystems
Microamp optical 96-well reaction plates with barcode	Applied Biosystems
Microamp optical adhesive film	Applied Biosystems
PowerUp SYBR Green master mix	Applied Biosystems
Gel cutting tips	Axygen
Brain heart infusion (BHI) broth	BD
AMPure XP magnetic beads	Beckman Coulter
JetSeq library quantification LO-ROX kit	Bioline
Sterile Petri dish	Celstar
Cotton-tipped, flocked, dual-headed swabs	COPAN
Black, flat bottomed assay plate	COSTAR
DNA LoBind tubes (2/5 ml)	Eppendorf
PCR primers	Eurofins
MiSeq 2x300bp sequencing kit	Illumina
1 kb plus DNA ladder	Invitrogen
E-gel dock	Invitrogen
E-gel sizeselect 2% agarose gel	Invitrogen
Quant-it PicoGreen dsDNA assay kit	Invitrogen
TOPO-TA cloning kit	Invitrogen
Tris-EDTA (TE) buffer	Invitrogen
KAPA quantification kit for Illumina platforms	KAPA
DNAzap DNA degradation solutions	Life Sciences
Fastdna spin kit for soil	MP
Molecular grade BSA	New England Biolabs
Q5 high-fidelity 2x master mix	New England Biolabs
Blood agar	Oxoid
Horse blood	Oxoid
Molecular grade water	Qiagen
Plasmid mini kit	Qiagen

QiaQuick gel extraction kit	Qiagen
Taq PCR core kit	Qiagen
Faststart HiFi PCR system dNTP pack	Roche
PCR grade water	Roche
Sterile plastic tubes 15/50 ml	SARSTEDT
Plastic tubes 2 ml	SARSTEDT
Betaine	Sigma-Aldritch
Foetal Bovine Serum	Sigma-Aldritch
Glycerol	Sigma-Aldritch
Sodium hydroxide solution (10 M)	Sigma-Aldritch
2-propanol	Sigma-Aldritch
Agar	Sigma-Aldritch
Agarose	Sigma-Aldritch
Trizma base	Sigma-Aldritch
1000/200/10-20 µl	Starlab
Screw-capped plastic tubes	Starlab
Sterile 25 ml reservoir	Thermo-Fisher
Ethanol absolute	VWR
Graduated Pasteur pipettes	VWR

2.2. List of equipment used

ITEM	MODEL	COMPANY
Electrophoresis gel tank	E-Gel dock	Invitrogen, Ashford, UK
	Wide Mini Ready	Bio-Rad, Hertfordshire, UK
Sequencer	MiSeq	Illumina, Cambridge, UK
	454 GS-FLX Titanium	Roche, Basel, Switzerland
Fluorescence plate reader	FLUOstar Omega	Eppendorf, Stevenage, UK
Polymerase chain reaction thermocyclers	Veriti	Invitrogen, Ashford, UK Kingdom
Centrifuge	Eppendorf 541R	Eppendorf, Stevenage, UK
Homogeniser	FastPrep-24	MP Biomedicals, Santa Ana, U.S.A.
Laboratory oven	Hybridization oven	Thermo Scientific, Paisley, UK
Nanodrop spectrophotometer	Nanodrop ND1000	Thermo Scientific, Paisley, UK
Real-time polymerase chain reaction thermocyclers	StepOne-Plus	Applied Biosystems, Ashford, UK
	ViiA-7	
Bioanalyser	Bioanalyzer 2100	Invitrogen, Ashford, UK
Gel imaging systems	Gel Doc EZ system	Bio-Rad, Hertfordshire, UKUK
	Ultraviolet trasniluminator	Genetic Research Instrumentation Ltd, Dusmow, UK

2.3. Reference sequences for rplF species identification

Reference sequences used to assign species to sequenced *rplF* amplicons were downloaded from the *Neisseria* PubMLST database (https://pubmlst.org/neisseria/). Reference sequences and IDs listed in Table 2 were used as reference for all analyses during this project. Of the 196 sequences available, only 164 have an assigned species. For this project only, the profiles with an assigned species were included in the final data base.

Table 2 Reference *rplF* amplicon sequences obtained from the PubMLST *Neisseria* database. The first column shows the species included in the database, the ID numbers assigned to the sequences and the total number of sequences for each species. IDs marked with '*' refer to sequences without a definite species assignment, while '**' identifies sequences with more than one species assignment.

Species	ID	Total
Neisseria meningitidis	1, 2, 3, 4, 8, 18, 64, 66, 67, 88, 91, 98, 106, 107, 108, 109, 110, 122, 123, 124, 125, 126, 127, 129, 132, 133, 136, 149, 153, 159, 161, 167, 169 ,170, 171, 172, 173, 174, 175	39
Neisseria lactamica	6, 32, 33, 34, 51, 52, 65, 80, 81, 85, 86, 89, 92, 93, 94, 111, 113, 144	18
Neisseria subflava	11, 12, 23, 25, 31, 38, 40, 42, 43, 53, 56, 57, 58, 59, 71, 72, 82*, 95, 97, 99, 100, 128, 130, 131, 142, 147, 148, 156, 181, 182, 183, 184, 185, 186	34
Neisseria polysaccharea	9, 39, 44, 63, 90, 96, 120, 160	8
Neisseria mucosa	13, 14, 17, 21**, 22, 27, 28, 29, 30, 35, 41,54, 87, 119, 177, 178, 179	17
Neisseria cinerea	10, 19, 20, 45, 73, 74, 121, 158, 180, 187	10
Neisseria bergeri	16, 62, 69, 70, 112*	5
Neisseria oralis	21**, 26, 36, 68, 102	5
Neisseria elongata	15,37,60,75	4
Neisseria animalis	76	1
Neisseria animaloris	192	1
Neisseria bacilliformis	46,49	2
Neisseria canis	48,196	2
Neisseria dentiae	47	1
Neisseria dentrificans	188	1
Neisseria gonorrhoeae	5, 7, 83, 84, 157, 176	6
Neisseria musculi	101	1
Neisseria shayegani	78	1
Neisseria wadsworthii	77	1

Neisseria weaveri	24	1
Neisseria zoodegmatis	168	1
Neisseria spp.	50, 79, 189, 190, 191	5
Bergiella dentrificans	55	1
Moraxella spp.	61	1

2.4. Reference sequences used for 16S rRNA gene taxonomy assignment

Reference sequences used for 16S rRNA gene OTU assignment were downloaded from the SILVA database (https://www.arb-silva.de). Release version 111 for QIIME was used for OTU assignment for all samples.

2.5. Samples used

Human infant throat swab samples used for this project were collected for the Dormice study (IRAS project ID: 105537, REC: 12/LO/1362), and from a Bexsero immunization study led by Professor Andrew Pollard at Oxford University, as part of the EUCLIDS project (90). For the Dormice study, samples from 204 healthy term infants were collected at the following time points: birth, 6 weeks, 6, 9, 12, 18, 24 and 36 months of age. Seventy-two of these samples also include a 3- and 4.5-month time points. Extensive metadata, including home visit questionnaires and notes collected from each participant's general practitioner were collected and included during the analysis as metadata. Samples were collected using dual-headed, cotton-tipped flocked swabs (COPAN diagnostics, USA), and were stored in a freezer at -80 ºC for less than 10 days before DNA extraction.

For the samples collected as part of the Oxford immunogenicity study described above (from here on referred to as the Oxford study), the collection strategy is described in Figure 3. Briefly, infants in Group 1 were given a dose of Bexsero® at 2, 4 and 12 months, while those in Group 2 had injections at 6, 8 and 13 months. All infants had an oropharyngeal throat swab taken at 6 and 13months of life. Samples were collected in Oxford using the same protocol as for the Dormice study. Swabs were sent frozen in six different batches and extracted using the same protocol as for the Dormice samples, described below.

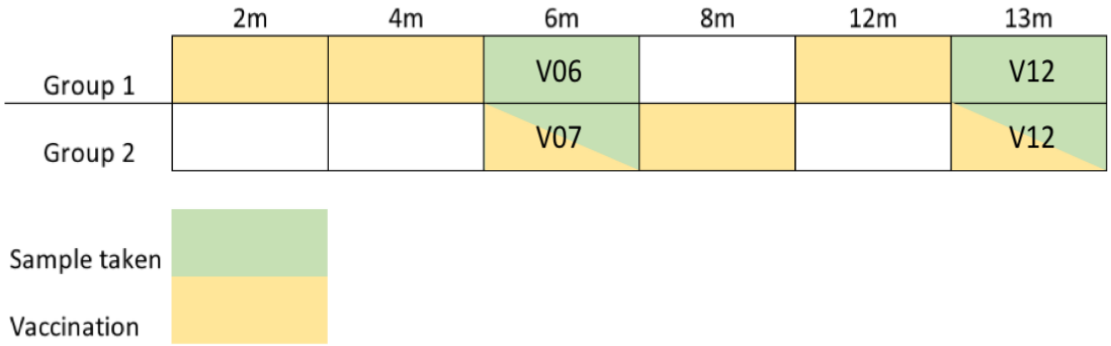


Figure 3 Vaccination strategy and sample collection for the Oxford study samples. Each column represents the month of life a sample was taken, or a vaccine given. The labels inside the boxes represent the codes given to each group of samples collected. Yellow sections represent a time point in which the vaccine was administered and green secion when a oropharyngeal throat swab sample was collected. Sections with both colors indicate time points where the vaccine was administered the same day a oropharyngeal swab was collected.

2.5.1. Adult swabs collected from healthy volunteers

Oropharyngeal swab samples were collected from 25 healthy adult volunteers using the same type of swabs as for the Dormice study. DNA was extracted within 24 hours using the protocol described below and kept at -80ºC until needed.

2.5.2. Bacterial isolates used as reference sequences

N. meningitidis (reference isolates MC58, 8013, Z2491), *N. mucosa, N. sicca, N. subflava, N. flava, N. polysaccharea, N. lactamica* were selected from the MIDG collection as reference isolates. Brain-Heart Infusion (BHI) media supplemented with 10% Foetal Bovine Serum (FBS) was used to culture all isolates.

2.5.3. Artificial DNA mixture samples

Artificial bacterial DNA mixtures were prepared to test the performance of the *rplF* PCR protocol. DNA obtained from the *Neisseria* isolates mentioned above were quantified using the Quant-iT PicoGreen dsDNA Assay Kit (Invitrogen). Each sample was diluted to 10000 genome copies/µl and were mixed in a total of 200 µl according to the following ratios: 0.01%/99.99%, 0.1%/99.9%, 1%/99%, 2%/98%, 5%/95%, 7.5%/92.5%, 10%/90%, 20%/80%, 40%/60%, 50%/50%, 60%/40%, 80%/20%, 90%/10%, 92.5%/7.5%, 95%/5%, 98%/2%, 99%/1%, 99.9%/0.01%, 99.99%/0.001%. An extra set of samples containing mixtures from three different species were prepared using the following ratios: 50%/25%/25%, 50%/40%/10%, 50%/10%/40%, 25%/50%/25%, 10%/50%/40%, 40%/50%/10%, 25%/25%/50%, 10%/40%/50, and 40%/10%/50%.

2.6. DNA extraction protocol

DNA was extracted from human oropharyngeal swab samples using the FastDNA SPIN Kit for Soil (MP). Extractions were performed as per the manufacturer's recommendations. In summary, 122 µl

of MT buffer and 978 µl of sodium phosphate buffer were mixed in a lysing matrix E (LME) tube containing a swab head. Samples were homogenized twice for 40 seconds using the speed setting 6 in the FastPrep-24 instrument, and between the two steps, samples were placed on ice for 30 seconds. Samples were centrifuged at 14,000 g and the supernatant was transferred to a 2 ml plastic disposable tube containing 250 µl of PPE buffer. This was mixed by inversion 10 times to promote protein precipitation. These tubes were centrifuged at 14,000 g for 5 minutes to pellet the precipitated protein component. The supernatant was collected and mixed with 1 ml of resuspended binding matrix solution in a 15 ml plastic conical tube, placed in a mixer for two minutes to enhance DNA binding to the matrix. The tubes were left standing for at least 3 minutes to allow the binding matrix to precipitate. After discarding excess supernatant, the binding matrix was resuspended in 1ml of the remaining supernatant plus binding matrix were resuspended, transferred to a 2 ml plastic tube and centrifuged at 14,000 g for two minutes. The supernatant was removed, and the binding matrix resuspended in 500 µl of SEWS-M. The SEWS-M mixture was transferred into the SPIN column and centrifuged for 1 minute at 14,000 g. Samples were allowed to air-dry for 5 minutes and were centrifuged again at 14,000 g for two minutes. The binding matrix was resuspended in 30 µl of low-EDTA TE buffer and placed in an incubator at 55ºC for 5 minutes. Samples were eluted by centrifuging at 14,000 g for 1 minute. The final purified DNA samples were transferred to labeled, rubber-sealed, screw-capped tubes. Finally, DNA quantity and purity were measured using NanoDrop. Samples were then stored at -80ºC.

2.7. Amplification of the rplF gene

The PCR reaction conditions used to amplify the *rplF* gene were the same as the ones published by Bennett *et al.* (40). Briefly, the QIAGEN Core PCR Kit was used as per the manufacturer's specifications, incorporating the published *rplF* gene primers (Forward primer: 5'-

CAGTGACTGTTCCCGCTGGTGT-3', Reverse primer: 5'-AGGYTCAGGAGKWCGGAAHG-3'). The *rplF* fragment was amplified for 35 cycles, each cycle consisting of 95ºC for 30 seconds, 55ºC for 30 seconds, and 72ºC for 1 minute. This reaction yields a 404-base pair (bp) long amplicon. PCR products were purified using 2% E-gel SizeSelect gels (Invitrogen). When amplifying *rplF* fragments from human oropharyngeal swabs, 5% of the water volume was replaced with DMSO (Roche). Modified versions of the Forward primer were also tested (5'-CAGTGACTGTTCCYGCTGGTGT-3' [rplF_Fm],5'-CAGTGACKGTTCCCGCTGGTGT [rplF_Fm2],5'- CAGTGACKGTTCCYGCTGGTGT-3' [rplF_Fm3]) to take account of possible sequence mismatches.

2.8. Chromatogram sequence analysis

Purified PCR products were sequenced using Sanger sequencing at the Imperial College London BRC Genomics Facility or at Genewiz UK Ltd. Samples were sequenced from the 3' and 5' end using the primers described above. The resulting AB1 chromatogram files were imported and visualized using Geneious (V9.1.3). Sequences obtained were aligned to *rplF* reference sequences using MAFFT (V1.3.6). Alignments were visually inspected and edited to ensure a correct alignment of all sequences. This alignment was used to create a Neighbour-Joining tree, using the Tamura-Nei model, bootstrapping the tree 1000 times. The consensus tree was constructed using a 70% support threshold for the branches. The resulting tree was visually inspected, and species were assigned to each of the sequenced samples based on the closest reference sequence.

2.9. Sequencing library preparation for 16S rRNA sequencing

To assess the bacterial diversity within oropharyngeal throat swab samples, the V3 to V5 regions of the 16S rRNA gene were sequenced. The V3 to V5 hypervariable regions were chosen as they have been shown to provide a great degree of taxonomic resolution compared to others (91) Roche 454 FLX compatible primers were designed using the Adaptor B for the Reverse primer (5'-CTATCCCCTGTGTGCCTTGGCAGTCTCAGCCTACGGGAGGCAGCAG-3') and Forward primers (5'-CCATCTCATCCCTGCGTGTCTCCGACTCAGNNNNNNNNNNNNCCGTCAA TTCMTTTRAGT-3', where the N track indicates a unique 12 bp long barcode sequence). These primers are described in Sim *et al.* (92). Each sequencing run included negative controls (negative controls included in different DNA extraction batches), and positive controls (samples previously amplified in other sequencing runs).

2.9.1. PCR amplification

To ensure no contamination of the PCR reactions, all were prepared in a UV-sterilized microbial safety cabinet. All equipment used during this process was treated with DNAzap DNA Degradation Solutions (Invitrogen). The reaction master mix, per sample, included 0.25 µl of FastStart HiFi polymerase, 1 µl of a 10 µM dilution of both Forward and Reverse sequencing primer, 0.5 µl of dNTP mix, 2.5 µl of FastStart 10x PCR buffer, 11.25 µl of molecular biology grade water, 1 µl of 20g/ml BSA and 6.5 µl 5 M betaine. To ensure a sufficient yield of PCR amplicons for all samples, each sample was amplified in quadruplicate. The reaction master mix was distributed to labelled 96-well plates, and 4 µl of DNA sample was added to each well replicate. Plates were sealed, mixed and centrifuged for 30 seconds at 4000 g. These were placed on Veriti thermal cyclers (Applied Biosystems) at 94ºC for 2 minutes, followed by 35 cycles at 94ºC for 20 seconds, 50ºC for 30 seconds and 72ºC for 5

minutes. Once completed, 5 µl of all negative and positive controls were visualized on an agarose gel and inspected for the presence of bands (positive control) or absence (negative controls).

2.9.2. Gel extraction and purification

After pooling all sample replicates into new 96-well plates, 70 µl of AMPure XP beads was added to each sample and mixed. After mixing, and a 10-minute incubation at room temperature, the plate was transferred to a magnetic stand for 5 minutes. The supernatant was discarded, and the beads were resuspended with 100 µl of 80% ethanol. Plates were placed in the magnetic stand for 5 minutes to separate the beads from the supernatant. After another ethanol wash cycle, the supernatant was discarded, and the pellets were air-dried for 5 minutes. Pellets were resuspended in 30 µl of low-EDTA TE buffer, incubated for 5 minutes at room temperature, and separated from the supernatant. The supernatants were transferred to a 96-well plate and the pellets discarded. Amplicons were purified by placing them in the top row of 2% E-Gel SizeSelect gels, while the bottom row was filled with 25 µl molecular grade water (and refilled as necessary). When the appropriately sized bands reached the bottom wells, the water in the bottom wells was extracted and placed in a new 96-well plate.

2.9.3. Quantification and pooling

Purified amplicons were quantified using Quant-iT PicoGreen protocol dsDNA Assay kit (Invitrogen) as per manufacturer's specifications. Each sample was quantified in duplicate. Finally, samples were pooled in equimolar amounts. The remaining library preparation and sequencing was performed at the DNA Sequencing Facility at Cambridge University.

2.10. Sequencing library preparation for rplF gene sequencing

Libraries were prepared following the 16S Metagenomic Sequencing Library Preparation protocols published by Illumina (https://support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_docfumentation/16s/16s-metagenomic-library-prep-guide-15044223-b.pdf), including a few modifications to accommodate the *rplF* gene. Two different versions of this protocol were developed to adapt this protocol to the *rplF* gene.

2.10.1. Protocol version 1

2.10.1.1. PCR amplification

Amplicons from the *rplF* gene were obtained using custom, MiSeq-compatible *rplF* primers. To ensure a sufficient amount of sample amplicon, each PCR was done in quadruplicate. A reaction master mix was prepared using Q5 High Fidelity 2X PCR master mix (New England Biolabs Inc.) and molecular grade water aliquot tubes (Cambio) to accommodate 18.5 µl of master mix per sample. The reaction master mix was distributed to the PCR plates, and 1.25 µl of each previously aliquoted MiSeq PCR primer (10 µM) was added to the wells, resulting in unique combinations for all of them. Finally, 4 µl of DNA sample was added to each well replicate. Plates were sealed, labelled, placed on a mixer and centrifuged for 30 seconds at 4000 g. These were placed on thermocyclers at 98ºC for 3 minutes, followed by 35 cycles of 98ºC for 30 seconds, 57ºC for 2 minutes and 72ºC for 30 seconds. After the PCR cycles, 5 µl of all negative and positive controls, plus 5 randomly selected samples from each

plate, were mixed with gel loading buffer, and pipetted into wells in a 1% agarose gel. Samples were electrophoretically separated and inspected for the presence of bands (positive control and samples) or absence (negative control).

2.10.1.2. Size selection and replicate pooling

All sample replicates were pooled into one single well in 96-well plates. A volume of 90 µl was transferred to a new round-bottomed 96-well plate. Each sample was mixed with 63 µl (70% mixture, v/v) of resuspended AMPure XP beads, incubated for 10 minutes at room temperature, and transferred to a magnetic stand for 5 minutes. The supernatant was discarded, and the magnetic bead pellet was resuspended in 100 µl of freshly prepared 80% ethanol, and the plates were placed in the magnetic stand again for 5 minutes. After another cycle of ethanol wash and resuspension, the plates were placed in the magnetic stand again, ethanol supernatant were discarded, and the pellets were left to dry for 5 minutes at room temperature. Pellets were resuspended in 30 µl of low-EDTA TE buffer, incubated for 5 minutes at room temperature, and placed back in the magnetic stand for 5 minutes. The supernatants were transferred to a new 96-well plate.

Samples were quantified using Quant-iT PicoGreen dsDNA Assay kit (Invitrogen) as per manufacturers specifications. Depending on the concentration of the samples, different volumes of each sample were taken to obtain a combined equimolar library pool. The library was mixed and centrifuged. A set volume of the library pool was transferred to a new tube, and it was concentrated to 30 µl using the AMPure XP protocol described above but applied to only one sample.

2.10.1.3. Gel purification

A 200 ml, 2% (W/V) agarose gel mixed with 5 µl of gel dye was prepared and poured in a deep gel tray and left to set. The 30 µl library was mixed with 10 µl of loading buffer and loaded into one of the gel wells. The library was electrophoretically separated and inspected, and the appropriately sized band was cut out from the gel. The QIAquick gel extraction kit (QIAGEN) was used to purify the extracted gel band. The gel band was weighed and mixed with three times its weight in volume of QG buffer. This was incubated at 50ºC for 10 minutes, inverting the sample every minute. One volume of isopropanol was mixed with the sample and centrifuged for 1 minute at 14.000 rcf and transferred to a QIAquick spin filter column. The sample flow-through was retained and passed through the filter again, to increase DNA retention yield. After the second centrifugation the flow-through was discarded. Another 500 µl were passed through the filter, followed by 750 µl of PE buffer, which was mixed for 5 minutes, then centrifuged for 1 minute at 14.000 rcf. The spin filter was placed in a new 1.5 ml plastic tube, and 30 µl of EB buffer was added to the filter. It was incubated for 4 minutes and centrifuged for 1 minute at 14.000 rcf. The remaining volume was transferred to a labeled, rubber-sealed, screw-cap plastic tube and kept at -80ºC until required.

2.10.1.4. Library quality control

The final library was quantified using the KAPA Library Quantification Kit for Illumina, as per manufacturer's specifications, using two-fold library dilutions from 1/1000 to 1/8000. After preparing the qPCR reactions, PCR plates were sealed, and fluorescence was read using the ViiA 7 Real-Time PCR system (Applied Biosystems). In parallel, library size was confirmed by identifying the library size in a Bioanalyzer (Agilent), using a High Sensitivity 1000 DNA Chip (Agilent), as per the manufacturer's specifications.

2.10.1.5. Preparation of library and PhiX spike

If the library quantification and size were satisfactory, the library was further processed to fit MiSeq specifications. First, a fresh 0.1N NaOH solution was prepared diluting a stock solution of 10 M NaOH (Sigma-Aldritch) using the Illumina-supplied HT1 buffer. Ten microliters of library were mixed with 10 µl of 0.1N NaOH and incubated for 5 minutes to denature the library. After incubation, the mixture was diluted with ice-cold HT1 buffer to stop the dehybridization process, and was further diluted as required, usually to a final concentration of 8 pM, using the HT1 buffer. The PhiX spike was prepared by mixing 2 µl of 10 M PhiX (Illumina) with 2 µl of EBT buffer (Illumina) and 3 µl of 0.1N NaOH, this mixture was also incubated for 5 minutes at room temperature before being diluted with 990 µl of ice-cold HT1 buffer. The denatured PhiX solution was further diluted as necessary, usually to 8 pM. The diluted library and PhiX were mixed in different proportions depending on the library being sequenced. For the majority of the libraries, an 80% library (800 µl) and 20% (200 µl) PhiX mixture prepared. Finally, 600 µl of the mixture were loaded into the sequencing cartridge, followed by the sequencing primers in the corresponding wells. This was loaded into the MiSeq and sequencing of the library was started as soon as possible.

2.10.2. Protocol version 2

An alternative protocol was also used. The main difference is that version 2 involves amplification of the *rplF* gene in a two-step PCR approach. The first step of the PCR required the same components as before but using a different set of primers that include the published *rplF* primers, plus a tag sequence used to attach the primers in the second PCR. The first step for this PCR was performed

using the same conditions and volumes as the one described for the previous protocol, but the PCR was performed for 20 cycles. After completion of the first PCR, the sealed plates were centrifuged for 30 seconds at 4000 g. These were placed on thermocyclers at 98ºC for 3 minutes, followed by 20 cycles of 98ºC for 30 seconds, 57ºC for 2 minutes and 72ºC for 30 seconds, plus a final extension at 72ºC for 5 minutes. Once finished, the replicate plates were pooled using AMPure XP beads as described before. A second round of PCRs was performed, but this time each sample was amplified only once and not in quadruplicate as described before. The PCR reaction was composed, per sample, of 25 µl of Q5 master mix, 21 µl of 2.5 pM primer mix and 4 µl of sample. The primers in this PCR contain the adaptor binding, barcode, and sequencing primer binding sequences. After all the samples were prepared, the PCR plates were sealed and centrifuged for 30 seconds at 4000 g. These were placed on thermocyclers at 98 ºC for 3 minutes, followed by 6 cycles of 98ºC for 30 seconds, 57ºC for 2 minutes and 72ºC for 30 seconds, plus a final extension at 72ºC for 5 minutes. The remaining steps of the library preparation and quality control remained the same, except that to try to further increase the quality of the reads obtained, and extra denaturation step was performed immediately before starting sequencing. Once the diluted library-PhiX mix was prepared, the library was placed in a heating block at 96ºC for two minutes, followed by a 5-minute incubation in ice-cold water. The library was loaded into the cartridge and proceeded immediately to sequencing.

2.11. Bioinformatics and statistics

Output files from either the 454 FLX sequencer (SFF files) or from the Illumina MiSeq (fastQ files) need to be processed in order to be able to assign reads to OTUs or sequence variants (SV). The MacQIIME distribution (V1.9.1) of QIIME (93) was used to process the output sequencing files, using the script necessary for each sequencing format. Once the sequencing files were processed, they were imported into R environment (V3.4.4) using the Phyloseq (94) package.

2.11.1. Sequencing file processing for Roche 454 SFF files (16S rRNA gene amplicons)

For each sequencing run a sample mapping file sheet, containing sample names and barcode combinations, plus other run related information, was generated. These files were converted into FlowClust (95) compatible files. This program was used for flowgram clustering, also referred to as denoising, and quality filtering of sample reads. Output files were scanned for chimeric sequences using USEARCH (V9.2.64) (96). Finally, the sequencing file was demultiplexed, OTUs were assigned to each read by aligning them to the SILVA 111 16S rRNA dataset using the UCLUST distribution bundled within QIIME, with a 97% similarity cut-off. A set of reference sequences was generated by selecting a representative sequence from each of the clusters generated. Finally, the resulting representative sequence set was aligned using MAFFT (97) and a rooted phylogenetic tree was generated using FastTree2 (98).

2.11.2. Sequencing file processing for Illumina MiSeq fastQ files (rplF amplicons)

Sample sheets containing information for each sample were generated for each run as necessary. These were used in conjunction with the corresponding multiplexed fastQ file to obtain quality filtered, demultiplexed files using of QIIME (Version 1.9.1). All the demultiplexed files obtained for 12 MiSeq runs were combined and clustered using the Minimum Entropy Decomposition algorithm (99). For each of the resulting nodes (clusters), representative sequences were selected, aligned using MAFFT (97)and a rooted phylogenetic tree was generated using FastTree2 (98).

2.11.3. Statistical analysis

All statistical test described were performed in the R environment (Version 3.4.4), a list of all R packages used can be found in Appendix 1. For all statistical tests, a p-value lower than 0.05 was considered significant, and the Bonferroni-holm method (100) was used to adjust p-values for multiple comparisons.

2.12. Real-time quantitative PCR for the rplF gene

To quantify the total amount of neisserial DNA in the samples studied, it was decided to translate the *rplF* PCR to a qPCR approach. To do this, the TOPO TA cloning kit was used to create a plasmid (rplfp) that could be used as a standard for quantification purposes. An *rplF* amplicon from a *Neisseria subflava* isolate amplified with the rplF_Fm primer was inserted into the TOPO TA cloning plasmid following the supplier's specifications. *Escherichia coli* competent cells were transformed with rplfp, and the recovered cells were grown on BHI agar overnight. The Plasmid Mini Kit (QIAGEN) was used to purify the plasmid.

This plasmid was used in ten-fold dilutions from 10^7 to 10^2 plasmid copies/µl to generate a standard curve for each qPCR plate used. All samples and standard dilutions were tested in triplicate. To obtain better quality results from the qPCR, the DNA extracted from oropharyngeal swabs was diluted using 2.5 µl of sample in 7.5 µl of water. The qPCR master mix consisted of 5 µl of PowerUp SYBRGreen MasterMix, 2 µl of sample, 0.6 µl of a 6.66 pM mixture of primers and 2.4 µl of water. Each qPCR plate was prepared immediately before quantification, which was performed in a StepOnePlus machine, using the same PCR conditions as the standard *rplF* PCR.

3. Chapter 3: Validation of rplF as a Neisseria speciation tool

3.1. Introduction

Isolation and identification of *Neisseria* species can involve a series of test, such as assessment of cell morphology and arrangement, biochemical tests (*i.e.* growth under different conditions and media (1), pigment production, hemolysis, acid production from sugars (101)), sequencing of conserved genes, such as MLST or rMLST schemes (23,102). However, identification of species may involve culturing isolates from specimens, which may induce bias in the results obtained, as not all species grow under the same conditions, and sequencing multiple loci per sample may not be economically or experimentally feasible.

Bennett *et al.* (40) described how by sequencing one ribosomal gene, *rplF*, amplified from pure bacterial cultures, it is possible to identify the species of a particular *Neisseria* isolate, obtaining results comparable to those from phylogenies constructed from sequences of 53 concatenated ribosomal genes. However, this article does not describe the applicability of *rplF* sequencing in complex mixtures of DNA.

This Chapter describes assessment of the validity of the *rplF* sequencing method, as proposed by Bennet *et al*. (40), to identify *Neisseria* species, and application to samples where the bacteria are not available, such as DNA extracted from human oropharyngeal throat swabs.

3.2. Aims

- Evaluate the validity of *rplF* amplicon sequences to identify *Neisseria* species.
- Validate the use of this method when applied to DNA extracted from human oropharyngeal throat swabs.
- Investigate potential issues this technique may have when applied to complex mixtures of DNA.

3.3. Summary of methods

3.3.1. Samples used and DNA extraction protocol

3.3.1.1. Adult oropharyngeal throat swabs

Oropharyngeal throat swab samples were collected from 25 healthy adult volunteers using cotton-tipped swabs. After DNA extraction, samples were stored at -80ºC for less than 24 hours.

3.3.1.2. Bacterial isolates

N. meningitidis (reference isolates MC58, 8013, Z2491), *N. mucosa, N. sicca, N. subflava, N. flava, N. polysaccharea,* and *N. lactamica* isolates were selected from the MIDG collection as reference

isolates. BHI media supplemented with 10% FBS was used to culture all isolates. All isolates were cultured at 37 ºC in a 10% CO_2 atmosphere.

3.3.1.3. DNA extraction protocol

These samples were extracted using the FastDNA Spin Kit for Soil, using the protocol described in section 2.6.

3.3.2. PCR amplification of the rplF gene

The PCR amplicons were obtained following the protocol described in (40). Amplicons were electrophoretically separated and purified using a 2% agarose E-gel (Invitrogen). The samples were sequenced at the Imperial College BRC Genomics facility.

3.3.3. Analysis of Sanger sequencing chromatograms

Sequencing files obtained were aligned using MAFFT (97), and a neighbour-joining tree using the Kimura 2-parameter model was reconstructed using MEGA7. Trees where further analysed and modified in the R environment (V.3.4.4) using the ggtree package (V1.12.0). A species was assigned to each read obtained based on the nearest neighbouring reference sequence or based on the closest BLASTn match (V2.2.26).

3.3.4. Bioinformatic analysis of genomic sequences

Reference *Neisseria rplF* sequences were downloaded from the PubMLST database, and genomic sequences were downloaded from Genbank. The files obtained were imported into Geneious for further analysis. ANI scores for selected genomes where calculated using the average_nucleotide_identity script from the pyani package (V0.2.7) (103) in the Python 3.6.4 environment, using MUMmer (V3.32) to calculate pairwise distances in between genomes.

3.4. Results

3.4.1. Reproducibility of rplF sequencing, and testing using bacterial genomic DNA and that extracted from oropharyngeal throat swabs

To assess the reproducibility of this method, all available *rplF* reference sequences were downloaded from the PubMLST database, aligned, and a dendrogram created to visualize their similarity. To help visualize results, the most divergent sequences (*N. animalis, N. animaloris, N. dentiae, N. musculi, N. canis, B. denitrificans, Moraxella sp., N. baciliformis, N. shayeganii, N. wadsworthii, N. waeveri* and *N. zoodegmatis*, plus one sequence labeled as "*N. oralis* or *N. mucosa"*) were excluded. Results can be seen in Figure 4. Similar to the results shown in Bennett's publication (40), species cluster together in single-species clades, allowing the differentiation of individual species, with the exceptions of *N. bergeri*, *N. oralis* and *N. mucosa*, all of which are spread across 2 clusters, and *N. cinerea* which appears in three different clades.

It is also important to highlight that there are reference sequences assigned to non-*Neisseria* species, these being *Bergeriella denitrificans* and *Moraxella* spp. *B. denitrificans* is part of the *Neisseriaceae* family and used to be included in the *Neisseria* genus *as Neisseria denitrificans* (104), which may explain why the primers designed for *Neisseria* can amplify the gene from this genus. On the other hand, *Moraxella* belongs to a completely different family, *Moraxellaceae*. Nevertheless, as *rplF* is a ribosomal gene, some degree of conservation of its sequence might be expected. *Moraxella* spp. cluster groups closely with a set of unidentified *rplF* sequences (reference ID: 135, 137, 139, 140, 143). Querying these sequences using BLAST against the NCBI non-redundant (nr) database shows that they most likely belong to *M. catarrhalis.*

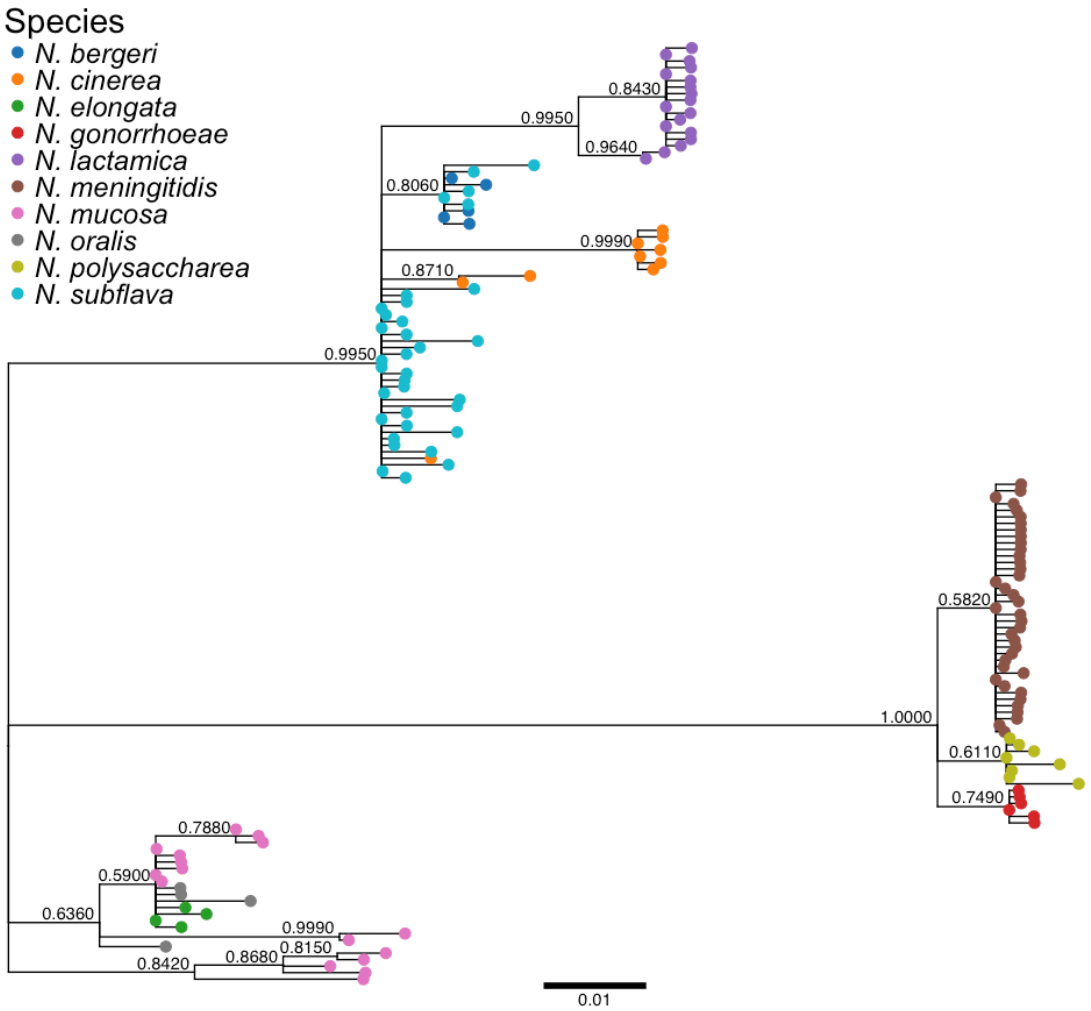


Figure 4 Dendrogram illustrating the differences among *rplF* reference sequences. Each species is colour-coded as shown in the legend, and the black bar at the bottom of the Figure shows a distance equivalent to a 1% difference among sequences (4.13 bp difference). The values on the tree branches indicate the consensus support of each branch.

The capability of the *rplF* sequencing method to separate species relies on generating dendrograms from *rplF* amplicon alignments, using the Kimura two-parameter model of evolution (105). However, on the website where this database is hosted (https://pubmlst.org/bigsdb?db=pubmlst_neisseria_seqdef&page=sequenceQuery), sequence is matched based on sequence similarity using BLAST (106). This approach could introduce some variability in species assignment as BLAST and the Kimura model rely on different algorithms to identify similar sequences. On the one hand, phylogenetic tress generated by the Kimura model scores dissimilarities based on the two parameters considered, *i.e.* transition and transversion rates among

nucleotides. On the other hand, BLAST considers identities, mismatches and gaps to calculate score values for DNA sequences but scores all nucleotide changes are scored the same. When aligning *rplF* sequences no gaps are introduced in the alignment, so only identities and mismatches will be relevant in further experiments. An example of an alignment of *rplF* sequences is shown in Figure 5.

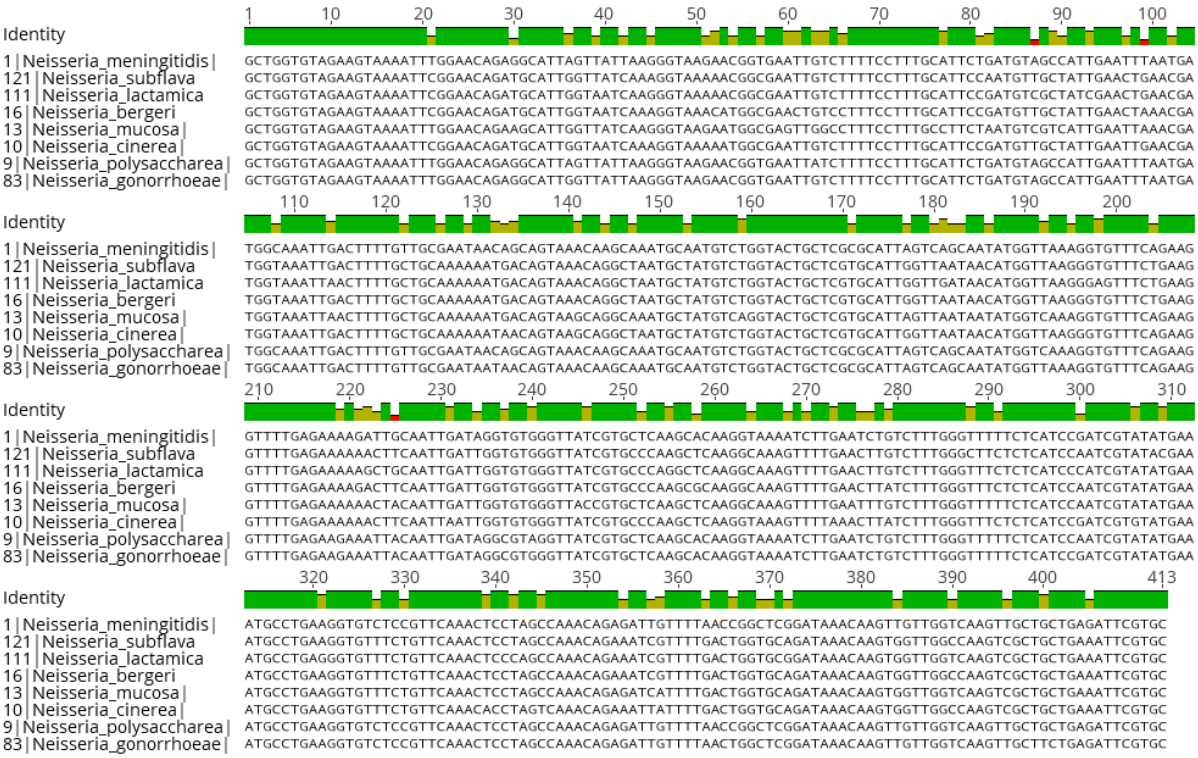


Figure 5 Alignment of *rplF* amplicon sequences. The alignment was constructed using MAFFT. Eight *rplF* reference sequences (ids: 1,121, 111, 16, 13, 10, 9 and 83) were used as representative sequences. The coloured bar above the sequence alignment indicates the level of conservation of each base, with higher values represented by taller green segments, less conserved bases in yellow, and the most variable bases by short red segments. The numbering of the bases is shown above the coloured bars.

To assess the validity of BLAST to correctly classify *Neisseria* species, the number of similarities in between sequences was further examined. A summary of the distances among *rplF* amplicon sequence, measured in bp differences, is shown in Table 3. The majority of the species included have a variation within sequences from the same species below 10 bases, with the exception of *N. mucosa*, which is expected, as it has been shown previously that the sequences from this species cluster in different groups (52). Some of these species, such as *N. animalis* and *N. animaloris,* had a value of 0 for their dissimilarities as they only have one reference sequence. *N. bergeri, N. cinerea, N. elongata,*

N. gonorrhoeae, N. meningitidis, N. mucosa, N. oralis, N. polysaccharea, all of them considered human commensals or pathogens, had an average distance to its nearest neighbour less than 16 bp. These sequences formed clusters of similarities. One is formed by *N. elongata, N. oralis* and *N. mucosa*, the second by *N. subflava, N. bergeri* and *N. cinerea*, and the last one by *N. polysaccharea, N. meningitidis* and *N. gonorrhoeae.* All species in these groups are closely phylogenetically related (40).

Table 3 Summary of the dissimilarities in between *rplF* amplicon sequences measured in bp differences. This Table excludes rplF reference sequences without a species assigned, or those labelled as *Neisseria* sp. The Table shows the average distance among sequences from the same species (intraSD), the average distance with the other species (interSD), the species with the closest distance (nearest group), and the average distance to the nearest group, which shows the average number of bp differences to the reference *rplF* sequences belonging to the closest species.

Species	IntraSD	InterSD	Nearest group	Average difference to nearest group
B. denitrificans	0	77.06	*N. animalis*	71.18
Moraxella sp.	0	175.48	*N. subflava*	170.5
N. animalis	0	67.51	*N. subflava*	56.74
N. animaloris	0	85.07	*N. cinerea*	80.5
N. bacilliformis	6	119.84	*N. polysaccharea*	112.63
N. bergeri	2.8	43.41	*N. subflava*	**10.2**
N. canis	8	103.26	*N. wadsworthii*	80
N. cinerea	9.33	45.30	*N. subflava*	**15.01**
N. dentiae	0	91.57	*N. musculi*	64
N. denitrificans	0	70.87	*N. subflava*	59.44
N. elongata	3.17	43.75	*N. mucosa*	**13.81**
N. gonorrhoeae	1.67	47.49	*N. polysaccharea*	**7.33**
N. lactamica	3.58	48.49	*N. subfalva*	18.08
N. meningitidis	2.67	59.02	*N. polysaccharea*	**7.84**
N. mucosa	18.04	48.04	*N. oralis*	**12.85**
N. musculi	0	90.21	*N. dentiae*	64
N. oralis	6	42.18	*N. elongata*	**6.56**
N. polysaccharea	3.07	48.47	*N. gonorrhoeae*	**7.33**
N. shayeganii	0	43.56	*N. subflava*	110.59
N. subflava	6.54	48.01	*N. bergeri*	**10.2**
N. wadsworthii	0	104.84	*N. canis*	78
N. weaveri	0	102.87	*N. zoodegmatis*	90
N. zoodegmatis	0	76.98	*N. mucosa*	67

3.4.2. Description of the rplF amplicon

In Bennett *et al.* (40), the *rplF* amplicon used to identify *Neisseria* species is described as being 413-bp. However, this amplicon described overlaps with 8 bases at the 3' end of the Forward primer binding sequence, and with 1 base at the 5' end of the Reverse primer sequence. If only the region between the primer binding sequences is considered as the amplicon, the total length is reduced to 404 bp. A graphical representation is displayed in Figure 6. The *rplF* gene sequence is represented by a green bar, the *rplF* amplicon described by Bennett *et al.* (40) is in blue, and the amplicon described here in yellow. The primer binding sequences (rplF_F for the Forward and rplF_R for the Reverse primers) are represented by red fragments. Although it is unclear why Bennett et al (39) determined the amplicon to be 413bp, trimming the amplicon the 404 bp proposed will not have an impact in the sequence diversity of this amplicon, and therefore, in the effectiveness of species identification, as the 9 bases that differ are conserved in all sequences, with the exception of two bases. The variation of these two bases is further characterized in Section 3.4.4.1.

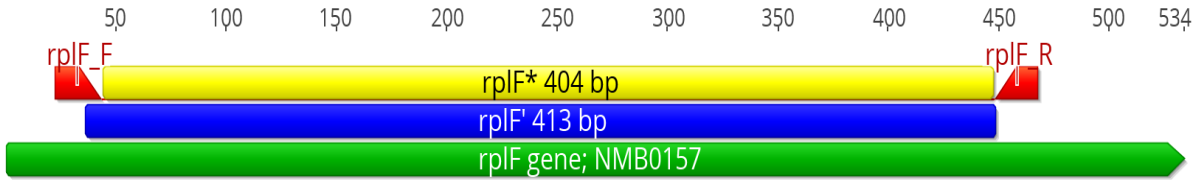


Figure 6 Graphical representation of the primer binding sequences to the *rplF* gene. The green segment represents the *rplF* gene from *N. meningitidis* MC58. The blue bar the *rplF* amplicon described by Bennett *et al.*, the yellow bar the amplicon proposed in this manuscript and the red segments the binding sites of the *rplF* forward (rplF_F) and reverse (prlF_R) primers.

3.4.3. Identification of Neisseria isolates from pure cultures and human oropharyngeal swabs.

To further characterize the validity of the *rplF* sequencing method, *N. meningitidis* (reference isolates MC58, 8013, Z2491) and two isolates each of *N. mucosa, N. sicca, N. subflava, N. flava, N. polysaccharea, N. lactamica* were selected for analysis. Isolates were cultured on BHI agar plates supplemented with a 10% FBS overnight, and 2-3 colonies were selected for DNA extraction. Amplicons from the *rplF* gene were purified and sequenced. Furthermore, oropharyngeal swab samples were collected from 25 adult volunteers, and DNA extracted. Fragments from the *rplF* gene were also obtained from these samples, purified and sequenced. The sequences were aligned alongside the reference *rplF* sequences and a dendrogram was generated to identify the species of each of the samples included. The results can be seen in Figure 7.

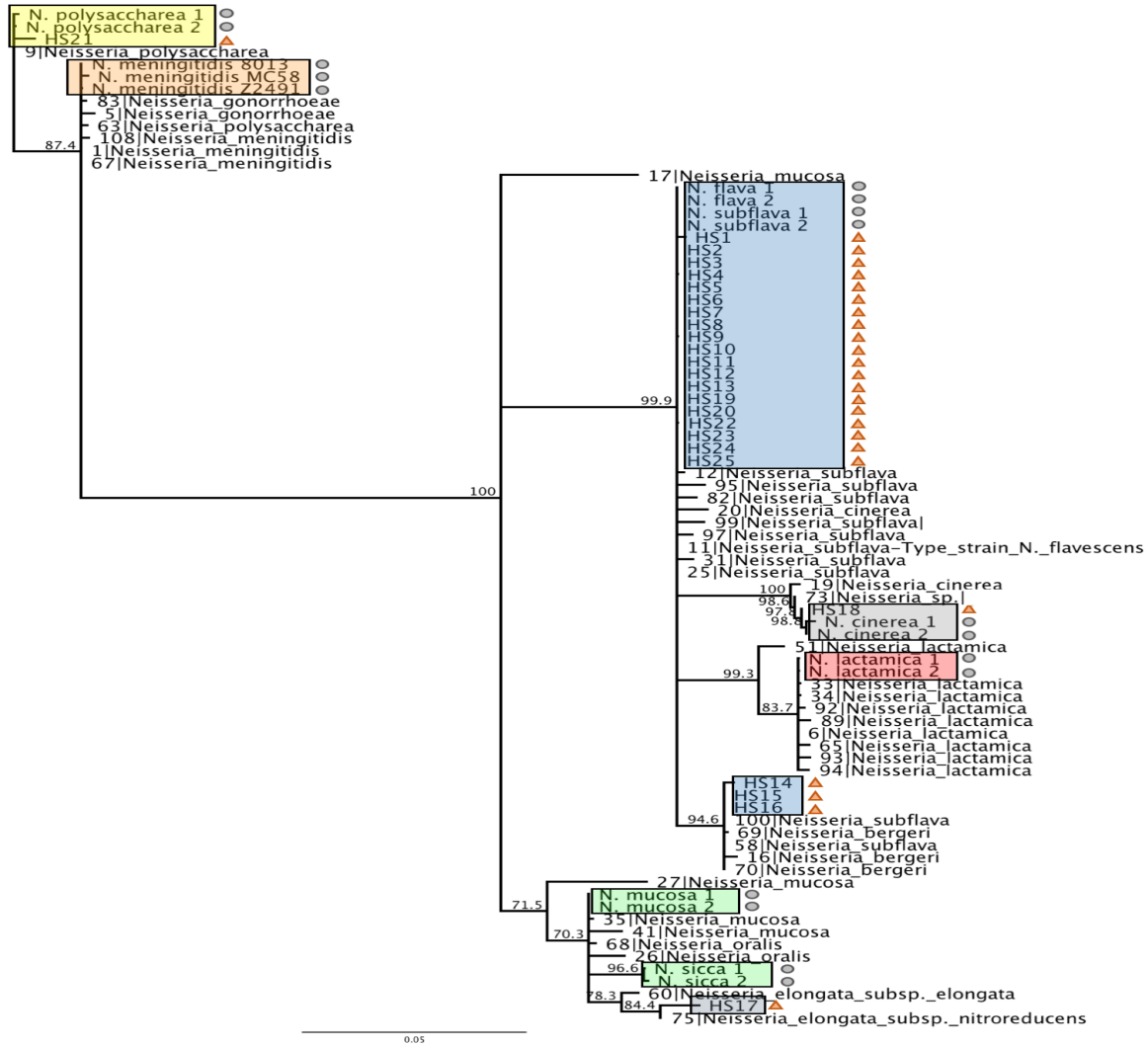


Figure 7 Dendrogram constructed using a subset of *rplF* reference and sequences obtained from bacterial isolates and oropharyngeal human throat swabs. The coloured blocks represent clusters of the samples being tested, each one representing one species (yellow: *N. polysaccharea*, orange: *N. meningitidis*, blue: *N. subflava*, grey: *N. cinerea*, red: *N. lactamica*, green: *N. mucosa,* light blue: *N. elongata*. Grey circles indicate that the sample was a bacterial isolate, while the orange triangles indicate a human sample. The bar underneath the graph represents a similarity of 5% in the sequence (20 bp). The values on the tree branches indicate the consensus support of each branch.

The *Neisseria* isolates tested clustered with reference sequences corresponding to the same species, with the exception of *N. sicca*, which clustered with *N. mucosa*. Regarding the results from human samples, the majority (22/25) were identified as *N. subflava*, one as *N. polysaccharea,* one as *N. cinerea,* and another one as *N. elongata*.

The sequences obtained from human samples had a lower overall quality compared to those from genomic sequences, having an average percentage of high-quality and low-quality bases of 30.92 ± 33.56%, and 50.27 ± 30.36%, compared to the genomic sequences, which were a 65.95±25.73% and 23.37±20.57%, respectively. An example of this can be seen in Figure 8 which shows a comparison of Sanger sequencing chromatograms obtained from four human samples (HS1, HS3, HS10, H20), two *N. subflava*, and two *N. flava*. The average quality from the chromatograms from genomic DNA was significantly higher (p=0.0005) than those from human samples.

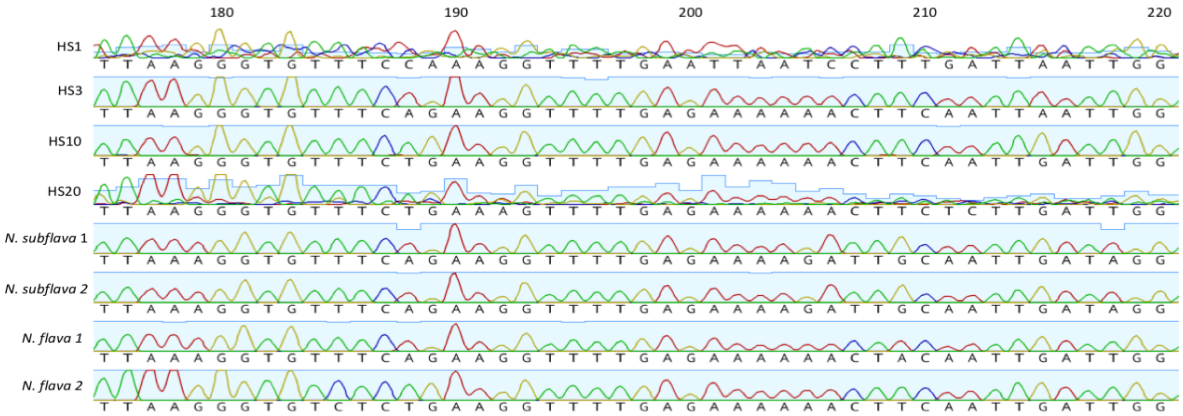


Figure 8 Comparison of chromatograms obtained from four each of human oropharyngeal samples, and bacterial genomic DNA. Each coloured line represents the fluorescent signal produced by each base as follows: T = green; A =red; C = blue; and G = yellow. The light-blue section behind the traces represents the quality for base sequenced, a higher bar representing a higher quality. The base corresponding to the most prominent peak for each position in shown below the traces. The numbers at the top of the figure represents the base of the amplicon being shown.

3.4.4. Primer specificity

In Bennett *et al.* (40), the degree of specificity of the *rplF* primers to known neisserial *rplF* gene sequences is not mentioned. The *rplF* gene is a ribosomal gene, and it could be expected that its sequence shows high conservation among different bacterial families (23). As one of the next steps was to apply the *rplF* sequencing method to DNA extracted from human oropharyngeal swabs, I investigated interactions the *rplF* primers could have with DNA from other upper respiratory tract commensals, or human DNA, when present in the sample.

3.4.4.1. Polymorphisms in primer binding sites

In Section 3.4.1, it was mentioned that there were two variable bases in the Forward primer binding site. I decided to further investigate how variable the bases in the Forward primer are. In order to test for a general identity in the primer binding sequence, I selected one *rplF* gene from each of the following species as a representative: *N. canis, N. arctica, N. wadsworthii, N. dentiae, N. zoodegmatis, N. dumasiana, N. zalophi, N. weaver, N. animaloris, N. gonorrhoeae, N. polysaccharea, N. meningitidis, N. elongata, N. cinerea, N. flavescens, N. lactamica, N. mucosa, N. sicca* and *N. subflava.* The region corresponding to the *rplF* Forward and Reverse primer binding sites were concatenated, and compared to the concatenated primer sequences (Forward primer and reverse complemented sequence of the Reverse primer). The resulting sequences were aligned and the total number of differences among sequences was obtained (Figure 9). Species associated with carriage in humans (*N. gonorrhoeae, N. polysaccharea, N. meningitidis, N. elongata, N. cinerea, N. flavescens, N. lactamica, N. mucosa, N. sicca* and *N. subflava*) had a much higher similarity to each other than the other species, comprising commensals of different animals, which all had more than three dissimilarities. Regarding the human-related *Neisseria, only N. polysaccharea* and *N. meningitidis* had a perfect match to the primer sequences, while all the others had at least one dissimilarity.

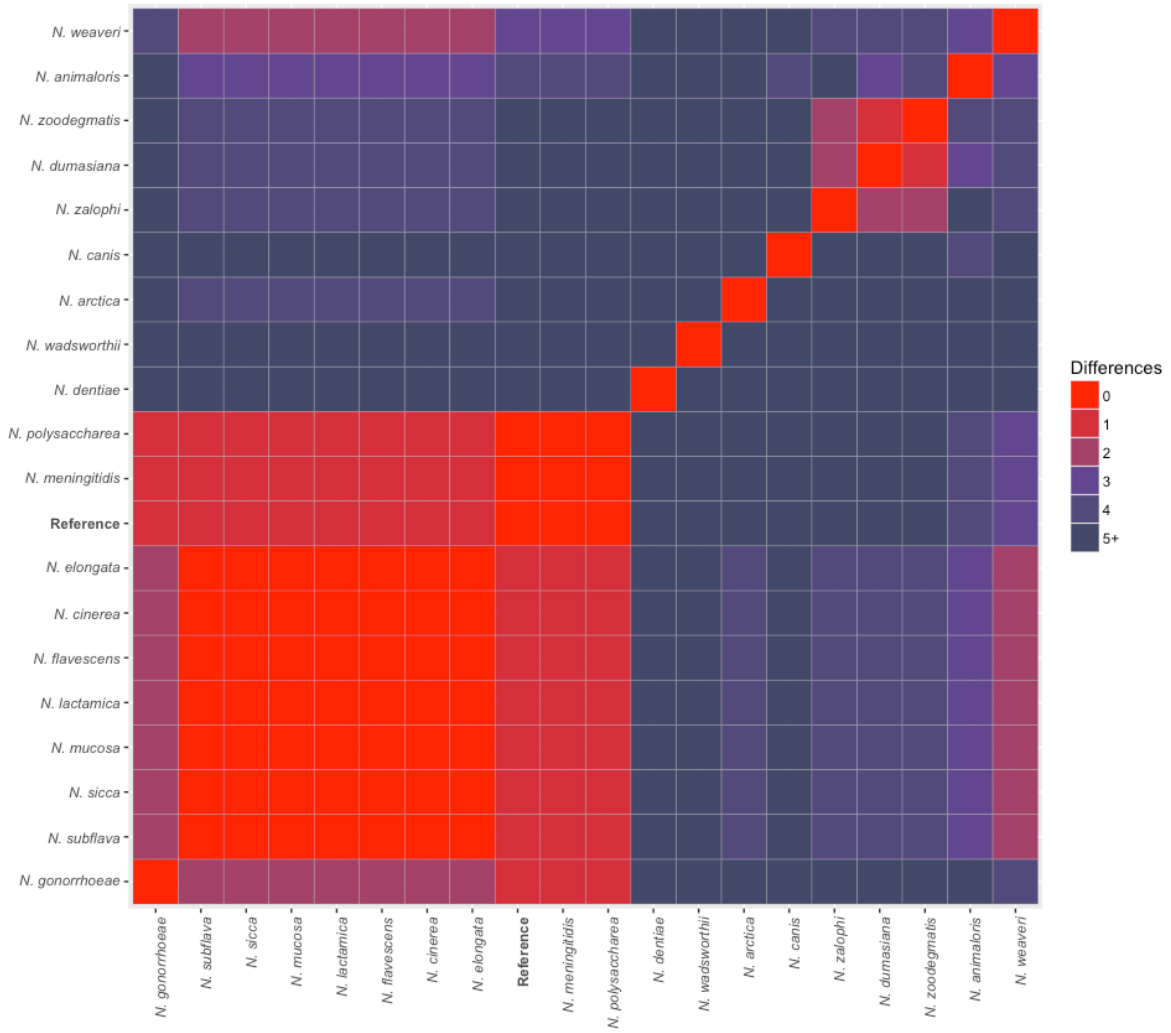


Figure 9 Heatmap showing the pairwise total number of differences among Neisseria *rplF* primer binding sequences. Brighter colours indicate a higher similarity of sequences.

To further corroborate these results, particularly for the human-related *Neisseria,* I downloaded the *rplF* gene records from the PubMLST *Neisseria* database, keeping up to 1000 for each species. Then, I selected only the sequences corresponding to the Forward primer binding site. A summary of the results obtained is shown in Table 4. Only *N. meningitidis* and *N. polysaccharea* had perfect matches to the published primers. All other species, except for *N. gonorrhoeae*, had the same base change, a substitution of a C to a T, at base 14 of the primer binding site, while *N. gonorrhoeae* had a substitution of a T to a G at base 8.

Table 4 Number of isolates with a perfect match to the published *rplF* Forward primer (Match), those with at least 1 mismatch (Mismatch) and the base change seen.

Species	**Match**	**Mismatch**	**Base Change**
N. meningitidis	1000	0	No change
N. polysaccharea	11	0	No change
N. lactamica	0	113	14 (C -> T)
N. subflava	0	19	14 (C -> T)
N. cinerea	0	9	14 (C -> T)
N. mucosa	0	16	14 (C -> T)
N. oralis	0	3	14 (C -> T)
N. sicca	0	4	14 (C -> T)
N. bergeri	0	9	14 (C -> T)
N. elongata	0	4	14 (C -> T)
N. gonorrhoeae	0	1000	8 (T -> G)

The differences identified in the primer binding region do not prevent the *rplF* primers from amplifying the gene in species without a perfect match, but it may not amplify the sequences as efficiently when the mismatches are present. With this information, I decided to test the following primer (rplF_Fm) to see if there could be differences in the amplification of *rplF* sequences:

- rplF_Fm: 5'-CAGTGACTGTTCCYGCTGGTGT-3'

This primer includes an ambiguous base (Y) at base 14 that accounts for the presence of T and C at that location. When tested on human and genomic samples, the new primer in conjunction with the published Reverse primer, amplified *rplF* sequences identical to those published. The Reverse primer, which already includes four ambiguous bases, did not show any bias towards a particular species.

3.4.5. Similarity of rplF sequences across upper respiratory tract commensal and other Neisseriaceae.

To further assess the impact of sequencing *rplF* in a complex mixture of DNA, I decided to test *in silico* the similarities of *rplF* sequences from other *Neisseria* and known upper respiratory tract commensals. For this, I downloaded available genomic sequences from GenBank of the bacteria indicated below (number of genomes is shown in parenthesis): (A) Different species in the *Neisseriaceae* genus: *Vitreoscilla* (6)*, Stenoxybacter* (1)*, Snodgrasella* (114)*, Simonsiella* (5)*, Morococcus* (6)*, Kingella* (61)*, Eikenella* (52)*, Crenobacter* (2)*, Conchiformibius* (2)*, Bergeriella* (2)*, Amantichitinum* (2) and *Alysiella* (3); (B) Upper respiratory commensals: *Veillonella* (26)*, Streptococcus* (328)*, Rothia* (42)*, Prevotella* (93)*, Moraxella (49), Haemophillus (74), Corynebacterium* (209) and *Brevibacterium* (2) as described by Biesbroek *et al*. (107).

The *rplF* gene from these genomes was extracted and aligned to the *rplF* gene from 100 *Neisseria* isolates. Based on the position of the primer binding sites in the *Neisseria* sequences, the sequences corresponding to the Forward and Reverse primer binding site, and the amplicon sequence (according to my definition of the amplicon) were extracted based on the sequence alignments. The primer binding sequences were concatenated to form one unique sequence per GenBank entry. At this point, sequences in these files split into upper respiratory tract and *Neisseriaceae* sequences, keeping the same 100 *Neisseria* records in all files. Finally, all the files were aligned individually, and distance matrices were calculated for each one measuring the total number of dissimilarities for the two primer binding site alignments, and the percentage of identity for the amplicon alignments. This allows the comparison of all sequences against each other. The results of this analysis are shown in Figure 10.

The *rplF* primer binding sequences among certain *Neisseriaceae* were more similar than those of upper respiratory tract commensals. *Simonsiella* spp., *Morococcus* spp., *Kingella* spp., *Conchiformibius* spp., and *Bergeriella* spp. sequences had 7.82±0.96, 3.61±2.08, 8.13±1.07, 6.80±1.70, 3.34±2.60, 4.10±3.14 bp differences on average to *Neisseria,* respectively. However, only *Morococcus* had a maximum identity to other *Neisseria* sequences higher than 90%, matching *N. sicca* sequence up to 99.6% and *N. macacae* up to 98.9%. *Neisseriaceae rplF* amplicon and primer binding sequences were more varied than those of upper respiratory tract commensals. *Prevotella* is the closest in terms of primer binding sequence similarity, being 12.91±1.95 bp different on average. All non-*Neisseria* amplicon sequences had a similarity below 85% to the *Neisseria* sequences.

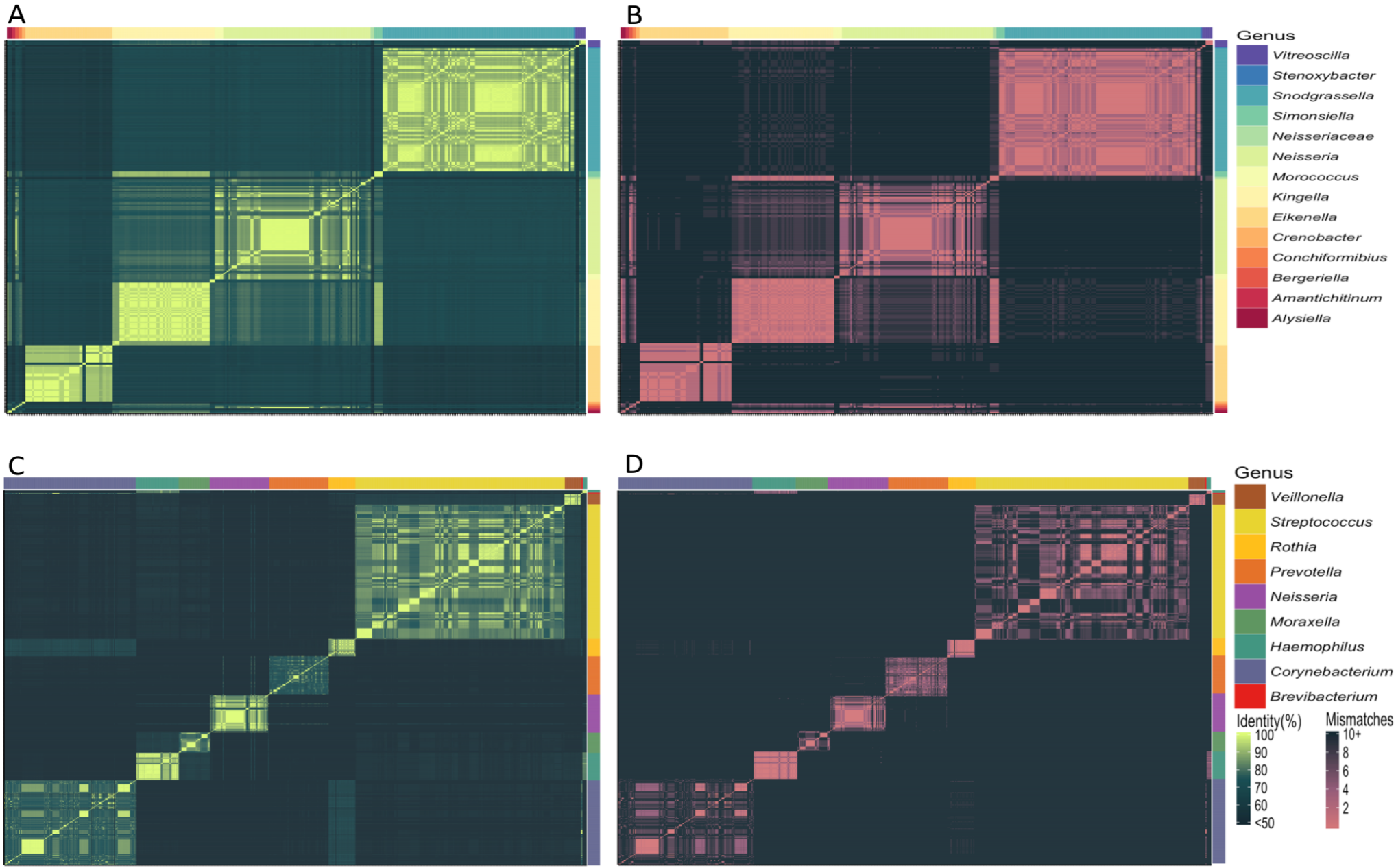


Figure 10 Heatmaps showing the pairwise dissimilarities among *rplF* amplicons, and rplF primer binding sequences in the Neisseriaceae and other upper respiratory tract commensals. Figures 9A and 9C show the percentage of identity of rplF amplicons among Neisseriaceae and upper respiratory tract commensals, respectively. Figures 9B and 9D show the number of dissimilarities (mismatches), in the same order as Figures 9A and 9C. All four images have a colour bar on the top and right margins, depicting the area of the picture occupied by each genus, with the legend at the right side of the Figure relating each genus to each colour. The percentage of identity is shown in green and the number of mismatches in pink, according to the legend at the bottom right corner.

3.4.5.1. Is Morococcus cerebrosus related to N. mucosa?

Based on the results from obtained after the *Neisseriaceae* comparisons, I wondered if there could be problems in downstream analysis regarding *Morococcus cerebrosus*, as it could be possible to confound *M. cereborsus* sequences with *N. sicca* or *N. mucosa* sequences. *M. cerebrosus* is a Gram-negative, oxidase-positive coccus originally isolated from a brain abscess (108). Due to the aggregate-forming morphology when cultured, it was initially compared to a reference *N. mucosa* isolate, and it was considered to be similar enough to it to be part of the *Neisseriaceae,* but not similar enough to be considered a *Neisseria* species. However, since the first publication, very little work has been done to further characterize this genus. One article (109) identified *Morococcus* as part of the oral microbiome, its sequencing reads forming the same OTU as other *Neisseria* sequences. Since we could expect to find this bacterium colonizing the upper respiratory tract, I decided to further characterize *M. morococcus* sequences.

To compare *Neisseria* and *Morococcus* sequences, I decided to use the average nucleotide identity (ANI) score. Briefly, ANI is a method that allows the comparison of genomes at the nucleotide level. After pairwise genome alignments are computed, reciprocal best hits (RBH) are identified (*i.e.* two genes from two different genomes that are the best match to each other when comparing all genes from both genomes). Finally, the ANI score is defined as the sum of the product of the identity of RBHs multiplied by the alignment length, divided by the total length of the RBH genes. So, for two genomic sequences that are identical the score should be 1, and the more diverse the sequences are, the lower the score would be. Scores above 0.95 are considered acceptable to delimit genomes from the same species (110). The ANI method has been extensively used to identify similar species, and correlates well with DNA to DNA hybridization values (111).

Using the pyani package, I compared *M. cerebrosus* genomic sequences (N=5), with *N. mucosa* (N=6), *N. sicca* (N=9), *N. lactamica* (N=3), *N. subflava* (N=3), and *N. meningitidis* (N=5) downloaded from NCBI. *N. lactamica, N. subflava* and *N. meningitidis* sequences were used as controls. The results are shown in Figure 11. *N. lactamica* and *N. meningitidis* formed single species clusters, and *N. subflava* clustered with two of the *N. mucosa* genomes. Regarding *M. cerebrosus*, it clustered very closely to some of the *N. sicca* and *N. mucosa* in two different clusters, while having scores very close to 0.95 or higher to the *N. sicca* and *N. mucosa* genomes, and sequence identity almost identical to a *N. sicca* isolate.

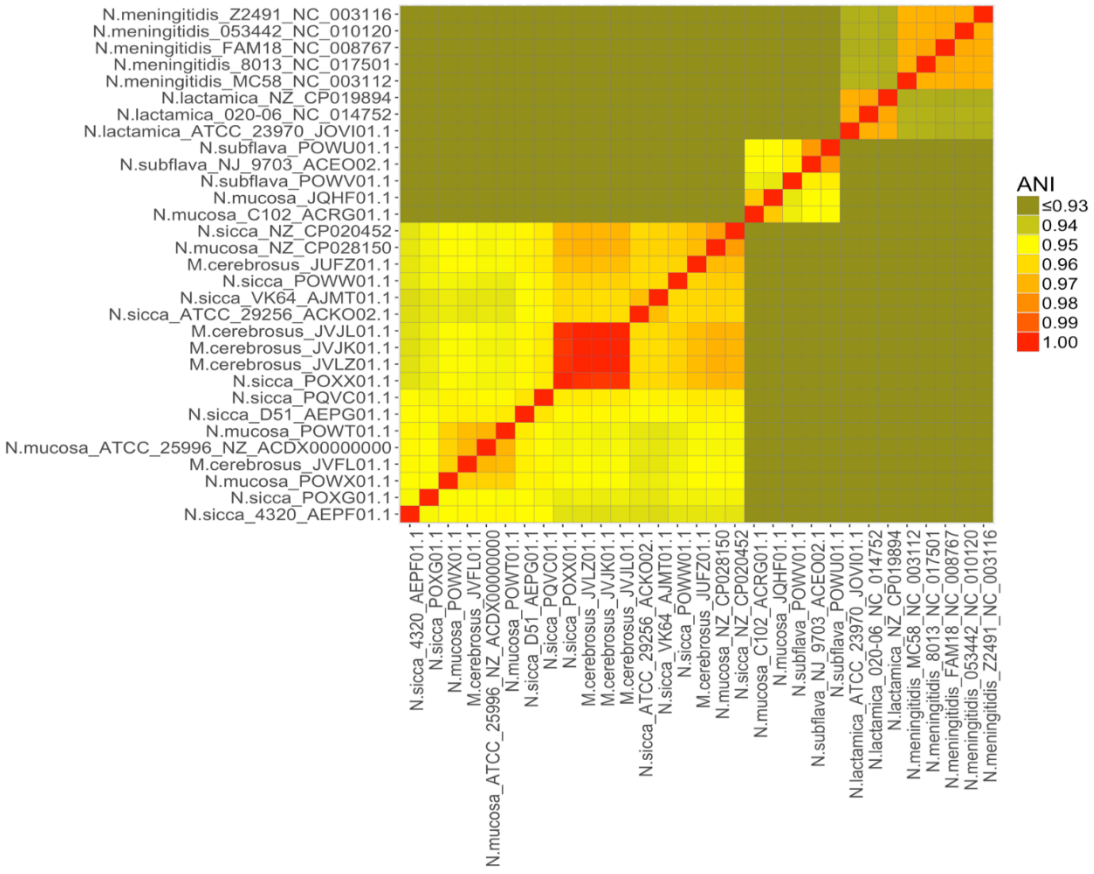


Figure 11 Heatmap showing the pairwise ANI scores comparing *Neisseria* and *Morococcus* species. Each tile is coloured according to the legend, wth higher ANI scores shown in brighter colours.

3.5. Discussion

The development of the *rplF* sequencing method by Bennet *et al*. (39) has been used as a tool for the identification of *Neisseria* isolates, to identify species in large scale studies (51,112), and distinguish novel *Neisseria* species from others (43,44). However, it is important to recognize its limitations to be able improve the method.

I was able to reproduce the phylogenetic trees shown in Bennett *et al.* (39) , showing how it is possible to obtain clades that lead to unique species. As mentioned previously, most species form a single clade in which all the sequences from a particular species cluster, with the exception of *N. mucosa, N. cinerea, N. oralis* and *N. bergeri.*

One of the complications with the Bennett et al. (39) method is the misidentification of species based on the method used to determine the closest match to a reference sequence. For a "lower scale" study (samples sequenced using the Sanger method), identifying the nearest match based on a tree could be feasible, as this is not very computationally intensive, however, using BLAST or a BLAST-like algorithm could also be possible. It is important however, to highlight the closeness of sequences belonging to different species, such as *N. bergeri* and *N. cinerea* with certain *N. subflava*. This may not pose a difficulty as long as the quality of the chromatograms (for Sanger sequencing) is satisfactory.

With regards to the classification of the *Neisseria* isolates tested and shown in Figure 7, with the exception of *N. sicca* and *N. flava*, all species were clustered as expected. *Neisseria* taxonomy has not been investigated thoroughly outside of the most well studied organisms: *N. lactamica*, *N.*

meningitidis, and *N. gonorrhoeae.* In Bennett *et al.* (52) this issue is briefly discussed, and it is mentioned that isolates previously identified as *N. sicca* were almost identical to *N. mucosa* when classified using 246 conserved genes, or rMLST scheme genes, suggesting that the *N. sicca* isolates where not classified properly when first discovered, and should be included within *N. mucosa.* Our *N. sicca* isolates followed this pattern and were clustered with *N. mucosa* reference sequences.

The case for *N. flava* is different, as *N. flava*, *N. flavescens*, *N. perflava* and *N. subflava* have traditionally been classified as biovars of *N. subflava* with the exception of *N. flavescens*, which was associated closer to other pathogenic *Neisseria* (1,113). My analysis followed this trend and classified the *N. flava* isolate with *N. subflava*. Recent analysis using MLST and core genome analysis schemes have classified *N. flavescens* to be indistinguishable from *N. subflava* or *N. mucosa* depending on the isolate (114).

For this study, I will follow the guidelines suggested by Bennett *et al.* (52) to classify *Neisseria* species. In summary, *N. gonorrhoeae, N. meningitidis, N. polysaccharea, N. lactamica, N. cinerea, N. subflava* and *N. mucosa* are considered the principal *Neisseria* species, but only *N. gonorrhoeae, N. meningitidis, N. lactamica* and *N. polysaccharea* form clusters that only include isolates previously described as those species. *N. cinerea* contains also *N. flavescens* isolates. *N. subflava* includes *N. subflava, N. flavenscens* and *N. flava* isolates. *N. mucosa* is comprised of *N. mucosa, N. perflava, N. sicca* and *N. macacae.* Others, such as *N. mucosa* var. *heidelbergensis, N. bacilliformis, N. elongata, N. dentiae, N. canis* and *N. weaveri* formed individual clusters each, but far more distant to the species first mentioned. However, in (114) it is demonstrated that *N. mucosa* var. *heidelbergensis* is identical to *N. oralis.* Finally, *N. bergeri*, even though it has not yet been formally acknowledged as a species,

was identified originally as a *N. polysaccharea* isolate, but differed from other isolates of this species when comparing rMLST profiles (52).

The results obtained from sequencing *rplF* amplicons from human oropharyngeal throat swabs are more difficult to interpret. The majority of the samples matched *N. subflava* sequences, while others matched *N. polysaccharea, N. cinerea,* and *N. elongata*, but, as described in Figure 7, all have been identified previously as commensals of the human oropharynx. Amplicons obtained from pure genomic DNAs should have a high quality when sequenced, but as seen with the human samples sequenced, the quality can vary significantly. There could be different explanations as to why one could obtain lower quality sequences when testing DNA extract from throat swabs:

1) Lack of primer specificity that leads to secondary amplification.

2) Presence of multiple *Neisseria* and *rplF* genes in a sample, which lowers sequencing quality.

I did not find unspecific bands generated during the amplification of *rplF* with the PCR conditions established by Bennett *et al*. (39). However, it could be possible that bands are present at a very low concentration, making the detection of unspecific bands difficult when amplicons are electrophoretically separated. One of the purification steps undertaken before sequencing the amplicons involved a gel extraction after separating the samples using a 2% agarose gel, so unless a theoretical secondary amplicon of the same size as the main amplicon exists it would not be able to be completely separated from the main one. So, it is possible that more than one *rplF* profile could be present in samples, be it from the same species or from another.

So far, the overall number of different *Neisseria* species co-habiting the pharynx at a given time is not known or has not been estimated. There have been several studies in which the carriage of *N. lactamica* and *N. meningitidis* have been studied over time (115-120), but only the overall number of isolates per time point were determined. Only two articles mention co-occurrence of *N. lactamica* with *N. meningitidis* (6,121). There is one study which used the *rplF* gene to identify *Neisseria* species from oropharyngeal swabs cultured on Thayer Martin agar (51). However, the number of isolates identified per swab was not specified. The results obtained by sequencing the *rplF* gene from human samples indicates the presence of more than one species.

This study identified great variability in the Forward primer binding sequence of bacteria, a point not mentioned by Bennet's *et al*. (40). While human-colonizing *Neisseria* had a relatively similar primer binding sequence for the Forward primer, animal-colonizing *Neisseria* were far more varied, having three or more differences. This could suggest that only human-colonizing *Neisseria* will be amplified when using the original Bennett et al. (39) primers. Taking only into account the human-colonizing *Neisseria* group, the variability described for the two bases in the Forward primer binding region could have a significant impact when applied to a NGS platform. If one were to just use Sanger sequencing to identify species using genomic DNA, this would not be an issue, but when dealing with complex mixtures of DNA, the mismatch could result in an imbalanced representation of the species present in a sample. Previous publications, using 16S rRNA sequencing, have shown that a single base pair mismatch could have a significant impact on the sequencing results, completely missing one genus present in the samples (92). Therefore, we developed a new Forward primer (rplF_Fm) that could account for this possible problem by changing the 14th base of the primer to from a C to a Y. Even though I know that this primer would not consider the base change present in *N. gonorrhoeae* (8th base changed from T to a G), it was decided that this would not be tested as it is not expected to find

this particular species in the infant population. However, it may need to be tested in the future if this is to be extended to other samples where *N. gonorrhoeae* would expect to be present.

The *rplF* gene is a ribosomal gene, and it is one of the genes included in the rMLST scheme used to identify different bacterial genera (23), but the question of how conserved its sequence is remained to be answered. I decided to test the *Neisseria rplF* amplicon sequences to other *Neisseriaceae* and known upper respiratory tract commensals in infants. Choosing to compare the *Neisseria* and *Neisseriaceae* sequences serves two purposes: these are the sequences that are the most likely to be the most similar to *Neisseria*, and genera in this family, such as *Eikenella* (33), *Bergeriella* (122), *Simonsiella* (123) and *Kingella* (124), have been identified in oral microbiome studies, so it could be possible that these could be identified in the samples being analyzed.

The results comparing *rplF* sequence similarity among URT commensals did not show any similarity of concern, neither regarding the primer binding or the amplicon sequences. The results suggest that *rplF* sequences from relatively distant bacterial families are sufficiently dissimilar and misidentification as a *Neisseria* species is unlikely. This is similar to the aforementioned *rplF* reference sequences belonging to *B. denitrificans* and *Moraxella*; their *rplF* sequences are distant from those of *Neisseria.*

The similarity of the of the *Neisseria rplF* primer binding and amplicon sequences to other *Neisseriaceae* showed some degree of similarity, especially to the four genera mentioned above. The shared similarity is not of great concern, as the results suggest that it could be possible to amplify some of these sequences based on the primer binding sequence similarity, but the amplicon

sequences, although more similar to *Neisseria* than the ones from URT commensals, have an identity score below 90%. The only exception to these results was *Morococcus*. This genus has not been greatly studied. Its first description (108) suggested the bacterium was related to *Neisseria*, but a different genus. For example, in the article where it was first described, it was shown that the *Morococcus* isolate and *N. mucosa* had different growth patterns, and antibiotic resistance profiles compared to *N. mucosa* (ATCC 19696, GenBank entry NZ_CP028150). Furthermore, its characteristic 'mulberry-like' aggregates, helped separate it from *Neisseria*. The ANI scores shown in Figure 11, suggest that the *Morococcus* genome is very similar to other *N. sicca* and *N. mucosa* (which could all be classified as *N. mucosa*), including one *N.* sicca genome which was identical to three of the *Morococcus* genomes. Although these results are not definitive proof that the *Morococcus* genus should be part of *Neisseria,* it is highly suggestive. For the analysis here presented, it is impossible to distinguish *Morococcus* and *N. mucosa* by *rplF* sequencing as described and *Morococcus*, should it appear in the samples collected, will be considered as *N. mucosa*.

4. Chapter 4: Adaptation of the rplF sequencing method to the Illumina MiSeq platform

4.1.introduction

NGS is a term generally used to describe high-throughput sequencing methods, which allow sequencing multiple DNA fragments in parallel. The development of faster and more accurate sequencing platforms since the mid-2000s has allowed the scientific community to obtain high quality genomes (125), sequence PCR amplicons, analyse transcriptomes or identify genes responsible for disease (126). Developments have led to an increase in the number and quality of reads obtained per run, while lowering the price per base sequenced (127).

Without taking the cost of the equipment into consideration, the two main variables that differ among platforms are read length size and number of reads obtained. NGS platforms can be divided into short-read and long-read platforms. Short-read platforms, like SOLiD 5500 xl or Illumina HiSeq 2500, can provide reads of 75 and 125 bp, respectively, while long-read platforms, like Pacific Biosciences RSII or Oxford Nanopore MK1 MinION, can sequence DNA fragments several kb long. However, there is usually a trade-off between sequencing length, average quality of the reads, and quantity of reads obtained, where short-read platforms tend to have higher sequencing quality and output. The scientific community has exploited this trade-off to sequence specific fragments of DNA rather than whole genomes, the main purpose of these machines, adapting different read length capabilities and platforms to different applications. Several publications show that sequencing custom amplicons is possible in Illumina platforms, such as fish DNA metabarcoding (128), Honey bee pollen

characterization (129), sequencing foot-and-month disease related viruses (130), and identification of bacterial populations in environmental samples (131). With these ideas in mind, I set out to translate the *rplF* sequencing method to a NGS platform (Illumina MiSeq), and develop a reproducible protocol that would allow me to obtain large numbers of high quality sequences.

4.2. Aim

- Develop a reliable protocol to amplify and sequence *rplF* amplicons using the MiSeq platform.

4.3. Summary of methods

Library preparation methods are described in detail in Section 2.10. A summary of such methods is displayed in Figure 12. For each protocol version, the first line shows the steps taken to generate the library, and the second one quality control (QC) and library and PhiX steps taken before loading the sequencing cartridge into the MiSeq. For the two-step library protocol, libraries 8 to 12 underwent an extra heat denaturation step before sequencing to help improve the quality of the sequencing cycles.

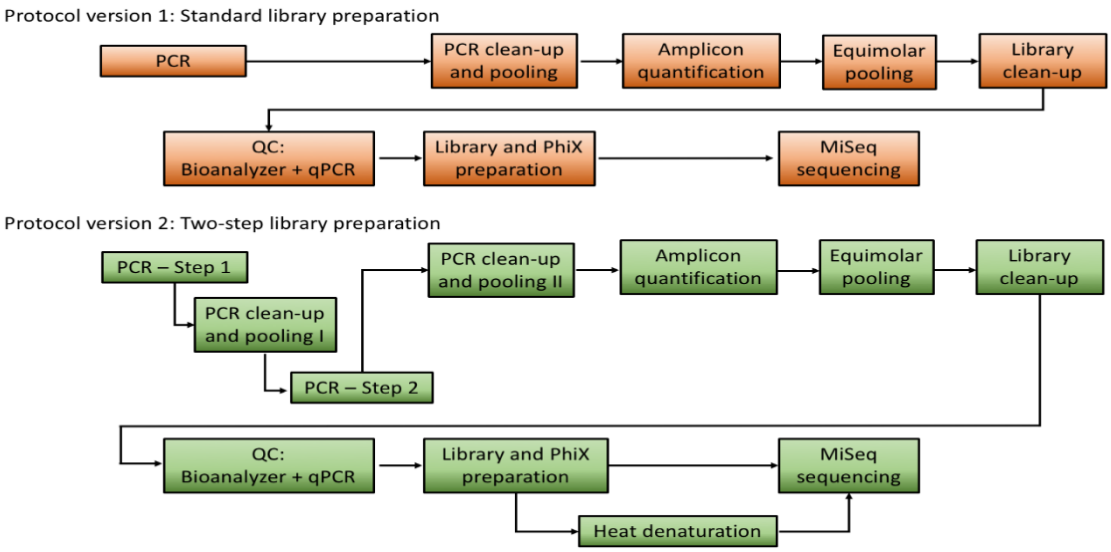


Figure 12 Comparison of Protocol version 1 and 2 used to generate MiSeq-compatible libraries. Each block represents one of the steps of the library preparation protocol. The standard library preparation, also refered to as single-step library protocol, is shown in orange, and the two-step library protocol in green.

4.4. Next generation sequencing: Illumina MiSeq sequencing method

The DNA molecules present in MiSeq libraries require two adapter sequences, referred to as P5 and P7. These adapter sequences are designed to be complementary to the nucleotide sequences present in the Illumina flow cells. The DNA molecules are followed by index sequences that are used to identify DNA sequences belonging to particular samples included in the library pool. Downstream of the adapter and index sequences are the sequencing and PCR primers which, depending on the design, could be the same or two different sequences. Furthermore, PCR primer sequences can be separated by a linker sequence from the rest of the oligonucleotide and prevents the other elements present upstream of the primer sequence interfering with the PCR amplification to avoid possible biases.

Once the sequencing cartridge containing the library is loaded into the MiSeq, the library goes through a process in which each DNA molecule is amplified for 35 cycles, in a process called bridge amplification. During this process, DNA molecules are incorporated into the flow cell, and bind to the oligonucleotides in it due to their complementarity to P5 and P7 sequences. DNA molecules are amplified using the oligonucleotides present in the flow cell as primers. This process yields clusters of identical sequences, which are also referred to as polonies. In the final step before sequencing, only the molecules bound by the P7 adapter are selected. This process is described in Figure 13.

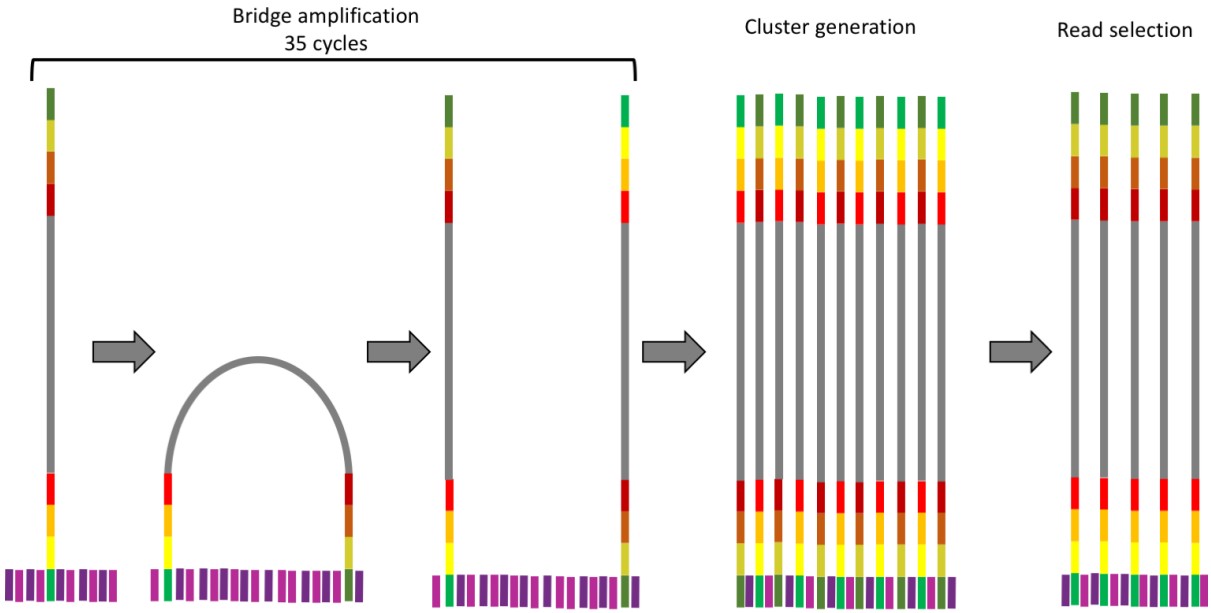


Figure 13 Bridge amplification of DNA molecules present in in a sequencing library. P5 and P7 binding sequences are shown in purple, P5 and P7 sequences are shown in green, index sequences in yellow, pad sequences in orange, and primer sequences in red, amplicon sequences are shown in grey.

MiSeq instruments perform four different read sequencing steps, sequencing different segments of the DNA molecules in each. For each step, a nucleotide mix is passed through the cell, and the nucleotide bound to the DNA molecules are identified thanks to fluorophores of different wavelengths bound to them. This process is repeated for a number of cycles, depending on the kit used, for Read 1 and Read 2. The first sequencing primer is used to sequence the Forward read of the DNA molecules. Following this, the indexing primer is passed through to sequence the Index1. The DNA molecules are then turned over, so that the molecules are bound by the P5 adapter, which is

used as the primer to sequence Index2. Finally, the second sequencing primer is passed through, and Read2 is sequenced. These four steps are represented in Figure 14.

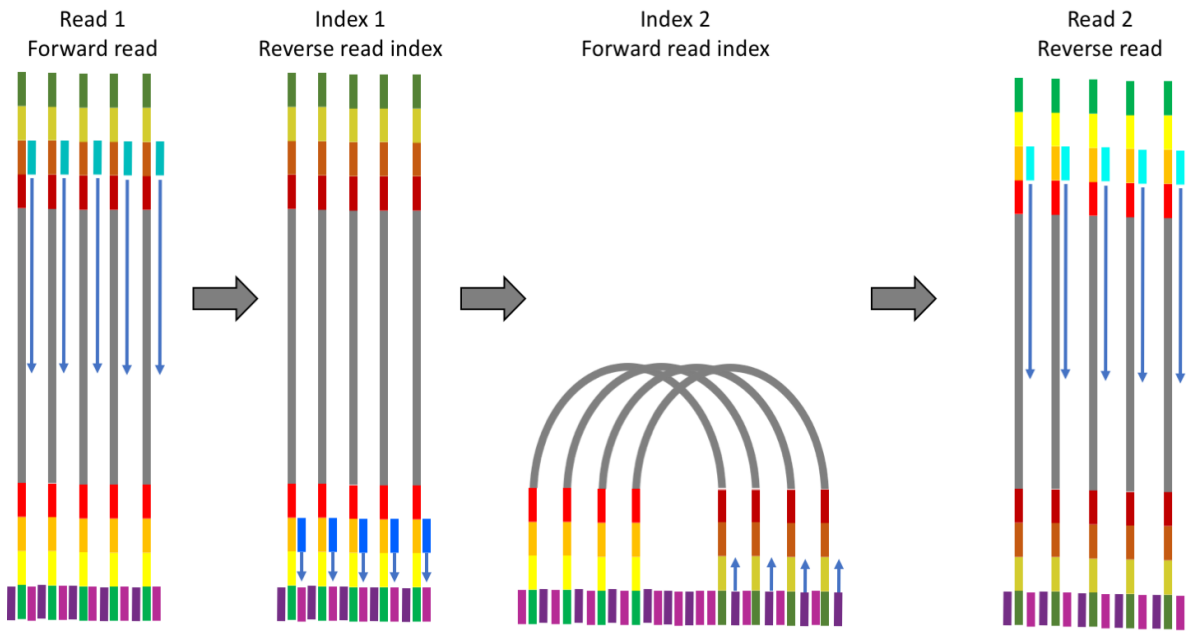


Figure 14 Description of the four different sequencing steps taking place during a MiSeq sequencing experiment. P5 and P7 binding sequences are shown in purple, P5 and P7 sequences are shown in green, index sequences in yellow, pad sequences in orange, and primer sequences in red, amplicon sequences are shown in grey. Blue segments and arrows indicate primer bindings and the direction of the amplidication of the DNA, respectively.

4.5. Primer design for custom amplicon sequencing in the MiSeq platform

Illumina's MiSeq and HiSeq platforms are capable of sequencing custom amplicons, however, there are several considerations one has to consider when designing the primers used to amplify the target DNA, and those used to sequence the amplicon in the instrument. While Illumina has published the adapter sequences and index sequences used in their kits (https://support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/experiment-design/illumina-adapter-sequences-1000000002694-06.pdf), there is no clear guide as to how to design a new sequencing protocol. Illumina suggests using one of

its kits, Nextera or TruSeq, to sequence DNA fragments in their platforms, although other companies such as New England Biolabs have created kits that emulate the results from those of Illumina (132). However, the use of these kits may impose an economic burden on large scale studies.

One crucial aspect that needs to be considered for the successful development of a new sequencing protocol is primer design. As for standard PCR approaches, it is necessary to assess the specificity of the primer sequences towards the desired DNA sequence. This aspect is more important when the DNA being amplified is present in a complex mixture of DNA, as it provides a higher possibility for similar sequences to appear in the mixture, which can lead to unspecific amplification. Another important aspect is the variability within the primer binding sequences. Previous studies (133) have highlighted the variability in results obtained when using different variable regions of the 16S rRNA gene to identify bacteria. It has also been shown how a single nucleotide ambiguity in the primer sequences can have a deleterious effect in the identification of a genus (92). It is therefore of great importance to identify potential sources of variation in the primer binding sequences, investigate the potential effect of their variation and, if possible, address this variation when designing the primer pairs to be used.

The variability and length of the amplicon being sequenced should also be considered when designing a sequencing protocol. Different sequencing platforms have biases or difficulties when sequencing particular types of sequences. As an example, homopolymeric regions, in particular those rich in A and T nucleotides, had larger error rates when sequenced in a Roche 454 platform (134). On the other hand, if one is interested in a longer DNA sequence or a palindromic sequence, 454 sequencing could prove more advantageous, as this platform provided longer reads than Illumina platforms (134). Illumina platforms can also struggle when sequencing low-diversity libraries (135),

and although it is possible to offset the low diversity by spiking libraries with a PhiX control, it is probable to obtain poor sequencing results due to poor base calling quality (135).

4.5.1. rplFV1 primer set development

The first primer set designed to sequence *rplF* amplicons in the MiSeq platform was designed using the published adaptor and index sequences published by Illumina. The overall design of the primer and PCR setup is as described in Figure 15.

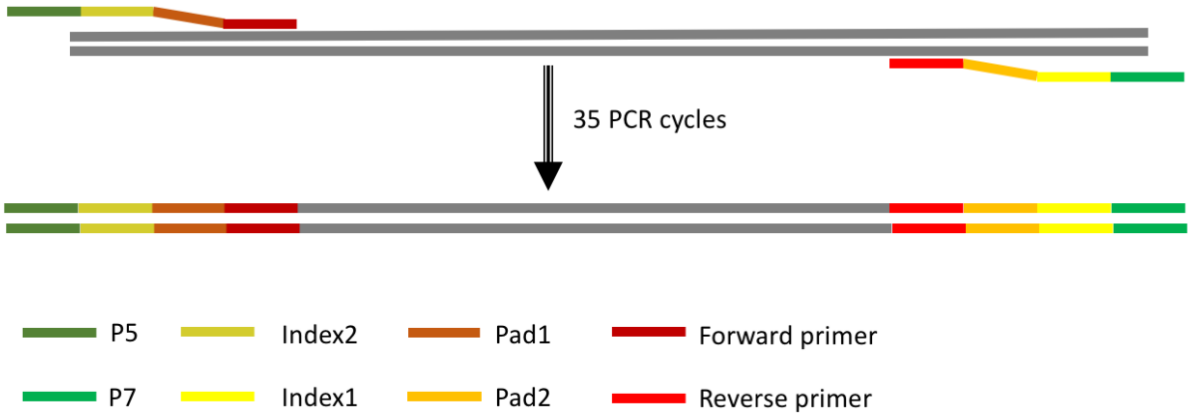


Figure 15 PCR and primer setup for the rplFV1 primer set. P5 and P7 sequences are shown in green, index sequences in yellow, pad sequences in orange, and primer sequences in red, amplicon sequences are shown in grey.

The rplFV1 primer set sequences (Table 5) were selected from the Illumina sequence set whenever possible to avoid problems due to incompatibility. The Adapter sequences, labelled P5 and P7, are complementary sequences to the oligonucleotides present in the MiSeq flow cells, and it should not be altered. The index sequences used are 8 bp long, and they allow tracing of each sequence obtained during the sequencing steps to the samples it originated from. The Forward and Reverse primers were the same ones as described in Bennett *et al.* (40), with the addition of the rplF_Fm primer.

These three primers also contain a padding sequence (Pad1 for the Forward primer and Pad2 for the Reverse, Table 5), which increases the overall melting temperature (Tm) of the sequencing primer. MiSeq instruments operate at a relatively high temperature, around 65ºC. Adding the padding sequences helps raise the normal Tm of the primers (Table 6) to be compatible with MiSeq temperatures. The Table shows the minimum Tm (Min Tm), the average Tm (Avg Tm) and maximum Tm (Max Tm) for all rplFV1 primer sequences including ambiguities. The primer used to sequence the Index2 sequences (Indexing primer) is also included in this Table, its sequence is the reverse complementary to the sequencing primer used for Read1.

Table 5 Sequences used in the rplFV1 primer set

Segment	Sequence
Adapter (P5)	AATGATACGGCGACCACCGAGATCTACAC
Adapter (P7)	CAAGCAGAAGACGGCATACGAGAT
Index2	TAGATCGC, CTCTCTAT, TATCCTCT, AGAGTAGA, GTAAGGAG, ACTGCATA, AAGGAGTA, CTAAGCCT
Index1	TAAGGCGA, CGTACTAG, AGGCAGAA, TCCTGAGC, GGACTCCT, TAGGCATG, CTCTCTAC, CAGAGAGG, GCTACGCT, CGAGGCTG, AAGAGGCA, GTAGAGGA
Pad1	GCCGTGCTCGTT
Pad2	CCAGCCACTGCCGC
Forward primer (standard)	CAGTGACTGTTCCCGCTGGTGT
Forward primer (modified)	CAGTGACTGTTCCYGCTGGTGT
Reverse PCR primer	AGGYTCAGGAGKWCGGAAHG

Table 6 Melting temperature of *rplF* primers and sequencing and indexing primers using in the rplFV1 primer set. For each sequence, the minimum calculated Tm (Min Tm), the average Tm (Avg Tm) and highest possible (Max Tm) is shown. The minimum and maximum Tm values are shown only for those sequences with ambiguous bases in its sequence.

Primer	Sequence	Min	Avg	Max
rplF Forward	CAGTGACTGTTCCCGCTGGTGT		61.7	
rplF Forward (modified)	CAGTGACTGTTCCYGCTGGTGT	59.4	60.6	61.7
rplF Reverse	AGGYTCAGGAGKWCGGAAHG	53.9	57.4	61.4
MiSeq *rplF* Forward	GCCGTGCTCGTTCAGTGACTGTTCCCGCTGGTGT		70.7	
MiSeq *rplF* Forward (modified)	GCCGTGCTCGTTCAGTGACTGTTCCYGCTGGTGT	69.4	70	70.7
MiSeq *rplF* Reverse	CCAGCCACTGCCGCAGGYTCAGGAGKWCGGAAHG	69.5	71.3	73.4
Indexing primer	CDTTCCGWMCTCCTGARCCTGCGGCAGTGGCTGG	69.5	71.3	73.4

4.5.1.1. Results obtained from sequencing libraries prepared using the rplfV1 primers

Using the artificial genomic mixtures described earlier (Section 2.5.3), I prepared a total of two libraries that were sent to the Department of Biochemistry at Cambridge University to be sequenced using the V3 2x300 bp kit (Illumina). Using this kit one can theoretically sequence 300 bp from each end of an amplicon, thus, in the case of the *rplF* amplicons (404 bp), would provide full coverage of the amplicon plus a 196 bp amplicon overlap. Both libraries were prepared with the same samples, but one was prepared using the standard *rplF* Forward primer and the other one with the modified primer.

The first batch of libraries generated was sent for sequencing, but I was informed that no results could be obtained from these libraries, as the MiSeq failed to process the sequencing reads due to a lack of binding of the library to the flow cell, and no output files could be obtained. A second batch of libraries was created using the same primers and the same protocol as the first batch. This time I was told that the MiSeq failed to complete the sequencing cycles due to an error that was later clarified as an internal error due to lack of sequence diversity. A summary of the sequencing quality scores for the third library sequenced are shown in Figure 16. The quality scores obtained were not satisfactory, especially for the Reverse read.

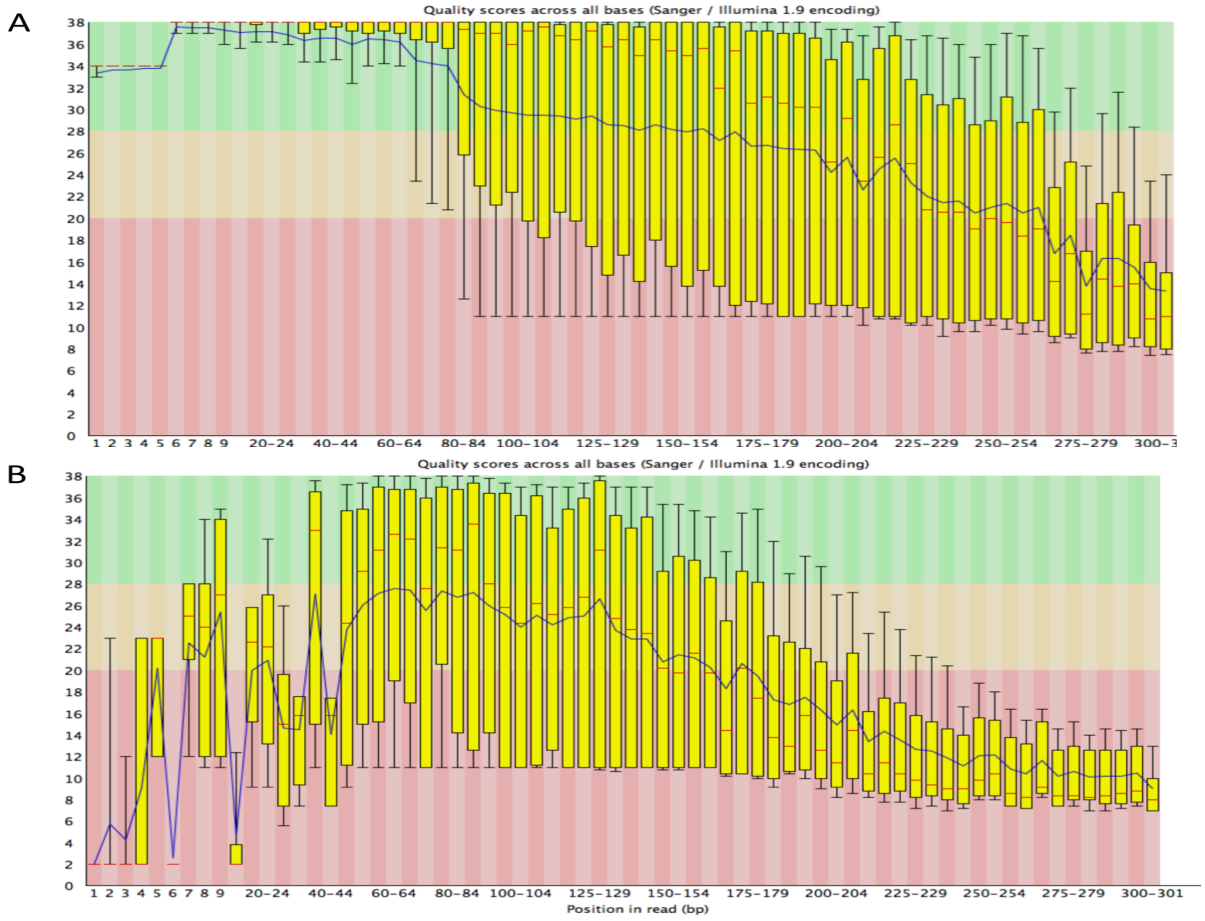


Figure 16 Summary of the quality scores obtained for the forward (A) and reverse (B) reads for the third *rplF* library sequenced. The X axis shows the base, or group of bases, sequenced in each direction, and the Y axis represents the quality of the scores. The red, yellow and green background segments indicate quality scores in between 0 and 20,20 and 20 and above 30, respectively. The sequencing quality scores of each base, or group of bases is represented with a boxplot (shown in yellow), with a red line indicating the median score. The blue line indicates the average quality value.

During a MiSeq sequencing experiment, a decrease in the sequencing in quality over time is expected. However, Q values of 20 (99% base accuracy), are considered the lowest acceptable value used for quality filtering. This also means that Forward and Reverse reads may not overlap and form a full-length amplicon, as the low-quality bases would be trimmed.

After the failed initial attempts to sequence the *rplF* libraries, I was not sure as to what could be the underlying problem present in the libraries. On the one hand, two libraries not binding to the flow cell could mean that the P5 and P7 sequences used could be erroneous but given that this problem did not appear in the second batch of libraries it could not be, as the same primers were used to generate both all libraries. The P5 and P7 sequences present in the rplFV1 primers were checked again, and the sequences were identical to those published by Illumina. On the other hand, the lack of diversity could be an intrinsic problem of the *rplF* amplicons.

Most of the variation in the *rplF* sequence arises from subtle nucleotide changes, and some of the reference sequences only differ in one nucleotide. Figure 17 shows the percentage of times each base appears in the *rplF* amplicon (according to Bennett et al (39)) for all the available *rplF* reference sequences. The majority of the bases appear at a frequency of 10% or less. Another issue faced with the first libraries prepared is that the diversity will be artificially lower than what could be expected to be found when sequencing human samples.

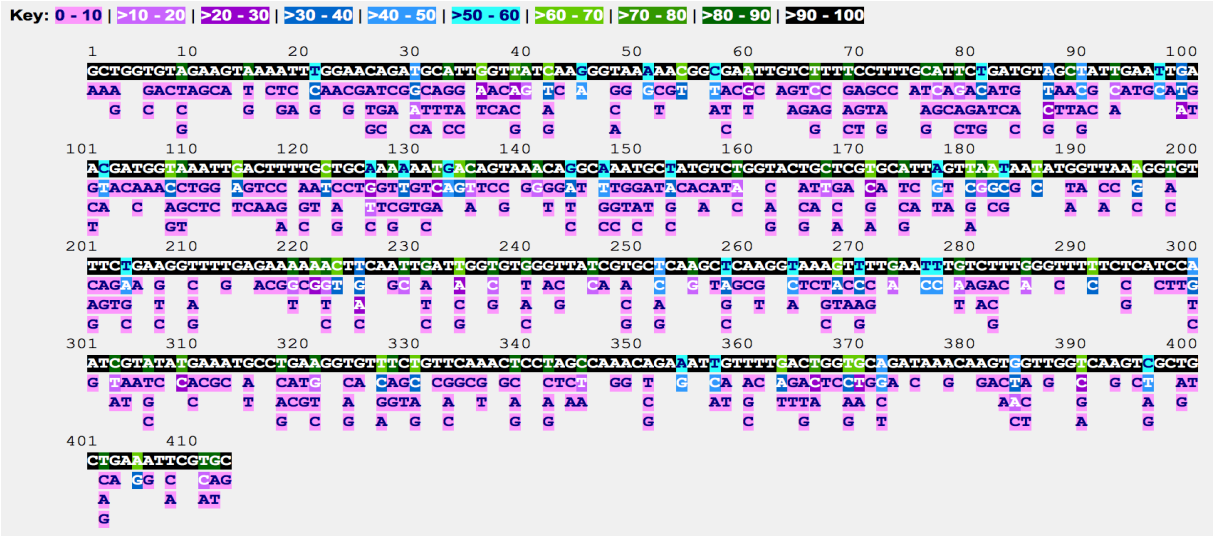


Figure 17 Estimated ratio of appearance of each base in the *rplF* amplicon, image obtained from the PubMLST *Neisseria* database. The appearance of each bease is represented with colours according to the color key on the top left corner.

However, it is possible to identify single nucleotide polymorphisms using MiSeq by increasing the amount of PhiX added to the library before sequencing (136), so it should be feasible to sequence *rplF* amplicons. The role of the PhiX in the library is to increase the diversity within the library when sequenced. Illumina suggests having a minimum amount of 5% PhiX for low diversity libraries. However, this value is dependent on the library sequenced, ranging from 1% (recommended for all libraries by Illumina as an internal control, http://emea.support.illumina.com/sequencing/sequencing_kits/phix_control_v3/questions.html) to 5-30% to most low diversity or monotemplate libraries, depending on the software version on the MiSeq. Libraries were sequenced using a 20% PhiX spike at 8 pM.

The PhiX added as a control (reference sequence: NC_001422) is a ssDNA microvirus bacteriophage genome. It is supplied by Illumina as an already prepared library, meaning that the reagent consists of 375 bp fragments of the PhiX genome with the P5 and P7 adapters attached to it,

but no barcodes are included. This genome is used as a control during MiSeq sequencing for several reasons:

- It has a balanced nucleotide sequence (44.8% GC content) which helps ensure some degree of nucleotide variability within the library at each cycle.
- It is a small genome (5386 bp), which allows the MiSeq to map the PhiX reads to the reference genome without incurring significant memory usage by the machine. This helps identify potential problems during sequencing, as the percentage of PhiX reads obtained should be similar to the percentage of PhiX added to the library
- It provides the MiSeq a method of calculating other useful metrics such as the colour matrix estimation, used to calculate the quality (Q) scores for each base sequenced, and the prephasing and phasing corrections, which account for the proportion of reads which are ahead or lagging within a cluster during the sequencing by synthesis steps.

Due to the lack of reliability of library preparation and sequencing approach used up until now, it was decided to create a new primer set. Furthermore, it was agreed with our collaborators at the Section of Genomic Medicine (Department of Medicine, Imperial College London) for me to be able to sequence the *rplF* libraries using their MiSeq, which would prove helpful in troubleshooting any issues arriving.

4.5.2. rplFV2 primer set development

The main difference included in the rplV2 primer set was the addition of the link sequences between the pad and primer sequences, as described in Figure 18. The link sequences are designed to prevent interactions between other sequences in the oligonucleotide and the target DNA, apart from the PCR primer. This is achieved by generating an anti-complementary sequence to the three upstream bases of the primer binding region in the target DNA. In the case of *rplF*, the three upstream bases for the Forward and Reverse primers are 5'-ACC-3' and 5'-TAT-3', respectively. By using the sequences 5'-GCA-3' and 5'-GCG-3', correspondingly, we ensure that only the primer sequence will bind to the *rplF* gene, and that the pad sequences do not interact with the gene, avoiding possible biases caused by upstream sequences in the oligonucleotide. The sequences and Tm of the sequencing and indexing primers of the rplFV2 primer set are shown in Table 7 and Table 8 . As described for the rplFV1 primer set, the Tm of these primers was also kept above the recommended temperature for MiSeq compatibility.

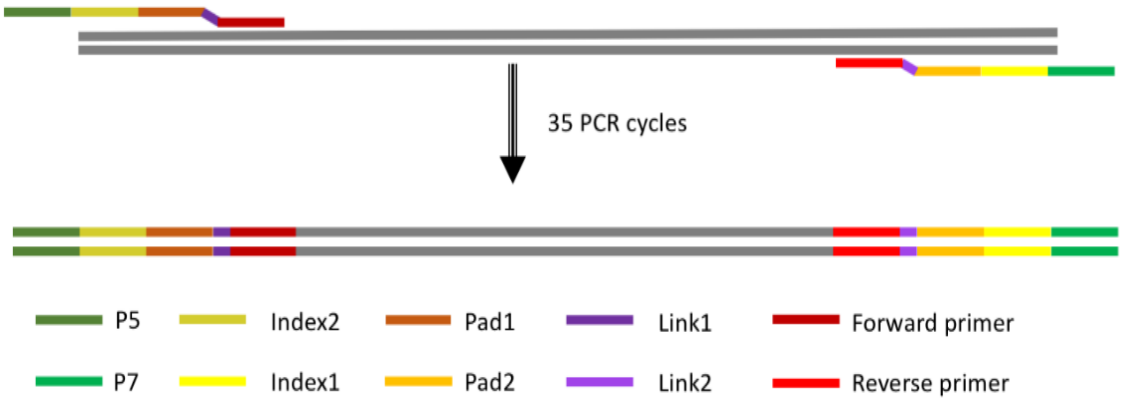


Figure 18 PCR and primer set up for the rplFV2 primer set. P5 and P7 sequences are shown in green, index sequences in yellow, pad sequences in orange, link sequences in purple and primer sequences in red, amplicon sequences are shown in grey.

Table 7 Comparison of sequences making up the rplFV1 and rplFV2, all sequences shown are in the 5' to 3' direction.

Segment	rplFV1	rplFV2
P5 adapter	AATGATACGGCGACCACCGAGATCTACAC	
P7 adapter	CAAGCAGAAGACGGCATACGAGAT	
Index1	TAAGGCGA, CGTACTAG, AGGCAGAA, TCCTGAGC, GGACTCCT, TAGGCATG, CTCTCTAC, CAGAGAGG, GCTACGCT, CGAGGCTG, AAGAGGCA, GTAGAGGA	TAAGGCGA, CGTACTAG, AGGCAGAA, TCCTGAGC, GGACTCCT, TAGGCATG, CTCTCTAC, CGAGGCTG, AAGAGGCA, GTAGAGGA, GCTCATGA, ATCTCAGG
Index2	TAGATCGC, CTCTCTAT, TATCCTCT, AGAGTAGA, GTAAGGAG, ACTGCATA, AAGGAGTA, CTAAGCCT	CTCTCTAT, TATCCTCT, GTAAGGAG, ACTGCATA, AAGGAGTA, CTAAGCCT, CGTCTAAT, TCTCTCCG
Pad1	TCCGGTCGCG	GCCGTGCTCGTT
Pad2	AGGCGGCCG	CCAGCCACTGCCGC
Link1		GCA
Link2		GCG
Forward primer	CAGTGACTGTTCCCGCTGGTGT	
Forward primer	CAGTGACTGTTCCYGCTGGTGT	
Reverse primer	AGGYTCAGGAGKWCGGAAHG	

Table 8 Sequence and melting temperatures for the rplFV2 sequencing and indexing primers. All temperatures are shown in ºC. For each sequence, the minimum calculated Tm (Min Tm), the average Tm (Avg Tm) and highest possible (Max Tm) is shown. The minimum and maximum Tm values are shown only for those sequences with ambiguous bases in its sequence.

Primer	Sequence	Min Tm	Avg Tm	Max
Forward primer	GCCGTGCTCGTTCAGTGACTGTTCCCGCTGGTGT		73.7	
Forward primer (modified)	GCCGTGCTCGTTCAGTGACTGTTCCYGCTGGTGT	72.4	73.0	73.7
Reverse primer	CCAGCCACTGCCGCAGGYTCAGGAGKWCGGAAHG	71.2	73.2	75.5
Indexing primer	CDTTCCGWMCTCCTGARCCTGCGGCAGTGGCTGG	71.2	73.2	75.5

As a summary of the MiSeq sequencing experiments performed using the rplfV2 primer set, the results from the experiment (from now on referred to as 'runs') 1,3, 5 and 6 will be explained, and 2, 5, and 7 will be compared to the other runs that had similar results. The overall set-up of the rplFV2 MiSeq experiments are described in Table 9. To try to improve the results obtained, different concentrations of PhiX and library were tested. Also, two different MiSeq kits, the V3 2x300 and V2 2x250 kits were used to try to sequence the full length of the *rplF* amplicon. All libraries were prepared as described in Chapter 2 (Section 2.10).

Table 9 Summary of MiSeq sequencing experiments for the rplFV2 runs. For each sequencing run, the library used, The MiSeq kit version, the percentage of PhiX spike in the final library, and the concentration, in pM, of of the PhiX spike and library used is shown.

Run	Library	Kit version	PhiX spike (%)	PhiX (pM)	Library (pM)
1	1	V3 2x300	20	8	8
2	1	V3 2x300	20	8	8
3	2	V3 2x300	20	6	6
4	2	V3 2x300	20	4	4
5	2	V2 2x250	30	2	2
6	3	V3 2x300	30	12.5	4
7	3	V2 2x250	30	12.5	4

4.5.2.1.1. Results for the rplFV2 sequencing run 1

For the first library tested, the Bioanalyzer trace and qPCR results were satisfactory, with a single peak on the Bioanalyzer trace of 548 bp, and a R^2 = 0.9987. The library was quantified to a final concentration of 30.53 pM, with a standard deviation of 0.4. After the MiSeq finished the sequencing cycles, a cluster density of 1056±32 1000 clusters/mm^2 was obtained, which is within the expected range for a V3 kit (1000 to 1200 clusters per mm^2). However, the quality obtained was much lower than desired for the Reverse read. Figure 19 shows the Q scores per base, 1-300 are the first 300 bases

from the 5’ end of the amplicon, 301 to 316 the index reads, and 317 to 618 the first 300 bases from the 3’ end of the amplicon. Furthermore, the percentage of PhiX identified for the Forward and Reverse reads were 10±1.8 and 0%, which were values lower than the 20% expected. The results obtained for the Run2 were very simillar to the Run1 results.

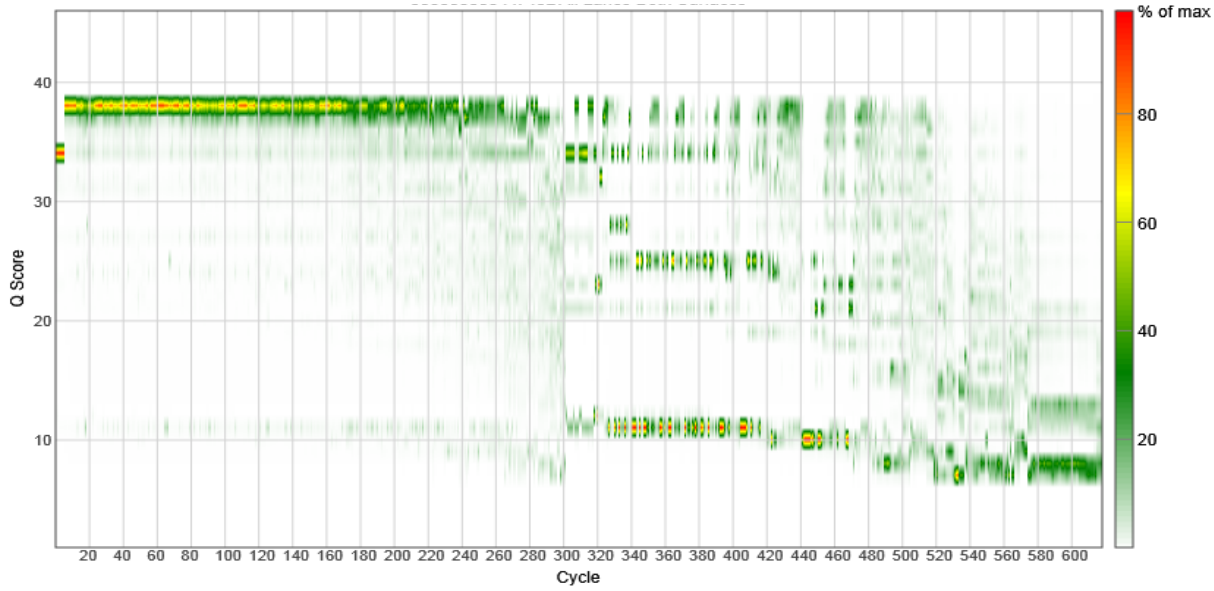


Figure 19 Quality per base obtained for all the bases sequenced for the first sequencing run for the rplFV2 primer set. The X axis indicates the bases sequenced during the sequencing steps in a MiSeq sequencing experiment, and the Y axis the mean sequence quality score. Segments are colored according the percentage of reads that shared that same quality scores for each base.

With this is mind, it was decided that it was possible that the underlying issue in the two runs was an over clustering of our library. This could happen either due to underestimation in the library quantification or due to the library binding preferentially to the flow cell compared to the PhiX. As there is very little space in between clusters, MiSeq software struggles to identify the boundries separating them. This can lead to the machine identifying two neighbouring clusters as one, which will lead to low quality, as the instrument will eventually detect different signals coming fron the two clusters as if it were coming from one. The instrument assumes that each clusters is formed by identical molecules, and if it identifies two or more different nucleotides coming fron the same cluster,

it will assign a low quality to that base. To solve this problem, the next libraries were loaded into the MiSeq at a lower concentration.

4.5.2.1.2. Results from sequencing run 3 using the rplFV2 primer set

The results from the QC steps taken before loading the MiSeq were also successful for this run, again obtaining single peaks in the Bioanalyzer trace, this time at a size of 554 bp, an R^2 of 0.997 for the quantification, and the library was quantified to 24.2±0.51 pM. After the sequencing was finished, a cluster density of 992±36 1000 clusters/mm^2 was achieved, but only 41.4±40.1 % of these clusters passed quality filtering. Furthermore, only 4.1±4.0 % of the reads were matched to the PhiX genome, out of the 20% expected. The quality of the Forward read was similar to that of the first run, while the Reverse read was very variable (Figure 20), and due to the very poor sequencing quality, the full amplicon sequence could not be assembled.

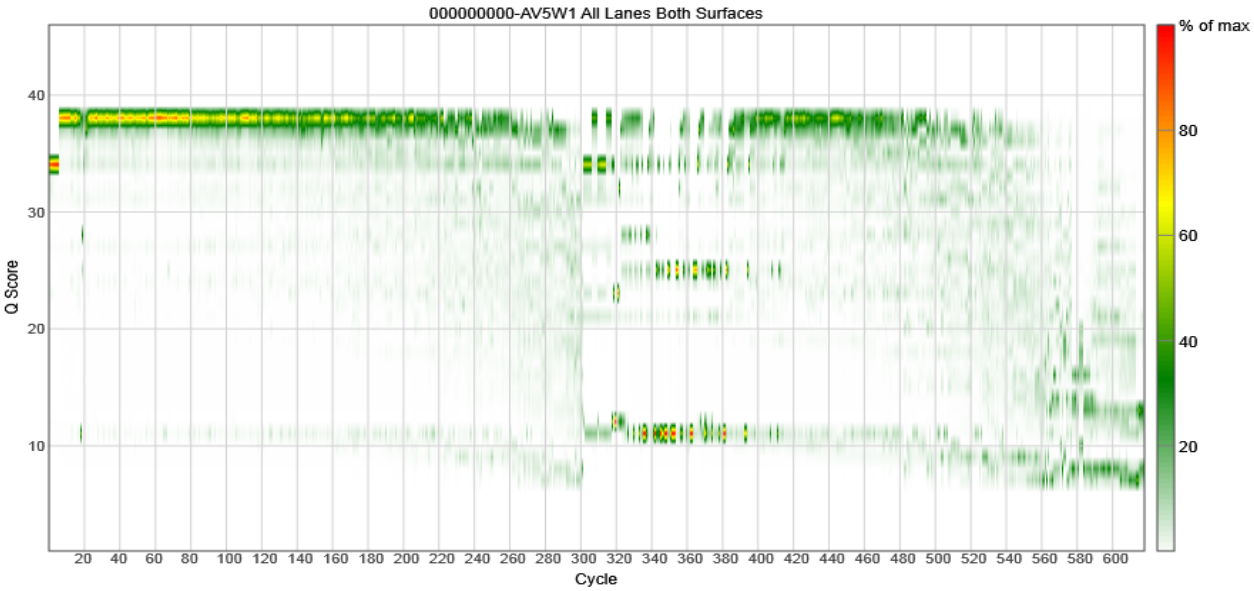


Figure 20 Quality per base obtained for all the bases sequenced for the third sequencing run for the rplFV2 primer set. The X axis indicates the bases sequenced during the sequencing steps in a MiSeq sequencing experiment, and the Y axis the mean sequence quality score. Segments are colored according the percentage of reads that shared that same quality scores for each base.

After further inspection of the potential issues identified in the runs so far, three major issues were identified. First, the indexing and Reverse reads had a very low quality, followed by a very low quality of the reverse read (Figure 21). Secondly, the cluster count during the sequencing of the Reverse reads was far more varied than the Forward reads (Figure 22). Finally, the phasing and prephasing values were relatively high, ranging from 0.123 and 0.71%, respectively, for the Forward read, and 0.54 and 0.244% for the Reverse read. The first two results pointed towards a potential problem during the steps at which MiSeq binds the amplicons by the P5 adapter, which happens in between the index sequencing steps. This could again be a secondary effect of over-clustering, as clusters in the MiSeq are thought to expand slightly during this shift from the P7 to the P5 adapter, if the flow cell is already very tightly clustered, this may cause the MiSeq to not delimit the boundaries in between clusters. The phasing and prephasing problem could also be a sign of a low diversity library.

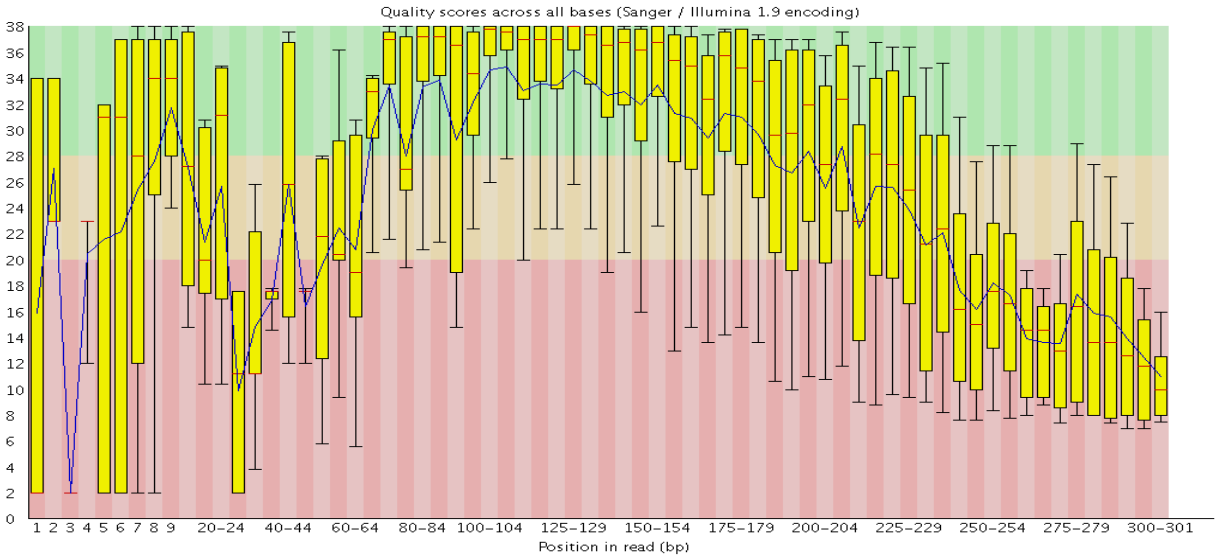


Figure 21 Quality scores per base reverser read for the third rplFV2 library sequenced. The X axis shows the base, or group of bases, sequenced in each direction, and the Y axis represents the quality of the scores. The red, yellow and green background segments indicate quality scores in between 0 and 20,20 and 20 and above 30, respectively. The sequencing quality scores of each base, or group of bases is represented with a boxplot (shown in yellow), with a red line indicating the median score. The blue line indicates the average quality value.

Another factor that could contribute to lower quality could also be an underlying problem with the indexing primer and the Reverse sequencing primer. As described in Figure 14, Index2 is

sequenced using the P5 adaptor as a primer, while Index1 is sequenced using one of the supplied primers. Both index sequences were sequenced poorly, we could exclude the indexing primer at least from being a problem, as the Index2 reads should not be affected by this, but they could be affected by over-clustering. Nevertheless, issues caused by the Reverse primer could still not be disregarded. For run 4, it was decided to use a 4 pM concentration for both the library and the PhiX, as opposed to the 6 pM concentration used in run 3, however, over-clustering issues still persisted.

4.5.2.1.3. Results for sequencing run 5

One problem that we could have faced so far was a potential problem with the V3 kits. After talks with Illumina representatives to try to solve some of the issues, it was suggested to try to sequence the *rplF* amplicons using a V2 kit, and increase the PhiX spike used to 30%. The rationale was that there were slight differences in the chemistry in the V2 kits compared to the V3, while the library preparation process was identical for both. As for previous runs, the QC steps were successful, obtaining single peaks in the Bioanalyzer trace (546 bp) and a library concentration of 27.04±3.1 pM. The cluster density for this run was lower than previous ones, 684±38 1000 clusters/mm^2, which is expected when using V2 kits, but still acceptable, and 68.3±13.9 % of these clusters passed quality filtering. Regarding PhiX reads, only 7.2±1.6% and 6.8±1.5% of the reads were aligned to the reference genome for the Forward and Reverse read, respectively, out of the expected 30%.

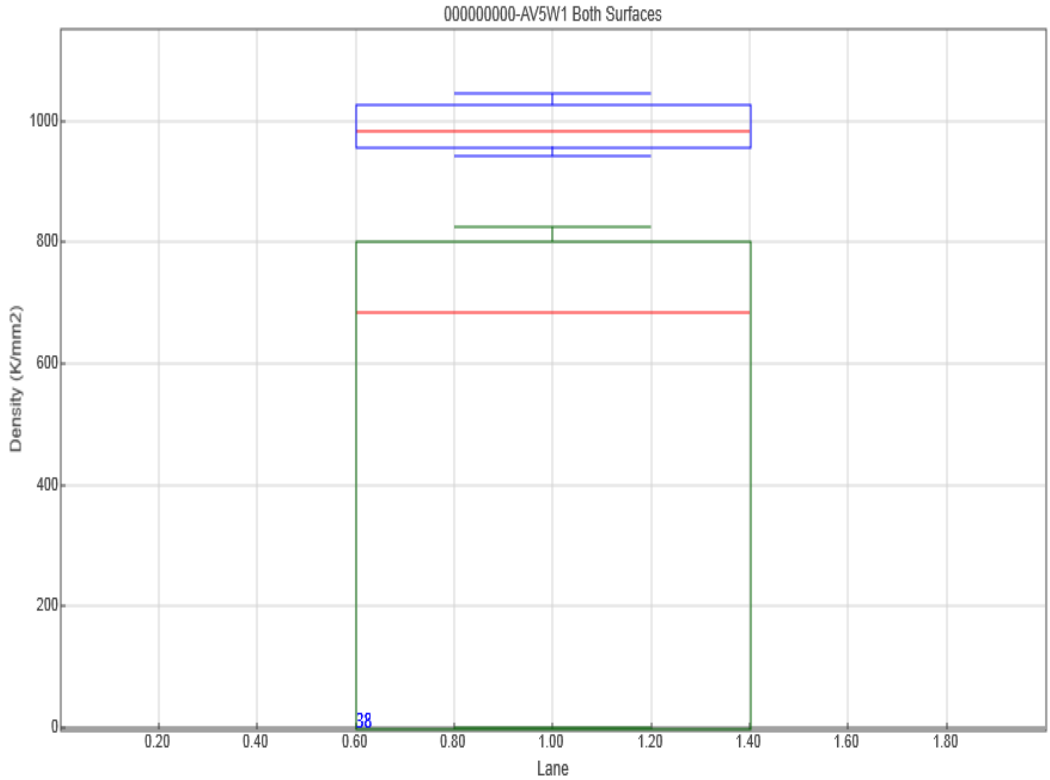


Figure 22 Boxplots showing the cluster counts for the Forward read (blue) and the Reverse (green). The Y axis indicates the density of clustering at a scale of 1000 clusters per mm^2.

The phasing and prephasing was 0% for both metrics for the Forward read, and 0.817 and 0.07%, respectively, for the Reverse read. These values point again towards the same problem experienced so far, *i.e.* there might be a really high library cluster density combined with low library diversity. The quality scores obtained for this run (Figure 23) show a more variable quality score for the Forward read, but still above the Q20 mark. On the other hand, the Reverse read, even though it was not optimal, saw a minor improvement in quality. At this point, the library, and PhiX, were loaded at a concentration of 2 pM, much lower than usual for most libraries, so there should not be a problem with over clustering.

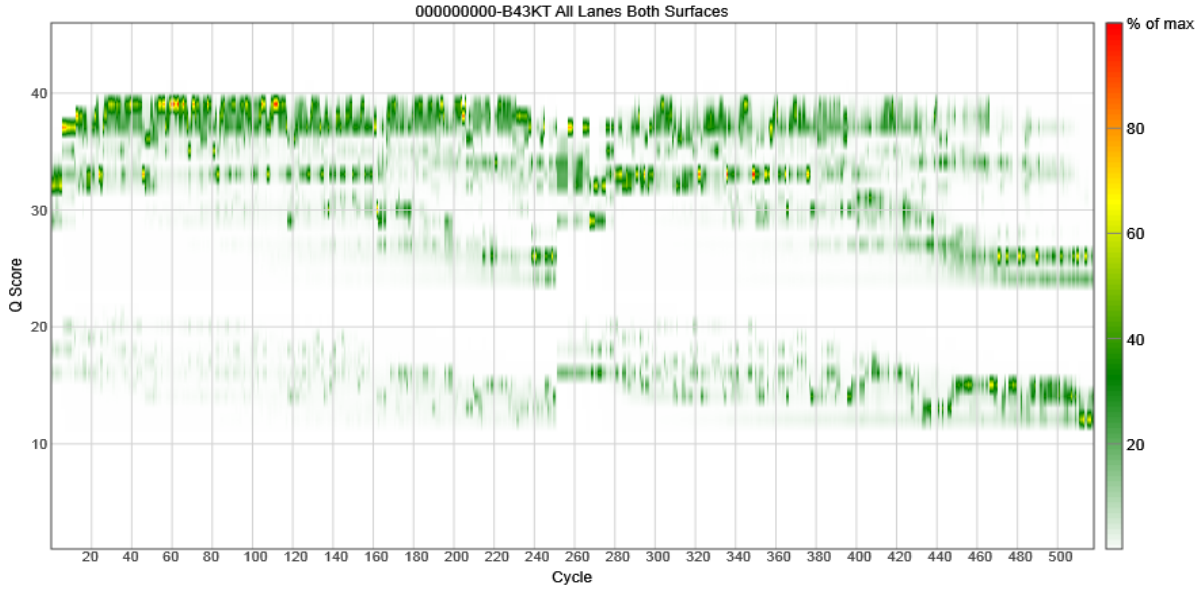


Figure 23 Quality score per base obtained for the fifth sequencing run using the rpIFV2 primers. The X axis indicates the bases sequenced during the sequencing steps in a MiSeq sequencing experiment, and the Y axis the mean sequence quality score. Segments are colored according the percentage of reads that shared that same quality scores for each base.

After further talks with Illumina representatives, it was suggested that our library might be binding preferentially to the flow cell compared to the PhiX, which may explain why we have seen a constant drop in the PhiX reads expected. It is considered that in a library pool containing DNA fragments of different sized, smaller fragments such as adapter or primer dimers would bind preferentially to the flow cell (137). This could mean that there could be primer dimers in the library preparations, or a secondary amplification that has not been identified but this idea was initially discarded, as Bioanalyzer traces only showed a peak of the estimated size (Figure 24). Our estimated library size was 540 bp, while PhiX has an average size of 500 bp according to Illumina. Based on these sizes, our library should not bind preferentially to the flow cell, and so at this point it is still not clear why our library continuously over-clustered on the flow cell. Regardless, for the next two sequencing experiments I tried increasing the ratio of PhiX to library to test any potential unaccounted problems caused by our library.

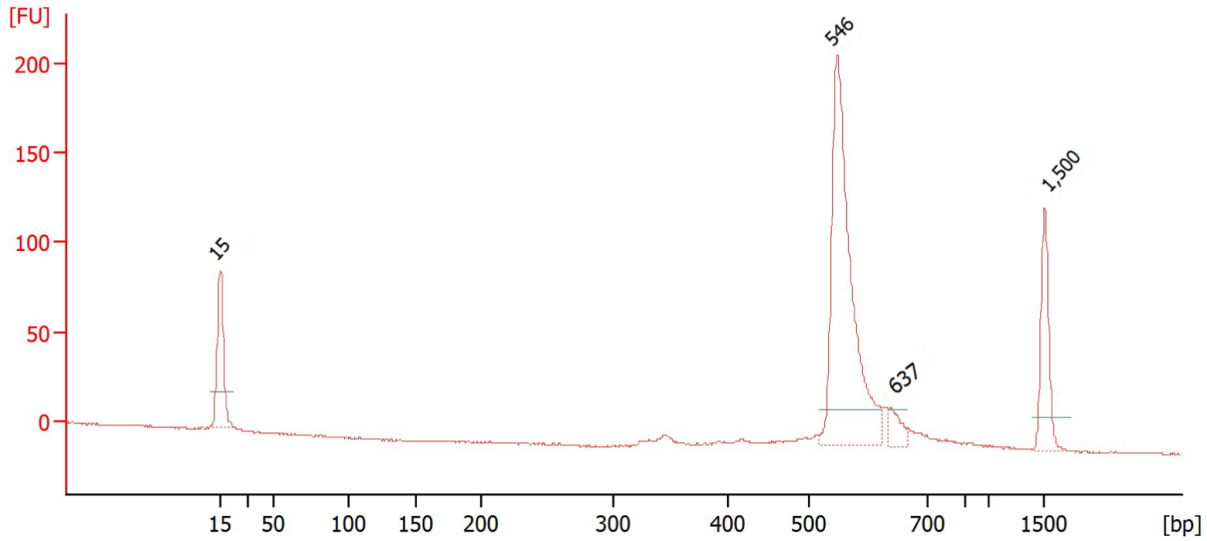


Figure 24 Bioanalyzer traces obtained for run 1. The X axis indictates the size of DNA fragments, measured in base pairs, and the Y axis the fluorescence units (FU) measured. A higher fluorescence intensity represents a higher DNA concentration. The peaks representing 15 and 1500 bp are the internal controls used in the experiment.

4.5.2.1.4. Results from library 6 and 7

For these two sequencing experiments a new library was prepared, and successful Bioanalyzer traces and qPCR quantifications obtained. The two libraries were sequenced using a 2x300 V3 and a 2x250 V2 kits, respectively. The V3 kit yielded a cluster density of 715±18 clusters per mm^2, 90.2±1.1 of these clusters passing quality filtering. However, there were still high phasing and prephasing values for the Forward (0.169 and 0.183%) and Reverse (0.602 and 0.167 %) reads. The percentage of PhiX reads was elevated to 17.6±0.9 for the Forward read, but 0% for the Reverse. On the other hand, the V2 kit had an 877±37 cluster density, with 85.9±1.9 of clusters passing quality filtering. Phasing and prephasing was lower for the Forward read (0.029 and 0.091 %) than the Reverse read (0.187 and 0.071 %). Yet, the quality for both of these sequencing runs was better than seen previously, but a sudden drop in quality was still observed towards the end of the end of the runs, which still did not allow the assembly of the *rplF* amplicon (Figure 25 A-B). Also, there was a substantial improvement in

the quality scores, especially for the Reverse read (Figure 25 C-D), showing a much lower variation range.

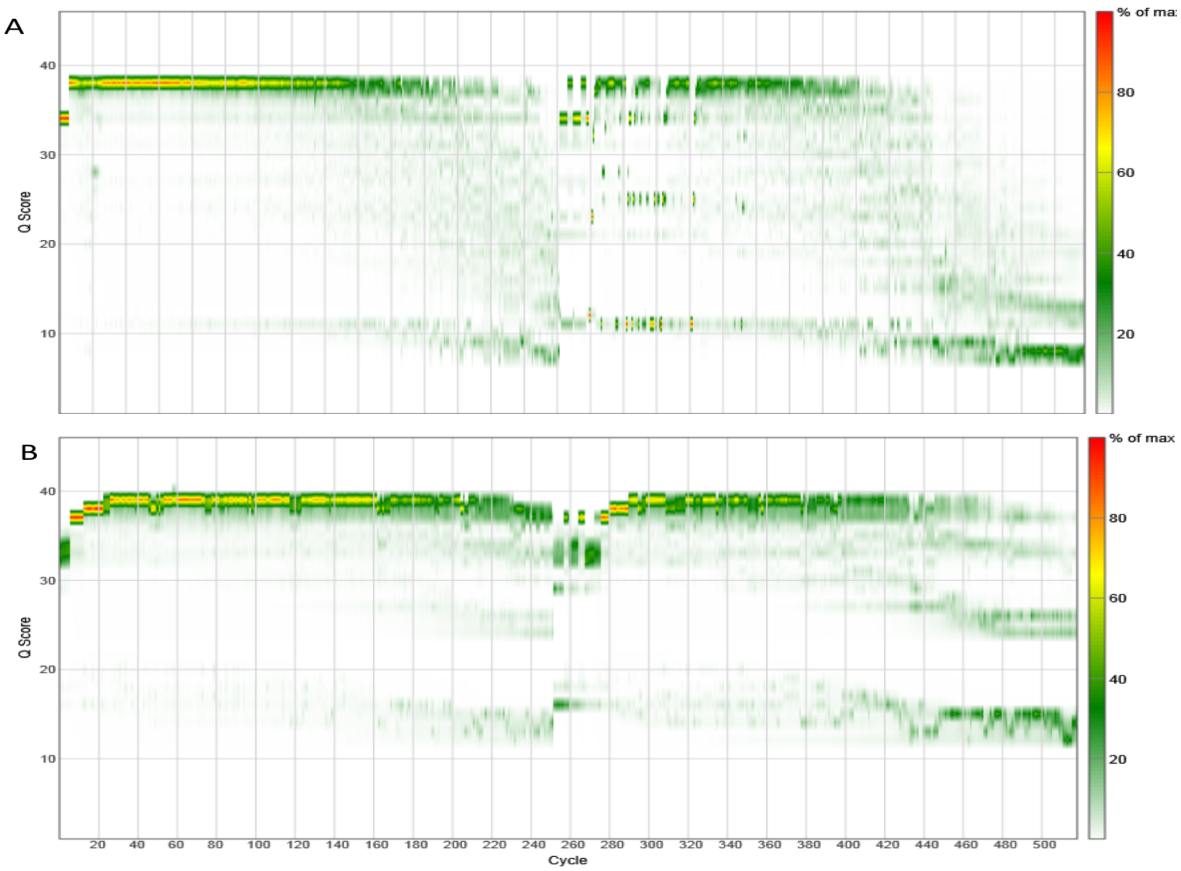


Figure 25 Quality scores per base obtained during sequencing using the V3 kit (A) and a V2 kit (B). The X axis indicates the bases sequenced during the sequencing steps in a MiSeq sequencing experiment, and the Y axis the mean sequence quality score. Segments are colored according the percentage of reads that shared that same quality scores for each base.

4.5.2.2. Summary of results for the rplFV2 primer set

The output obtained from the sequencing runs done up to this point was not satisfactory. There was no clear single reason as to why the experiments failed, but this might have been a combination of different issues that lead to the overall low quality. First, the diversity of these libraries prepared was low, but even following Illumina's recommendations for low diversity libraries, similar problems persisted throughout all the seven runs. Over-clustering, which is thought to be caused by a higher

affinity of the library compared to the PhiX control seemed unlikely, as the adapter sequences in both are the same, and the PhiX has on average smaller fragments than the *rplF* library. Most of the other issues found, such as the high and variable values of phasing and prephasing, could be a secondary effect of over-clustering, but they could also be a by-product of faulty reagents in the sequencing cartridges. There have been multiple reports of issues with 2x300 V3 Illumina kits in the past but given that this issue persisted when the library was tested with V2 kits, it might not be the cause of the problems found so far.

One final hypothesis proposed was that our primers interacted with the PhiX sequences. However, BLAST searches of our primers against the database showed no match in between the primer sequences (sequencing or indexing primers) and the PhiX genome. I also tested the PhiX genome in a PCR using the sequencing primers at an annealing temperature of 55ºC, the temperature of the *rplF* PCR, and 65ºC, the temperature at which the MiSeq functions, but no apparent bands could be identified for the PhiX sample. However, it is difficult to replicate the results that can be obtained during the amplification step in the MiSeq, as PCR reagents used (Q5 High-Fidelity 2X Master Mix, New England Biolabs) will be different from the buffers included in the MiSeq sequencing cartridge. As no cause was eventually identified, it was ultimately decided to abandon this approach to library preparation.

Two possible solutions were identified to try to solve the issues identified. One consisted of preparing the libraries using a two-step PCR, and the second one using the TruSeq Custom Amplicon kits supplied by Illumina. As mentioned earlier, using TruSeq kits not always a possible approach for larger scale options, so this was deemed a last-resort option. The two-step PCR options available have been used in other applications with great success (59). The basis for this method relies on first

amplifying the target DNA sequence using PCR primers with an overhang that that includes part of the HP10 and HP11 sequences in the Forward and Reverse primers, respectively. These sequences bind the primers included in the MiSeq cartridge and are used to sequence PhiX and libraries prepared using Illumina kits. This provides an advantage as there is no need to include custom sequencing and indexing primers in the sequencing cartridge, thus avoiding potential problems due to unforeseen interactions. Subsequently, a second PCR is performed on the purified products obtained from the first one amplification. This PCR is performed using primers with complementary sequences to the HP10 and HP11 sequences and are used to attach the index and adaptor sequences to the amplicon. This is detailed in Figure 26.

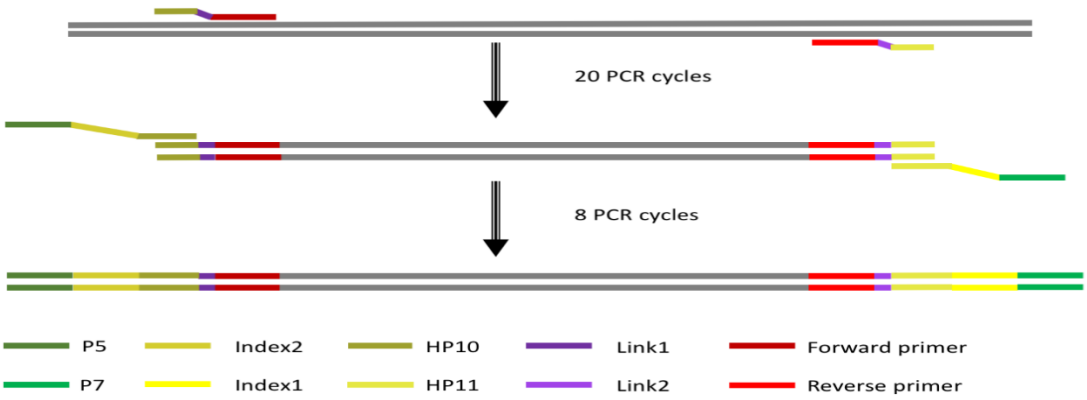


Figure 26 Diagram showing the two-step PCR approach for MiSeq library preparation. P5 and P7 sequences are shown in green, index sequences in yellow, pad sequences in orange, link sequences in purple and primer sequences in red, amplicon sequences are shown in grey.

Although the two-step approach seems promising, there are two potential obstacles that need to be considered. Firstly, the initial segments sequenced are the linker sequences (if included), followed by the PCR primer sequences, which should be identical for all the DNA molecules in the library, excluding the ambiguous bases included in the primers. The MiSeq uses the first seven bases to identify clusters (138), and this could prove problematic if all bases are identical. However, this should be offset by adding PhiX to the library. Second, the overlap shared between the two reads would be smaller, decreasing from 196 bp (estimated for the rplFV2 set) to 154 bp.

Another solution is using staggered primers to artificially increase the sequence diversity. With this solution, the protocol is identical to that for the two-step PCR, but it includes a spacer region of different size for each primer, which allows each DNA molecule to be sequenced a few bases ahead or behind the expected base, artificially increasing the diversity in each sequencing cycle. A graphical representation of the two-step PCR approach shown in (Figure 27).

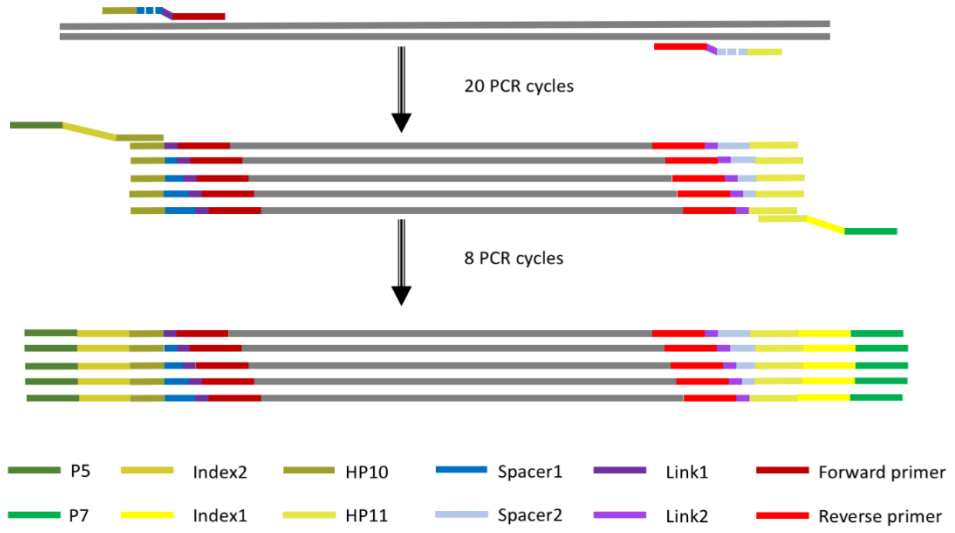


Figure 27 Schematic representation of the two-step PCR approach including a spacer segment. P5 and P7 sequences are shown in green, index sequences in yellow, pad sequences in orange, spacer sequences in blue,link sequences in purple and primer sequences in red, amplicon sequences are shown in grey.

The two-step PCR approach could bypass potential problems with diversity early on in the sequencing steps, as the spacer sequences can be designed to achieve complete base diversity and balance. This is especially relevant for the first seven cycles, as MiSeq uses these cycles to stablish the clusters present in the flow cell. As an example, I will describe the process I used to design spacer sequences that could accommodate *rplF* sequences, taking into account the linker sequences. There are several combinations that could be used to achieve a staggered library design, which could depend on whether or not a library of the same size is desired, and if a two-step PCR is used to generate this library. If a two-step PCR is used, one could design eight different primer versions for the Forward

primer, but then another set of eight different Reverse primers are required for the Reverse primer to compensate for the spacer length of the Forward primer. However, if a single-step PCR is used, 104 primers will be necessary, since eight different spacers could be used for the Forward primer, but 96 primers would be necessary to account for the index and spacer sequences, since each one of this would be a unique combination. Using all these primers would not be necessary if the library size is disregarded, but as mentioned earlier, the difference in size of the DNA molecules present in the library could skew sequencing results (137).

Considering the two-step, staggered library design, an example is explained in Table 10 and Table 11 . The spacer sequences can be designed so that the total count of nucleotide per sequencing cycle is as balanced as possible which, in addition to the variation added by the PhiX sequences, may be especially important for the first seven sequencing cycles. With the two-step PCR designs in mind, it was decided to test first test the standard two-step PCR approach, and should this fail, try the staggered PCR approach.

Table 10 Sequences designed for the Reverse primer spacers for a two-step, staggered *rplF* library design. Proposed spacer sequences are shown in black, linker sequence in purple, and spacer sequences in red. Nucleotides, and the total count for each nucleotide is displayed per sequencing cycle.

	Sequencing cycle										**Spacer sequence**	**Spacer length (bp)**
Spacer F1	T	A	C	T	G	T	T	G	C	A	TACTGTT	7
Spacer F2	C	G	T	A	T	T	G	C	A	C	CGTATT	6
Spacer F3	A	C	A	G	T	G	C	A	C	A	ACAGT	5
Spacer F4	G	T	G	T	G	C	A	C	A	G	GTGT	4
Spacer F5	C	A	T	G	C	A	C	A	G	T	CAT	3
Spacer F6	A	T	G	C	A	C	A	G	T	G	AT	2
Spacer F7	T	G	C	A	C	A	G	T	G	A	T	1
Spacer F8	G	C	A	C	A	G	T	G	A	C		0
Total A	2	2	2	2	2	2	2	2	3	3		
Total C	2	2	2	2	2	2	2	2	2	2		
Total T	2	2	2	2	2	2	2	1	1	1		
Total G	2	2	2	2	2	2	2	3	2	2		

Table 11 Sequences designed for the Forward primer spacers for a two-step, staggered *rplF* library design. Proposed spacer sequences are shown in black, linker sequence in purple, and spacer sequences in red. Nucleotides, and the total count for each nucleotide is displayed per sequencing cycle.

	Sequencing cycle										Spacer sequence	Spacer length (bp)
Spacer R1	A	G	A	T	A	T	A	G	C	G	AGATATA	7
Spacer R2	G	T	C	A	C	T	G	C	G	A	GTCACT	6
Spacer R3	T	A	T	C	T	G	C	G	A	G	TATCT	5
Spacer R4	A	C	A	T	G	C	G	A	G	G	ACAT	4
Spacer R5	C	A	T	G	C	G	A	G	G	Y	CAT	3
Spacer R6	C	T	G	C	G	A	G	G	Y	T	CT	2
Spacer R7	T	G	C	G	A	G	G	Y	T	C	T	1
Spacer R8	G	C	G	A	G	G	Y	T	C	A		0
Total A	2	2	2	2	2	1	2	1	1	2		
Total C	2	2	2	2	2	1	1.5	1.5	2.5	1.5		
Total T	2	2	2	2	1	2	0.5	1.5	1.5	1.5		
Total G	2	2	2	2	3	4	4	4	3	3		

4.5.3. Preparation of the rplFV3 primer set

Following the structure described in Figure 26, two sets of primers were generated, one to amplify the *rplF* gene, and the second one to attach the adapter, index, and HP10 and HP11 sequences. The sequences of all these fragments are described in Table 12. To further improve the quality of the sequencing for the index reads, I decided to generate custom index sequences that would be balanced at every base, to ensure that the quality of the sequencing could be as good as possible, instead of just picking primers from the list published by Illumina. To achieve this, I used python scripts described by Meyer *et al*. (139). Using the scripts described, I was able to generate a set of 16 index sequences for the Index2 and 24 indexes for Index 1, which was sufficient to create libraries for 192 samples.

Table 12 Sequences used to construct the rplFV3 PCR primers.

Segment	Sequence
P5	AATGATACGGCGACCACCGAGATCTACAC
P7	CAAGCAGAAGACGGCATACGAGAT
Index2	ACTATCTG, TAGCGAGT, CGTGAGTG, GGATATCT, CTACTATA, TACGAGAC, ACGTCTCG, GTCAGATA, CAGTCGGA, TAGAGCCT, ACCATTAT, CGATGCGC, TGACCAGC, ACCGTTCG, GTTGCCAC, GTTCAGAA
Index 1	TAAGGCGA, CGTACTAG, AGGCAGAA, TCCTGAGC, GGACTCCT, TAGGCATG, CTCTCTAC, CGAGGCTG, AAGAGGCA, GTAGAGGA, GCTCATGA, ATCTCAGG, AACCGCCA, ACTATGAG, ATGGTACT, CAACCATC, CCTCTCAT, CTCAGTCT, GATTACTT, GCGATACG, GTTAAGTT, TCAGATTC, TGCTCTGC, TGCTCTGC
HP10	ACACTCTTTCCCTACACGACGCTCTTCCGATCT
HP11	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
Link 1	GCA
Link 2	GCG
Forward PCR primer	CAGTGACTGTTCCCGCTGGTGT
Modified Forward	CAGTGACTGTTCCYGCTGGTGT
Reverse PCR primer	AGGYTCAGGAGKWCGGAAHG

4.5.3.1. Results from libraries generated using the rplFV3 primer set

To test the new rplFV3 primer set, two different libraries were prepared using the artificial mixture sample, one using the standard Forward *rplF* and the other using modified (rplF_Fm) primer, following the protocol described in Chapter 2 (Section 2.10). These libraries will be referred to as Library 1 and Library 2, respectively. As it was still not clear what the best combination of library and PhiX to use to sequence our libraries, I opted to use the combination that gave the best results with the rplFV2 primer set, 12.5 pM PhiX at and 4 pM library concentrations, having a final spike of PhiX of 20%.

After obtaining successful QC result for the Bioanalyzer trace and qPCR quantification, the libraries were sequenced. These two libraries yielded much better quality scored than those of the rplFV2 primer set (Figure 28), scores for Library 2 are not shown as they were almost identical to Library 1. A summary of the run statistics is shown in Table 13.

Table 13 Run metrics for rplFV3 Libraries 1 and 2. For each library and sequencing direction, the cluster density (in 1000 clusters/mm^2), the number of clusters passing quality filtering (CPF), the values of phasing and prephasing, the percentage of reads above a mean quality score of 30 (Q>30), and the percentage of PhiX reads identified (PhiX reads (%)) is displayed.

METRIC	**LIBRARY 1**		**LIBRARY2**	
READ	Forward	Reverse	Forward	Reverse
CLUSTER DENSITY	606±10		590±6	
CPF	92.0±0.7		93.1±0.5	
PHASING/PREPHASING	0.109/0.094	0.027/0.010	0.233/0.083	0.071/0.112
Q>30	92.8	83.1	88.8	81.6
PHIX READS (%)	39.0±1.1	38.6±1.1	43.9±0.8	43.9±0.8

One potential problem considered was the relative low cluster density of both libraries. Compared to the previous clustering values obtained, this seems relatively low, although it makes more sense to have a lower clustering density when using a relatively low library concentration. Also, the percentage of clusters passing quality filtering (CPF) and reads with an average quality above 30 was higher than those obtained with rplFV2. Furthermore, the aligned PhiX was much higher than seen previously, but still in line with the expected amount estimated.

After processing and analysing the data for these two runs, it was observed that the samples in Library 2 (amplified using the modified Forward primer) were more representative of their expected proportion, and thus, this primer was used to generate the rest of the libraries prepared. The

development of the pipeline to analyse these reads is discussed later in Chapter 5, and the analysis of the reads in Chapter 6.

All 10 libraries sequenced using the rplFV3 primers yielded good results. A summary of the output statistics is shown in Table 14. One trend observed throughout the sequencing experiments is the variation in yield and quality obtained based on the amount of PhiX spiked into the libraries. Starting with the last ratio tried for rplFV2 primers (30% of 12.5 pM PhiX and 4 pM library) yielded good quality results, but a low yield. After this result, I decided to test slightly different concentrations of PhiX, reducing the amount included in each run, which lead to a higher cluster density and yield. The final concentration used for the majority of the sequencing experiments was a 20% spike of a 12.5 pM PhiX dilution to an 8 pM library. This allowed successful sequencing of the libraries, with a high number of high-quality sequences (example show in *Figure 29*) being obtained. As suggested by Illumina, from library 8 onwards, I performed a heat denaturation step on the library-PhiX mix before loading the MiSeq cartridge. This helped enhance the quality of the sequencing, especially for the last sequencing cycles.

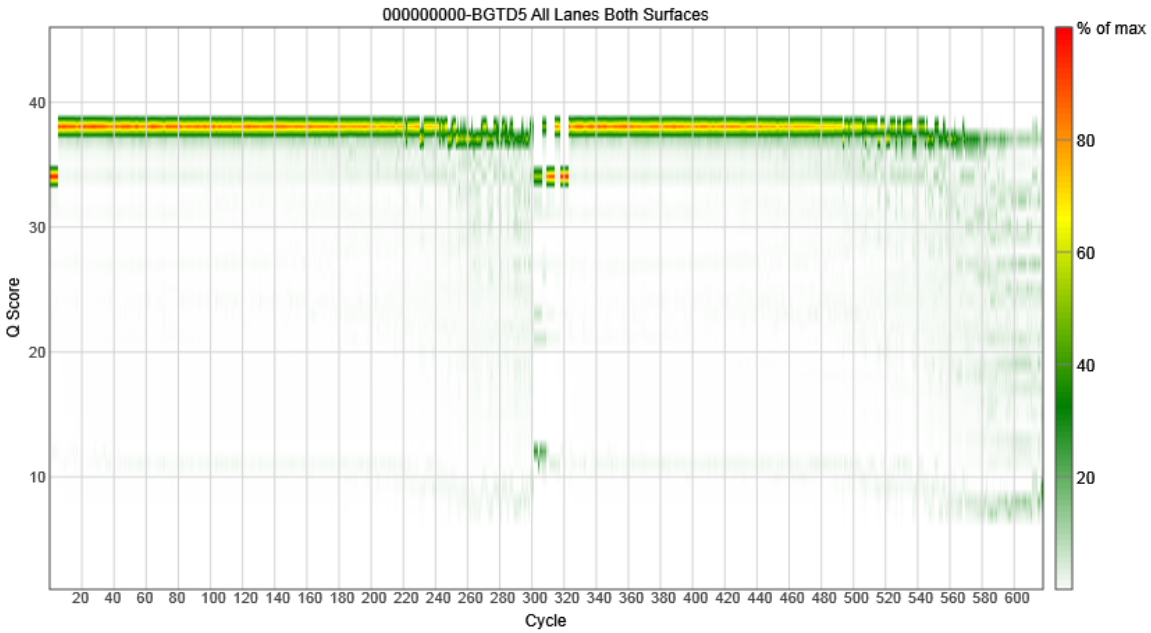


Figure 28 Quality scores per sequencing cycle for the rpFV3 Library1. The X axis indicates the bases sequenced during the sequencing steps in a MiSeq sequencing experiment, and the Y axis the mean sequence quality score. Segments are colored according the percentage of reads that shared that same quality scores for each base.

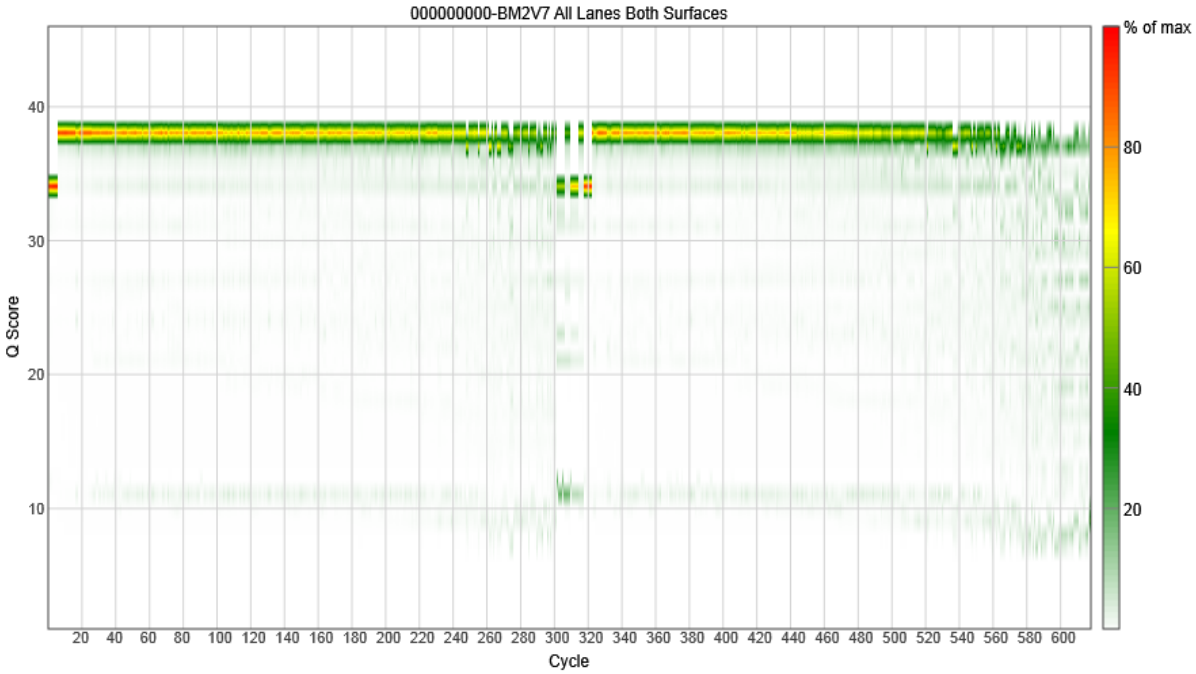


Figure 29 Quality scores per sequencing cycle for the rpFV3 Library 3. The X axis indicates the bases sequenced during the sequencing steps in a MiSeq sequencing experiment, and the Y axis the mean sequence quality score. Segments are colored according the percentage of reads that shared that same quality scores for each base.

Table 14 Sequencing statistics from 12 libraries prepared using human oropharyngeal throat swabs. For each library, the percentage PhiX spike, the library concentration used (pM), the cluster density (in 1000 clusters/mm^2) and number of clusters passing quality filtering (CPF), the values of phasing and prephasing, the percentage of reads above a mean quality score of 30 (Q>30), and the percentage of PhiX reads identified (PhiX reads (%)) and sequencing yield (in Gb) is displayed.

Library	PhiX spike (%)	Library concentration (pm)	Cluster density	CPF	Phasing/prephasing Forward read	Phasing/prephasing Reverse read	Average Q>30	*PhiX reads*	Yield (Gb)
3	30	4	480±10	96.8±0.3	0.275/0.113	0.105/0.10	81.1	63	7
4	20	4	721±3	95.4±0.7	0.193/0.410	0.075/0.164	83.8	43.7	7.9
5	20	6	606±10	92.0±0.7	0.109/0.094	0.027/0.010	93.2	38.8	8.8
6	20	8	649±16	91.6±0.7	0.141/0.401	0.016/0.00	80.35	24.15	10.5
7	20	8	860±27	95.1±0.4	0.219/0.390	0.012/0.001	83.4	24.8	13
8	20	8	934±35	91.1±2.8	0.134/0.136	0.176/0.018	74.4	14.5	14
9	20	8	838±32	89.1±15.0	0.241/0.396	0.138/0.00	85.0	23.5	11.9
10	20	8	943±34	93.8±2.0	0.258/0.466	0.083/0.022	81.5	22.7	14.1
11	20	8	867±28	93.8±1.4	0.244/0.406	0.055/0.004	81.35	24.2	13
12	20	8	829±33	93.6±1.0	0.160/0.299	0.059/0.00	78.1	24.6	12.3

4.6. Discussion

The development of a new protocol to apply the *rplF* sequencing strategy to an NGS platform proved to be a much more arduous task than expected. Even though Illumina has made the sequences used in their applications publicly available, there is no support or guarantee from the company towards the development of new applications using their platforms that do not rely on Illumina supplied kits for library preparation. After exploring the most common approaches to library preparation, single-step and two-step PCRs, I was able to sequence *rplF* amplicons obtained from DNA extracted from human oropharyngeal throat swabs.

The experience of sequencing the first libraries tested, generated with the rplFV1 primer set, gave no real explanation as to what could have been the underlying issue. The first sequencing experiments provided contradictory results, as one library could not be sequenced because the amplicons did not bind to the flow cell, but another one, prepared using the same primer set and samples, was able to bind to the flow cell. These results may hint at a technical error during the library preparation. The issues faced with the second batch of libraries sent was different, although more consistent with the results obtained for the rplFV2 libraries. However, the quality results obtained were slightly worse, having a more pronounced drop in quality for the Forward reads. Although it is still unclear why these libraries were not successful, and obviating possible technical problems, the same issues faced while testing the rplFV2 libraries may be responsible for the failure in this set of libraries.

For the rplFV2 libraries, even though the libraries did not yield completely satisfactory results, the quality seemed to improve overall slightly. The overarching issues identified, predominantly attributed to over-clustering, seemed to improve when drastically increasing the PhiX vs library ratio,

and marginally by using a V2 sequencing kit. The underlying cause of over-clustering is still not known. Two of the causes of low quality, suggested by Illumina, were the presence of smaller DNA fragments, such as primer dimers, present in our libraries, or poor library quantification. Both library QC steps were satisfactory, as a predominant peak of the expected size were identified for all reads, and the library concentration, and the library dilutions tested in the qPCR yielded consistent results. Another possibility is that our primers could interact with PhiX sequences, which could lead to poor sequencing results. Even though I did not find any sequences in the PhiX genome that could bind the primers used, either by mapping the primer reads to the PhiX genome or by PCR, this view cannot be discarded. The primer sequences could have a secondary structure that could increase its binding to the PhiX genome, although this is unlikely as the temperature at which the MiSeq functions is relatively high. It is also possible that some issues could have been caused by inherent issues of the V3 kits, as there have been several reports, of drops in quality in later cycles, but as these problems persisted when using the V2 kits, it is also unlikely that this was the cause. This approach was ultimately abandoned to pursue a new library preparation method using the two-step PCR approach, which yielded much better and consistent results.

The rplFV3 set of libraries provided much better results than the two previous ones. Although the two-step PCR approach is slightly more elaborate than the standard protocol, the results obtained with this approach were far superior than those obtained with previous protocols. A major factor that may be responsible for the successful approach is the use of the built-in sequencing and indexing primers present in the sequencing cartridge. Using these primers evades possible non-specific interactions other added primers may have and could hint that the issues in the rplFV2 set could have been related to them. Nonetheless, all the libraries tested were considered as sufficient quality, often surpassing the 70% of bases above Q30 assured by Illumina. Although we can still appreciate a trailing

lower quality score, particularly for the Reverse reads, there were still enough reads with high-enough quality to be processed through the data processing pipeline.

5. Chapter 5: Development of a pipeline for the analysis of rplF fastQ files and pre-processing of 454 sequencing reads.

5.1.introduction

The development of NGS platforms has been accompanied by the development of new data formats and programs used to process and analyse the data generated. The two most prominent data processing packages used for the analysis of 16S rRNA amplicons are QIIME (93), and MOTHUR (140). These two programs allow the processing of output files from different sequencing platforms to an operational taxonomic unit (OTU) table, which summarizes the data obtained to counts of each OTU per sample. The main metric used to distinguish and classify sequencing reads into different OTUs is the percentage of identity to reference sequences, the most common threshold used being 97% similarity to assign the taxonomy of the reference sequence to a read (141).

Although the main goal of these two programs is to analyse 16S rRNA amplicons, they can be applied to process other types of amplicons. There are several protocols published for the analysis of 16S rRNA reads, either published by the developers of the programs or by other users, there is no concise guide as to how to apply these programs to other amplicons. However, developing a guide that could fit most amplicons may not be possible, as different methods may be required to fit the necessities of each amplicon.

In this Chapter I will describe the steps considered to establish a reliable pipeline that can be used to process the output fastQ files obtained after sequencing the *rplF* libraries in MiSeq instruments. I will also describe the steps taken to process the Roche 454 output files generated after sequencing 16S rRNA gene amplicons.

5.2. Aims

- Develop a protocol to process *rplF* amplicon reads in MiSeq output files.
- Process 454 sequencing data for subsequent microbiome analysis.

5.3. Summary of methods

A summary of the scripts used to process the Illumina MiSeq fastQ files and 454 qual and fna files is shown in Figure 30.

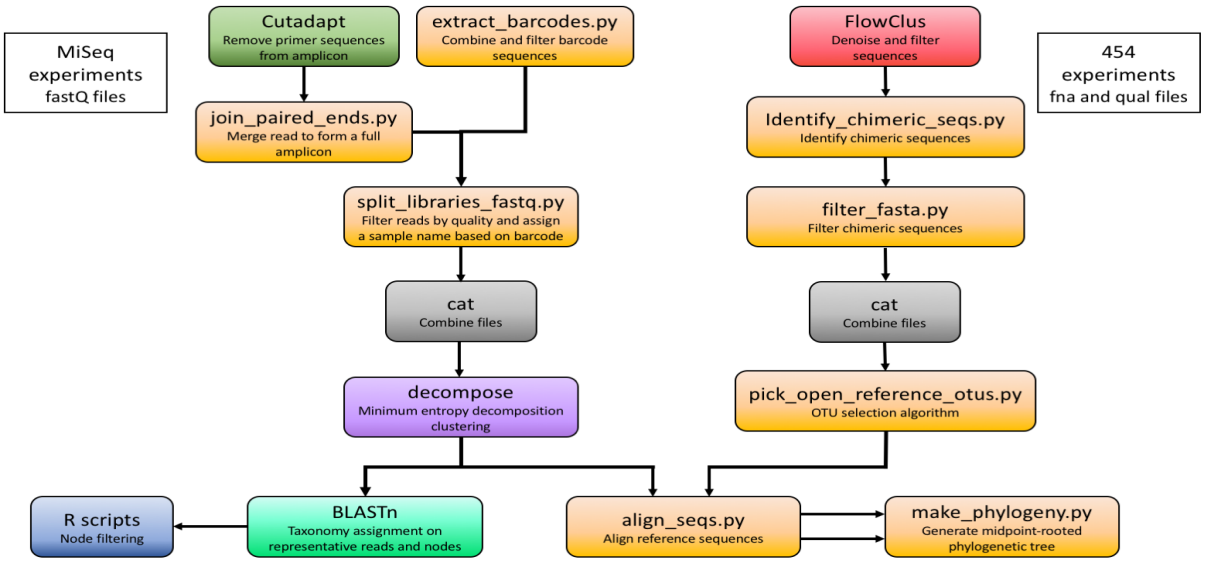


Figure 30 Summary of scripts used to process sequencing data files. The scripts used are shown in colour depending on the package or language used (QIIME = yellow, Cutadapt = dark green, FlowClus = red, Bash = grey, Oligotyping = purple, BLAST = light green, and R = blue).

5.4. Results

5.4.1. Proposed read processing pipeline

5.4.1.1. Data pre-processing using QIIME

Initial analyses of the sequencing data for the 12 MiSeq sequencing experiments used QIIME, as this was the main application used to process 16S rRNA sequencing reads. To process the amplicon reads, I trimmed the corresponding PCR primer sequences for each read using the cutadapt package (142), which also allows users to maintain those reads which have the primer sequences specified, eliminating PhiX reads in the process. Once the reads were trimmed, they were assembled into a full amplicon, using the join_paired_ends.py script, only keeping paired reads that had an overlap of at least 140 bp, allowing for a 5% variability on the overlapping region. It is possible to match the reads from of each cluster because MiSeq assign a X and Y coordinate to each cluster based on their location on the flow cell, and it is then passed on to each of the sequencing output files in the sequence ID section corresponding to each read. The scripts used take into consideration the coordinates in each read to assemble only reads obtained from the same cluster.

Index reads were filtered using the extract_barcodes.py script, which formats the index reads so that only those matching the ones in the sample sheets containing sample information, and therefore in the samples, are kept. Both barcode sequences are merged into a single 16 bp barcode. The final step performed before assigning a taxonomy to each read was to attach the full barcode sequences and sample names to the assembled reads, which was performed using the split_libraries_fastq.py script. In this process, sequences with an average quality below 20 were discarded.

The following step in the data processing pipeline was to assign a taxonomy to each read in the files and generate an OTU table. To achieve this, QIIME offers different routes, but the most commonly used is the pick_open_reference_otus.py script. In summary, the script does the following:

- Step 1: Align each sequence to a user-supplied set of reference sequences, keeping those with an identity above the selection threshold and the rest are filtered. This is also known as closed-reference OTU selection.
- Step 2: The sequences that failed to match the reference sequences in Step 1 are subsampled. These are aligned and clustered, and a reference sequence is selected for each cluster (the cluster centroid)
- Step 3: All the sequences that failed to match the database in Step 1 are then clustered using the Step 2 cluster centroids as reference sequences.
- Step 4: Sequences that failed to match the database in Step 3 are clustered using a *de novo* algorithm that does not rely on a database to cluster sequences, but the clustering is done among the sequences without a reference set of sequences.
- Step 5: The final OTU table is generated by concatenating the outputs from Steps 1, 3 and 4.
- Step 6: A taxonomy is assigned to each OTU based on its similarity to the reference sequences. A reference sequence is selected for each OTU, then these are aligned, and a phylogenetic tree is generated.

A parameters file is supplied to the pick_open_reference_outs.py script in which the user can specify different variables, such as the percentage thresholds used in steps 1, 3, and 4, the program

used to cluster the reads, the minimum length required for a match to be considered successful etc. By default, QIIME uses UCLUST (96) to align and cluster sequences into OTUs.

UCLUST works similarly to BLAST, but it is more efficient in terms of speed and memory usage due to its simpler algorithm. However, one of the potential disadvantages of UCLUST is that the best match identified is the first one found that meets the parameters specified, regardless of possible subsequent better matches identified in the reference sequence file. This disadvantage can be overcome by specifying an 'optimal' search in the parameters file, which forces UCLUST to match the reads to the best possible by aligning the queried read to all the sequences in the database. Such an approach can result in large memory usage and extended computational times. One can use UCLUST without the 'optimal' parameter if the database is properly formatted, and only if sequences that are further apart, in terms of percentage of identity, than the desired threshold are included. Such databases are provided for 16S rRNA sequences, examples being the SILVA database (143) or Greengenes (144). These databases are readily available and can be downloaded freely.

Developing a *rplF* database that is compatible with UCLUST proved challenging. As described previously, the variability in *rplF* amplicons is more limited compared to 16S rRNA reads, and thus the standard 97% threshold would not be optimal for this amplicon. For testing purposes, I randomly selected the results obtained from library 10 as a representative of all sequencing files. I processed the output files using the different clustering thresholds (97, 97.5, 98, 98.5, 99 and 99.5%) to assess the effectiveness of UCLUST. After processing the reads in QIIME, these were analyzed using the phyloseq package in R. An example of the output obtained is shown in Figure 31 which shows the species assigned to each sample when clustered at different thresholds. Some of the results obtained, such as samples 1442, 1478, 1580 and 1492 were relatively constant despite the threshold used, but

others like samples 1563, 1622, 1563 or 1446 seemed to shift in their species content depending on the values used. Furthermore, five samples contained reads that could not be assigned a species using this method (shown in light purple). The results obtained did not help establish an effective threshold that could be used to distinguish *rplF* sequences belonging to different *Neisseria* species.

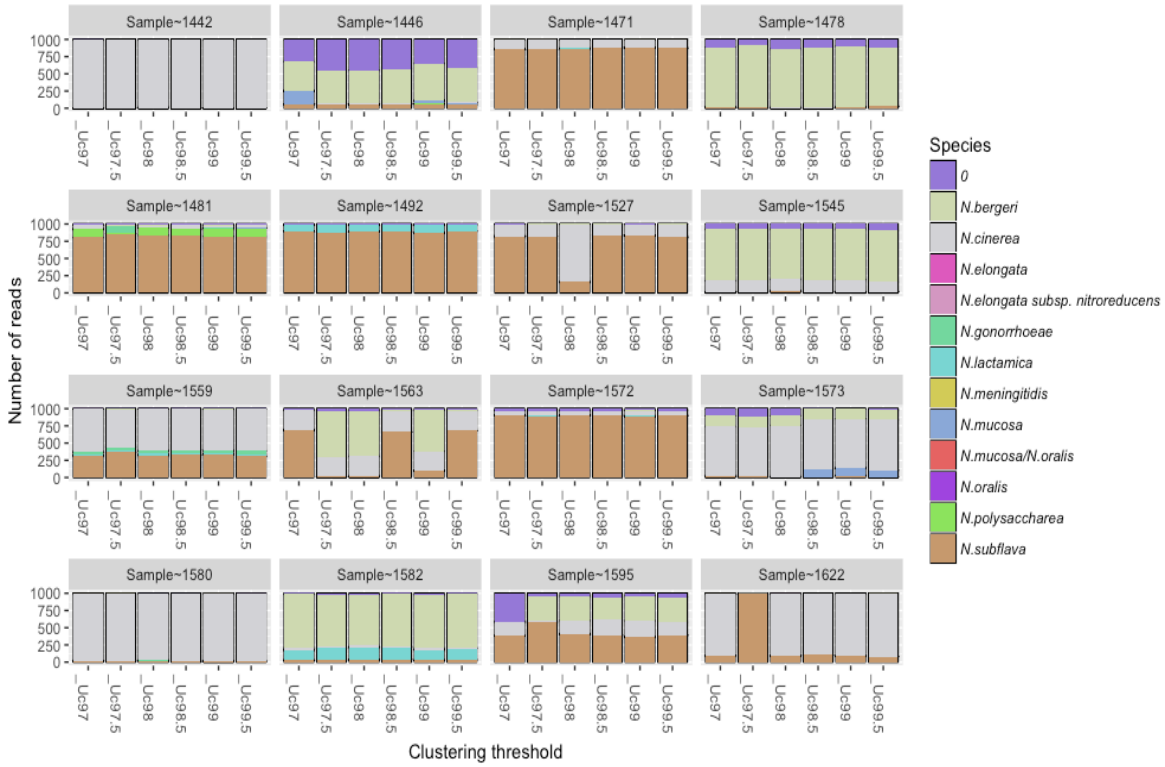


Figure 31 Results of 16 randomly selected samples obtained after processing the reads using different clustering thresholds. The X axis indicates the results for each of the UCLUST clustering thresholds used, and the Y axis the number of reads assigned to the different species, which are coloured according tto the legend on the left. Reads without a definite species assignment are classified as 0.

Another difficulty faced when trying to comprehend these results is that there is no reference as to what could be expected to be identified when sequencing *rplF* amplicons from oropharyngeal samples when using this method. To further characterize this method, I needed to identify potential samples that could be used as controls to test the methods used.

I could use the reads obtained from the two libraries prepared using artificial mixtures, but there was not a great sequence diversity among the reads obtained as there are only five different types of sequences, which I know were sufficiently different to be identified successfully. In order to solve this problem two solutions were considered: (1) Generate more artificial mixtures using DNA from isolates that belong to different *Neisseria* species but have very similar *rplF* amplicon sequences: (2) Generate synthetic samples using *rplF* reference sequences, using the most similar sequences. Ultimately, as generating new artificial mixtures would not provide any new information than what could be obtained from synthetic samples, the latter option was chosen.

5.4.1.1.1. Tests using synthetic samples

To generate the synthetic samples required, I created a file containing the reference number ID of the reference *rplF* sequences to be included in each sample, and the relative proportion of reads to use for each ID. Then I created a custom R script to which I can supply the reference *rplF* sequences in fasta format, the proportion of species, and the number of reads to simulate per sample, then the script creates a fasta file containing the synthetic samples specified as if they were obtained through the QIIME scripts described before. The fasta file containing reads for the synthetic samples was merged with the files from library 10 and were processed to ensure that the processing steps taken did not alter the results. Furthermore, the script also provides a taxonomy and OTU table that can be directly used with phyloseq and used as references of what should be expected. The species included in these mixtures were from two different *Neisseria* species that had a dissimilarity of six bases or fewer. For each combination of sequences (shown in Table 15), three different samples were generated using ratios of 0.05/0.95, 0.5/0.5 and 0.95/0.05. I also included one mixture, samples 40 to 42, that contains two sequences that are sufficiently different (*N. musculi* and *N. elongata* sequences 75 and 101 which are 86 bases apart) as a control.

Table 15 Combinations of *rplF* reference sequences used to create synthetic samples. Diff indicates the number of differences (Diff), in base pairs, in between the two sequences.

Sample	ID	Species	Diff (bp)	Sample	ID	Species	Diff (bp)
1-3	70, 79	*N. bergeri, N. spp*	2	**34-36**	121, 12	*N. cinerea, N subflava*	4
4-6	70, 57	*N. bergeri, N. subflava*	1	**37-39**	121, 130	*N. cinerea, N. subflava*	4
7-9	70, 58	*N. bergeri, N. subflava*	1	**40-42**	75, 101	*N. elongata, N. musculi*	86
10-12	16, 58	*N. bergeri, N. subflava*	2	**43-45**	37, 36	*N. spp, N. Subflava*	4
13-15	121, 40	*N. cinerea, N. subflava*	3	**46-48**	7, 90	*N. gonorrhoeae, N. polysaccharea*	5
16-18	121, 128	*N. cinerea, N. subflava*	3	**49-51**	1, 96	*N. meningitidis, N. polysaccharea*	5
19-21	37, 21	*N. elongata, N. mucosa or N. oralis*	3	**52-54**	66, 96	*N. meningitidis, N. polysaccharea*	4
22-24	79, 57	*N. spp, N. subflava*	3	**55-57**	35, 36	*N. mucosa, N. Oralis*	5
25-27	79, 58	*N. spp, N. subflava*	3	**58-60**	119, 102	*N. mucosa, N. Oralis*	9
28-30	69, 79	*N. bergeri, N. subflava*	2	**61-63**	79, 148	*N. spp, N. subflava*	4
31-33	62, 100	*N. bergeri, N. subflava*	4				

The fasta file containing the synthetic samples created was added to the files from library 10, and this new file was processed again using the same clustering thresholds as before. Once the files were processed using QIIME, they were imported to R using the phyloseq package. In parallel, a separate

phyloseq object was created using the files obtained from the synthetic sample generator script, which were not processed in QIIME, and served as a control as to what to expect to find in our samples.

As an example of the results obtained, three groups of samples, plus the control sample containing *N. musculi* reads, were selected. Results are shown in Figure 32, each block represent a sample processed with different clustering thresholds, plus the expected result which is shown in the leftmost bar for all blocks. The results show that the UCLUST approach does not accurately classify *rplF* amplicons. The relative abundance each species was identified in most groups of samples, but in some cases such as the *N. subflava* and *N. bergeri* mixture (samples 10 to 12), *N. subflava* was missed completely. For the other samples, the species identified varied depending on the threshold used. In cases such as the *N. mucosa* and *N. oralis* samples (Samples 58 to 60), the correct result was achieved at the 97% mark, but the *N. mucosa* reads were no longer identifiable in subsequent analyses with higher thresholds. Overall, there was no clear threshold that would allow complete separation of these samples, and I concluded that the UCLUST approach could not be applied to *rplF* amplicons.

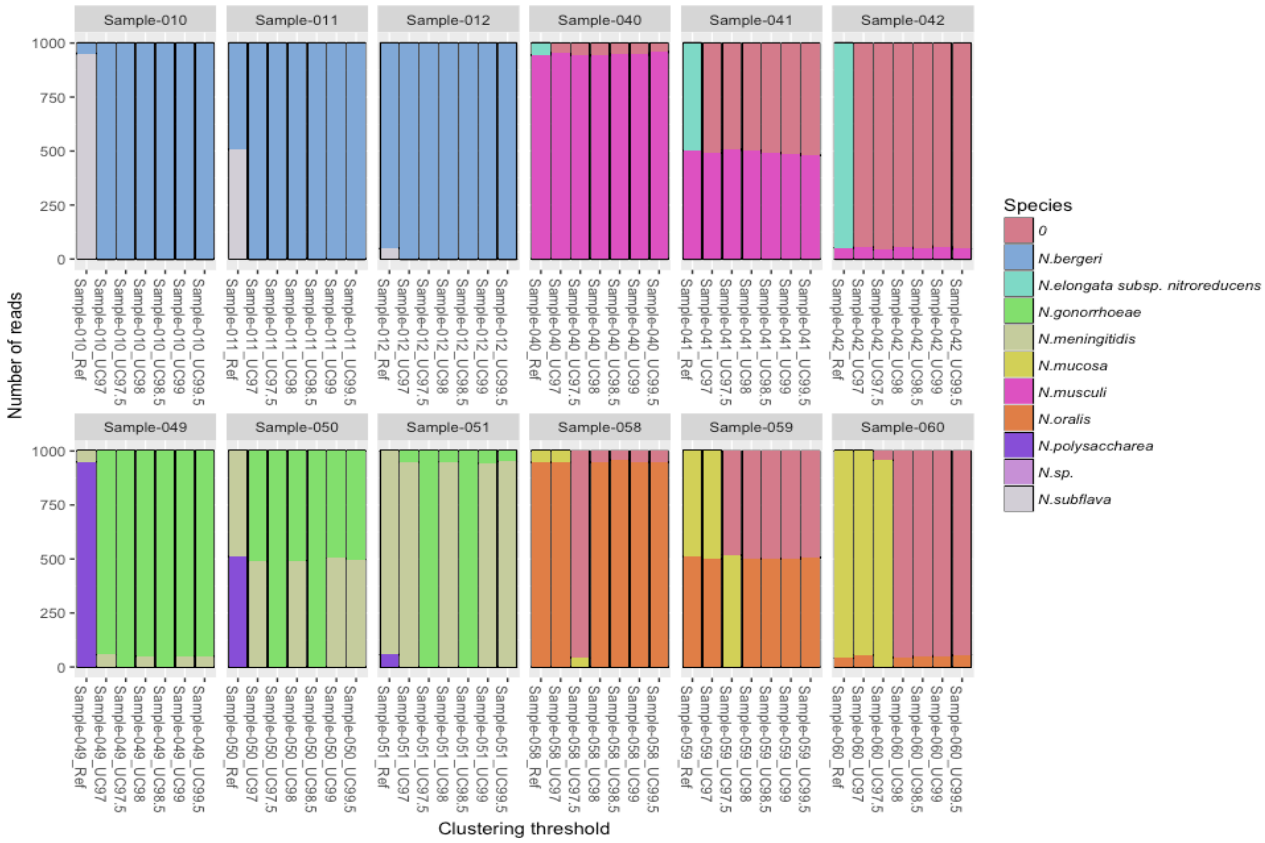


Figure 32 relative proportion of species identified in the synthetic samples created and processed with UCLUST. The X axis indicates the results for each of the UCLUST clustering thresholds used, and the Y axis the number of reads assigned to the different species, which are coloured according tto the legend on the left. Reads without a definite species assignment are classified as 0.

5.4.1.2. Different approaches tested for rplF sequence identification

I decided to test other methods for *rplF* amplicon analysis. Two possible options were identified, the minimum entropy decomposition (MED) method (99), or the dada2 package (145). These two methods offer a higher OTU resolution than UCLUST. Due to the increase in resolution provided by these packages, they do not refer to their output sequences as OTUs, but as sequence variants (SVs). These two methods rely on different approaches to identify SVs. Briefly, MED analyses the most variable bases in an alignment, and clusters the reads based on base composition of the most variable bases instead of the whole amplicon sequence. In contrast, dada2 includes a denoising algorithm that

corrects errors in the reads, and then partitions groups of unique reads based on different parameters the algorithm uses. Although both methods could potentially be applied to *rplF* sequence analysis and differentiate species, I decided to first test MED, as the clustering algorithm implemented in this package is more easily interpretable.

5.4.1.2.1. Use of the MED algorithm to cluster rplF sequences

The MED algorithm stems from the oligotyping method, originally published by the same author (99,146). The core method of the oligotyping algorithm relies on calculating the Shannon entropy value of each base in an alignment. This value describes the average entropy in a given data value, which in terms of DNA sequences, represents the average variability observed in a base in an alignment. An example is shown in Figure 33. The *rplF* amplicon contains certain bases that have a greater variation in their composition than others (the most variable bases are shown in Table 16), some, such as 302 having no variation at all. The oligotyping method divides the sequences based on a list of bases given by the user that should have the highest entropy values.

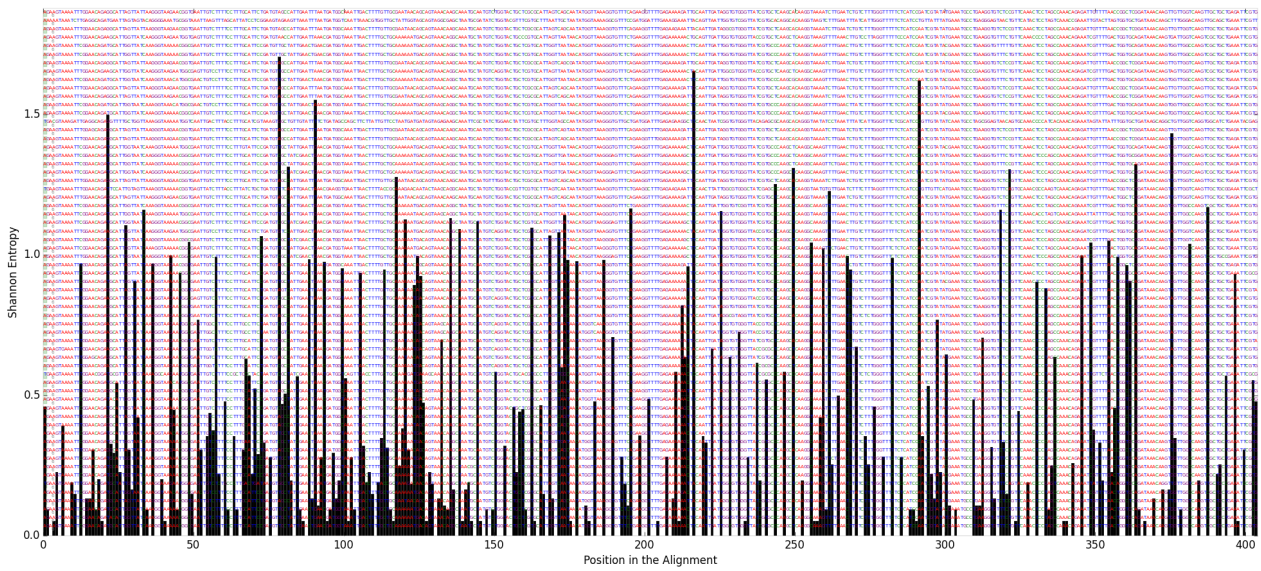


Figure 33 Shannon entropy values calculated for each base of the *rplF* amplicon using *rplF* reference sequences. Each base of the alignment on the X axis, and the entropy of each base on the Y axis.

Table 16 Most variable bases in the *rplF* amplicon.

Base	Entropy	Base	Entropy	Base	Entropy
78	1.7077	**90**	1.5525	**363**	1.3236
216	1.6551	**21**	1.5011	**81**	1.3144
291	1.6213	**375**	1.4344	**249**	1.3098

The oligotyping method is a supervised algorithm that relies on partitioning reads based on the nucleotide signature found in the sequences at the most entropic bases. The algorithm considers all possible nucleotide combinations present in the sequences being queried and splits them into groups that have identical nucleotide signatures without taking into account the reminder bases in the alignment. The MED algorithm requires several optimization steps, as one needs to test the number of bases being taking into account to cluster the data set if there is no previous knowledge of particularly variable bases. MED expands the oligotyping algorithm and applies it iteratively, querying all the variable bases.

The algorithm requires several inputs, the two most relevant being the number of bases being considered at a time, and the minimum entropy for a base to be considered variable. The algorithm identifies a set number of variable bases, checks that these have an entropy value over the threshold supplied, and splits the sequences based on the nucleotide signatures generated by these bases in different clusters (denominated nodes by the algorithm). On the second cycle, the algorithm repeats the process with the next set of bases, dividing the nodes previously generated into new nodes, and

the process is repeated until there are no bases left above the entropy threshold. The final step of this algorithm consists of checking the size of the nodes generated, and prunes those with a number of reads below an expected amount (supplied by the user, although by default it is the total number of reads supplied divided by 5000, which is 1802 for this library).

This MED method does not include a taxonomy assignment step. To achieve this, I used BLAST to identify the sequence that is the most similar to the *rplF* reference sequences. Using the BLAST output sequences, and the matrix of reads counts per node, I imported these files into phyloseq and were processed in the same way as the previous files. Similar to the different thresholds used when testing the UCLUST results, I used different parameters to test the effectiveness of the method. Initially, I used different cluster size values acceptable for the final pruning. Values from 0 (which defaults to the number of reads divided by 5000), to 800 in steps of 100 were used. The results are shown in Figure 34 with the same samples as those shown for the QIIME analysis (Figure 32). The MED algorithm provided excellent results, all the different variations tested providing almost identical results to the expected values.

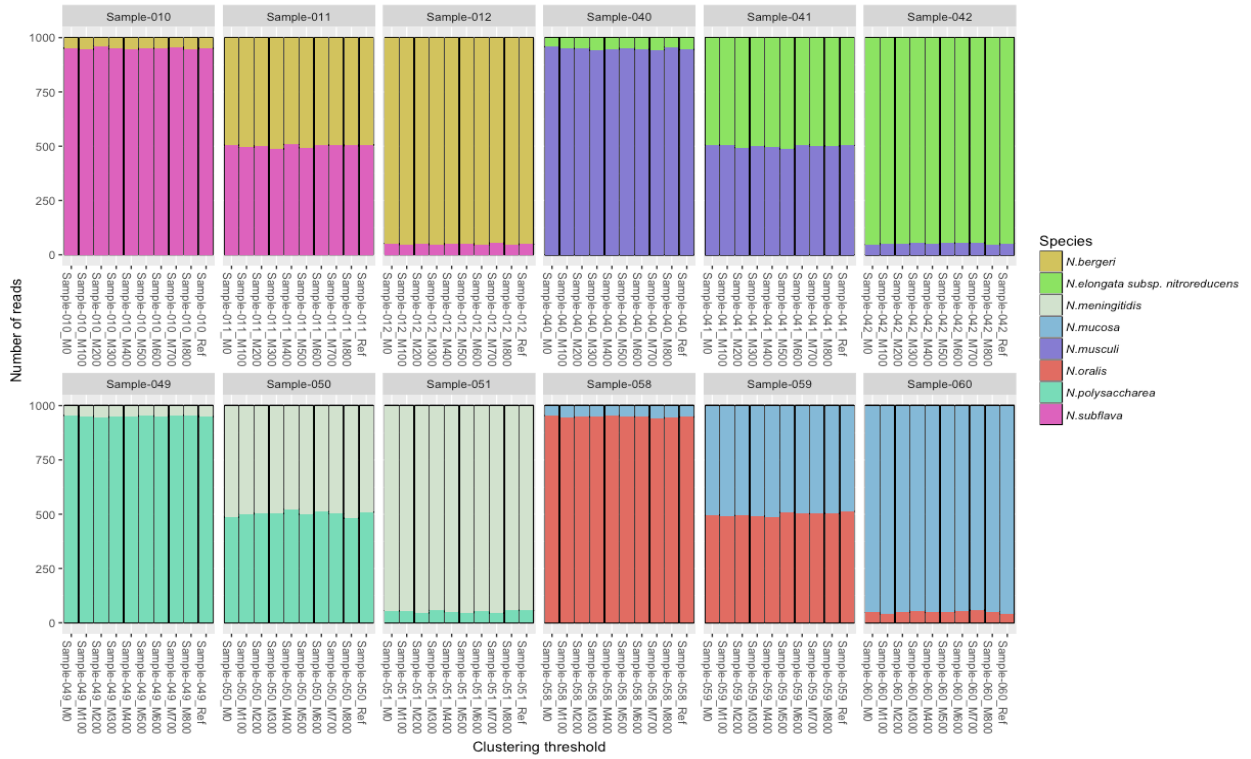


Figure 34 Relative proportion of species identified in the synthetic samples created and processed with MED for 3 sets of randomly selected *rplF* sequence combinations. The X axis indicates the results for each of the different cluster size filtering steps tested, and the Y axis the number of reads assigned to the different species, which are coloured according tto the legend on the left.

Although the first test of the MED algorithm was very promising, I wanted to test more parameters to identify what the best approach towards the analysis of *rplF* amplicons would be. Knowing that MED would give a more reliable answer than the previous methods used, I decided to test it using data from human samples and determine how the results may change depending on the parameters selected. Different combinations of the following parameters were assessed: 'm' which accounts for the minimum entropy required for a base to be considered for partitioning, 'd' which is the number of bases taken into consideration to assess the nucleotide signature, and 'M' the minimum number of reads required in a cluster. Using different combinations of these three parameters, I analysed the sequences from library 10 using the parameter combinations shown in Table 17

Table 17 Parameter combinations used to test the MED algorithm

m 0.10 | d 1 | M0	m 0.10 | d 1 | M500	m 0.20 | d 1 | M0	m 0.20 | d 1 | M500
m 0.10 | d 2 | M0	m 0.10 | d 2 | M500	m 0.20 | d 2 | M0	m 0.20 | d 2 | M500
m 0.10 | d 3 | M0	m 0.10 | d 3 | M500	m 0.20 | d 3 | M0	m 0.20 | d 3 | M500
m 0.10 | d 4 | M0	m 0.10 | d 4 | M500	m 0.20 | d 4 | M0	m 0.20 | d 4 | M500
m 0.10 | d 5 | M0	m 0.10 | d 5 | M500	m 0.20 | d 5 | M0	m 0.20 | d 5 | M500
m 0.10 | d 6 | M0	m 0.10 | d 6 | M500	m 0.20 | d 6 | M0	m 0.20 | d 6 | M500
m 0.10 | d 7 | M0	m 0.10 | d 7 | M500	m 0.20 | d 7 | M0	m 0.20 | d 7 | M500
m 0.10 | d 8 | M0	m 0.10 | d 8 | M500	m 0.20 | d 8 | M0	m 0.20 | d 8 | M500

After processing the 32 versions of the library tested, these were imported into phyloseq as before. Furthermore, I decided to implement a filtering step to these files when being imported, including only nodes in which the most abundant reads had a similarity to a *rplF* reference sequence above 98%. An example of the results obtained is shown in Figure 35. To facilitate the visualization of these results, only those from three random samples are shown. Each sample is shown twice, one processed without being filtered (labelled MED), and one including the filtering step (labelled MEDf). Each bar shows the results of each of the variable combinations mentioned in Table 17, but the labels are not shown to facilitate visualization. There was no observable difference among the results obtained when processing samples under different conditions.

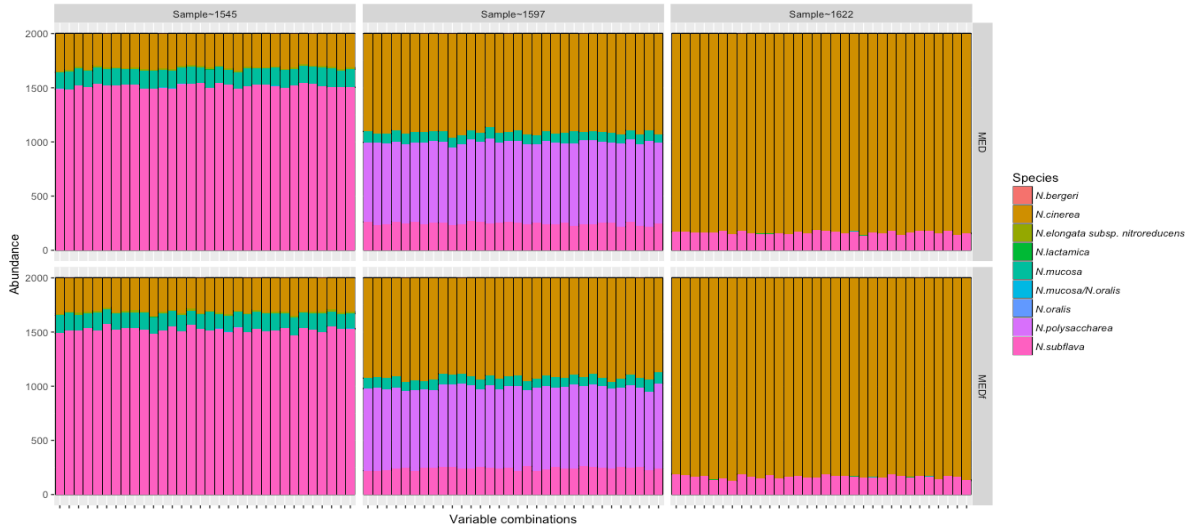


Figure 35 Relative abundances of species identified when testing different variable combinations in the MED algorithm. The X axis indicates the results for each of the different cluster size filtering steps tested, and the Y axis the number of reads assigned to the different species, which are coloured according to the legend on the left. The names of the different variables on the X axis are omitted to favour visualization.

Consistent results were found with the MED algorithm despite the variables used. However, I expected to identify some degree of variability depending on the variables used. Viewing the results obtained, it is possible to believe that the parameters altered did not lead to a significant change in the most abundant nodes as these will not be filtered out as the entropy caused by the variability of the sequences in these nodes will most likely surpass the 0.1 or 0.2 thresholds used, and the total number of reads in a given node will likely surpass the maximum amount set, which will be 1802 (for the default value M=0).

In order to have a greater understanding of the differences that may arise from the analysis of data with different MED parameters, I analysed the individual node files obtained for each of the combinations tested. I generated a new script that analyses all the node files present in an out folder and provided a Table with summary statistics (Table 18).

The results shown on Table 18 again show consistent results, indicating robustness of this method to identify *Neisseria* species. Analysing these results some trends were identified. As expected, analyses with more stringent parameters, such as using the default M parameter (M=0), or increasing the m parameter to 0.2 (default is 0.1), decreased the number of nodes obtained, nodes had slightly higher quality scores, as measured by the ASC and APC means. There was also a lower percentage of nodes with identities below 95%, and all nodes had a BLAST match. This is also accompanied by a lower average read count. The d parameter seemed to have a smaller effect on the overall quality of nodes identified, but larger values, usually from 5 onwards, appeared to stabilise the results obtained.

The use of slightly more lenient parameters (lower M values) did not significantly diminish the quality of the reads obtained. However, the number of clusters was almost doubled compared to the more conservative analyses. This could be due to the inclusion of the clusters with sizes in between 500 and 1802 reads. There was also an increase in nodes with identities below 95% and those with no BLAST match (these results will be discussed in more detail later). However, if we were to proceed with the results from these analyses, these nodes could easily be filtered.

There are arguments for either choosing a more stringent or flexible analysis. On the one hand, applying more stringent parameters will help ensure that the species identified for each sequence is not erroneous, and low-quality reads are properly excluded of the final analysis, reducing the noise in the data. Additionally, stringent analysis, may lead to loss of low prevalence nodes containing more rare sequences. It is also important to highlight that if one were to use the default parameters in the whole dataset including all the reads from the 10 libraries (a total of 66,135,760 reads), the minimum node size will be over 3 million reads, leading to almost a total lack of nodes

identified. Another parameter, usually considered for 16S rRNA survey analyses, is the presence of OTUs in more than one sample, which will be implement in downstream analyses to remove noise from the dataset.

As a final decision, I decided that I would use a less conservative approach for MED, using a d step of 5, m of 0.10 and M100. These values were considered too semi-conservative to be analysed as they were, so I decided to apply custom noise filtering steps instead of relying on arbitrary node sizes as the unique noise filtering step. This was implemented after the MED analysis.

Table 18 Summary statistics obtained after analysing 32 different MED outputs. The first three columns show the proportion of nodes in which the species identified for all the reads within them matched that of the most abundant read in each node at the 98, 99 and 100% level (98%, 99% and 100% in the Table). This script also provides the final number of nodes identified (Nodes), the average ratio of nodes with the same species as the most abundant SV in a node (average species count, ASC), the average ratio of nodes with the same rplF ID as the main sequence (average profile count, APC), the percentage of nodes that had a BLAST result with an identity to rplF sequences below 95% (Identity<95%), the percentage of nodes that did not match anything in the rplF database (lost), and the final number of reads obtained (total reads).

PARAMETERS	98%	99%	100%	NODES	ASC MEAN	ASC SD	APC MEAN	IDENTITY<95%	LOST	READS
m0.10-M0-d1	0.955	0.942	0.799	154	0.995	0.025	0.994	1.299	0.000	9,087,906
m0.10-M0-d2	0.955	0.942	0.812	154	0.995	0.021	0.995	1.299	0.000	8,997,748
m0.10-M0-d3	0.974	0.948	0.825	154	0.998	0.013	0.997	1.299	0.000	8,936,761
m0.10-M0-d4	0.968	0.961	0.838	154	0.997	0.015	0.997	1.299	0.000	8,886,790
m0.10-M0-d5	0.961	0.955	0.825	154	0.997	0.014	0.997	1.299	0.000	8,879,539
m0.10-M0-d6	0.974	0.968	0.883	154	0.997	0.016	0.997	1.299	0.000	8,837,379
m0.10-M0-d7	0.974	0.974	0.883	154	0.998	0.013	0.998	1.299	0.000	8,815,683
m0.10-M0-d8	0.955	0.955	0.864	154	0.996	0.025	0.995	1.299	0.000	8,797,192
m0.10-M500-d1	0.892	0.849	0.706	344	0.989	0.044	0.968	1.453	2.035	9,208,205
m0.10-M500-d2	0.897	0.874	0.721	348	0.989	0.049	0.969	1.437	2.011	9,151,931
m0.10-M500-d3	0.923	0.885	0.742	349	0.995	0.019	0.975	1.433	2.006	9,091,173
m0.10-M500-d4	0.915	0.898	0.764	352	0.993	0.036	0.973	1.420	1.989	9,071,430
m0.10-M500-d5	0.924	0.908	0.754	357	0.996	0.024	0.976	1.401	1.961	9,053,283
m0.10-M500-d6	0.925	0.911	0.783	360	0.994	0.034	0.974	1.389	1.944	9,041,392
m0.10-M500-d7	0.938	0.910	0.782	357	0.996	0.017	0.976	1.401	1.961	9,015,418
m0.10-M500-d8	0.931	0.909	0.790	362	0.995	0.022	0.975	1.381	1.934	9,003,714
m0.20-M0-d1	0.953	0.940	0.792	149	0.994	0.025	0.994	1.342	0.000	9,091,096
m0.20-M0-d2	0.953	0.940	0.805	149	0.995	0.022	0.995	1.342	0.000	9,010,977
m0.20-M0-d3	0.973	0.946	0.819	149	0.998	0.013	0.997	1.342	0.000	8,968,803
m0.20-M0-d4	0.966	0.960	0.832	149	0.997	0.015	0.997	1.342	0.000	8,929,220
m0.20-M0-d5	0.960	0.953	0.820	150	0.997	0.014	0.997	1.333	0.000	8,903,203
m0.20-M0-d6	0.973	0.966	0.879	149	0.997	0.017	0.997	1.342	0.000	8,869,638
m0.20-M0-d7	0.973	0.973	0.879	149	0.998	0.013	0.998	1.342	0.000	8,845,234
m0.20-M0-d8	0.954	0.954	0.854	151	0.996	0.025	0.995	1.325	0.000	8,834,543
m0.20-M500-d1	0.879	0.830	0.678	289	0.988	0.045	0.963	1.730	2.422	9,222,496
m0.20-M500-d2	0.878	0.851	0.695	295	0.987	0.053	0.963	1.695	2.373	9,171,392
m0.20-M500-d3	0.909	0.866	0.718	298	0.994	0.021	0.970	1.678	2.349	9,118,639
m0.20-M500-d4	0.900	0.880	0.740	300	0.992	0.039	0.969	1.667	2.333	9,098,394
m0.20-M500-d5	0.912	0.893	0.724	308	0.995	0.027	0.972	1.623	2.273	9,086,841
m0.20-M500-d6	0.914	0.898	0.754	313	0.993	0.037	0.970	1.597	2.236	9,070,899
m0.20-M500-d7	0.930	0.898	0.761	314	0.996	0.018	0.973	1.592	2.229	9,045,576
m0.20-M500-d8	0.922	0.898	0.773	322	0.994	0.023	0.972	1.553	2.174	9,035,177

5.4.1.3. Data processing in the R environment: Custom noise-filtering steps

The libraries were split in two different groups, one containing the processed files of two libraries prepared using artificial mixtures, and another including the processed files of the 10 libraries prepared using human samples. These two groups were processed identically. After processing the reads, the species assignment was performed using BLAST, as described before. A phylogenetic tree with a midpoint root was generated using FastTree, after aligning representative reads from each node using MAFFT. As a summary, the BLAST output files from each node, the matrix including counts of sequences per sample and node, the phylogenetic tree and a mapping file containing metadata of each sample, were used to generate phyloseq-compatible files, so that I could analyse the sample content. After processing the combined files with MED, I obtained an alignment of reads that showed a clear difference between highly entropic bases and those that are not (Figure 36).

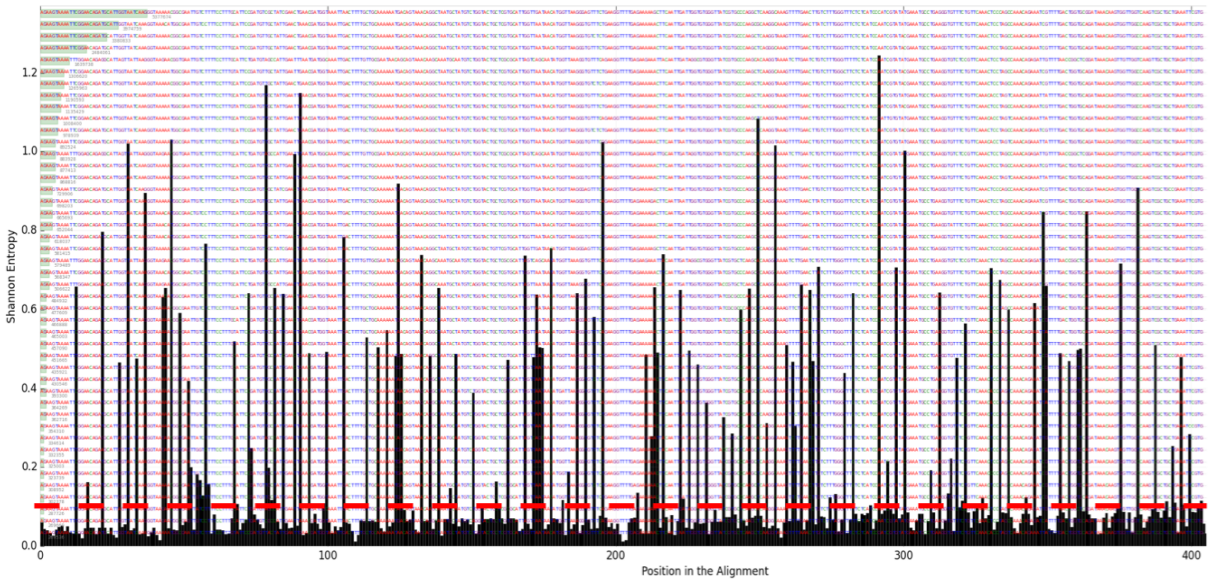


Figure 36 Entropy per base calculated for the analysis of 10 different MiSeq libraries. The red line shows the entropy threshold used to select bases considered to differentiate nodes), which is an indication the appropriateness of MED analyses for the rplF amplicon.

When importing files, the first step taken to filter noise out of the dataset was to exclude nodes with a BLAST identity below 97%, which resulted in a drop from 146 to 77 for the artificial mixtures, and 5782 to 5400 for the other libraries. A summary of the node identity values is shown in Figure 37 There were three groups of discarded nodes, those with an identity in between 96.79 and 92.57 (n=222), another group with values in between 82.87 and 78.66 (n=60), and those that did not match the database (n=100). A closer inspection of the discarded nodes revealed that the first group was formed by nodes that matched *Neisseria rplF* sequences, but these were either formed by more than one species or had a low total abundance. The second group matched sequence of the *rplF* gene of *Simonsiella muelleri*, and the third one was formed by sequences matching different human and bacterial sequences. A summary showing five random nodes from each group, is shown in Table 19.

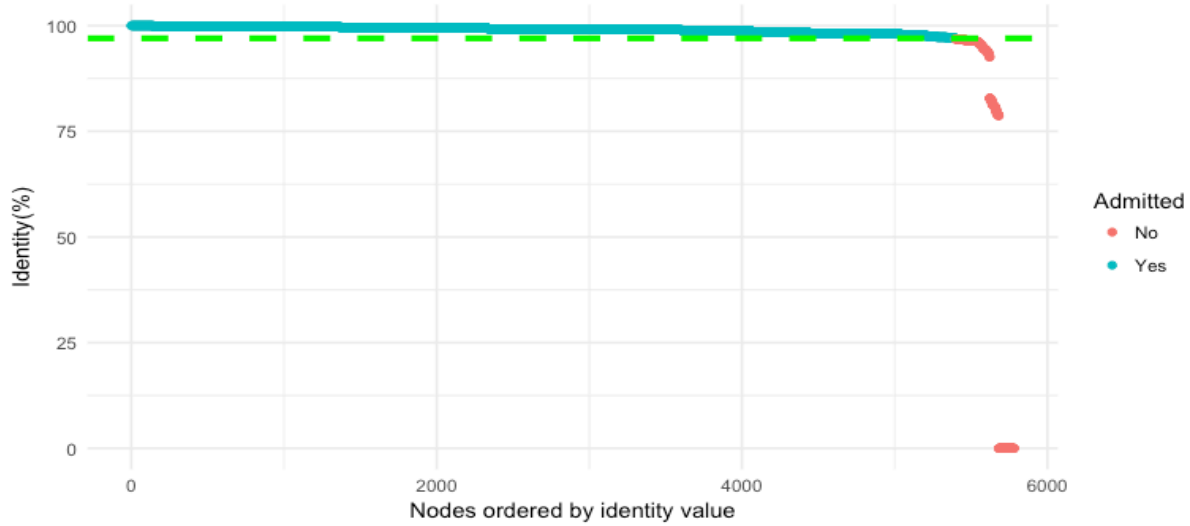


Figure 37 Identity values per node found in the libraries prepared with human samples. Nodes, shown in the X axis, are ordered by the identity percentage, displayed on the Y axis, of the most prominent read within each node. The nodes accepted appear in blue, the rejected ones in read, and the green line marks the 97% threshold used.

Table 19 Summary statistics of a subset of discarded nodes for 15 selected nodes. For each node, the total number of reads, the percentage of identity, the *rplF* prifiles found within that node, and the best *rplF* and BLAST match are displayed.

NODE	TOTAL READS	IDENTITY (%)	RPLF PROFILES	RPLF SPECIES	SPECIES OF BEST BLAST MATCH
NODE18002	191	97.79	*52*	*N. lactamica*	
NODE24847	210	97.78	80, 181	*N. lactamica, N. subflava*	
NODE41816	1186	96.53	9	*N. polysaccharea*	
NODE52552	217	96.29	37	*N. elongata*	
NODE27313	247	94.32	44, 160	*N. polysaccharea*	
NODE14356	1983	82.87	36	*N. oralis*	*Simonsiella muelleri*
NODE03827	327	82.57	74	*N. cinerea*	*Simonsiella muelleri*
NODE14411	616	82.19	58, 113	*N. lactamica, N. subflava*	*Simonsiella muelleri*
NODE34190	363	80.94	74,36,97	*N. cinerea, N. oralis, N. subflava*	*Simonsiella muelleri*
NODE00098	373	78.66	36	*N. oralis*	*Simonsiella muelleri*
NODE00019	410				*Homo sapiens*
NODE00244	471				*Rothia mucilaginosa*
NODE04027	238				*Homo sapiens*
NODE17253	231				*Homo sapiens*
NODE36249	183				*Homo sapiens*

Application of the filtering steps reduced the number of reads from 64,349,620 in 6,026 nodes (number of reads in the file when first imported) to 51,290,674 in 5644 nodes. I wanted to apply an extra filtering step to make sure I removed some of the noise in the dataset. One of the most commonly used steps when processing 16S rRNA reads involves removing OTUs that have a single count, also known as singletons. Applying this filtering step to my dataset revealed that 7.73% of the nodes found were singletons and removing these leads to a loss of 3,114,549 reads. At this point, as I did not have a reference as to how many nodes I could expect to find, I decided to investigate their composition more in detail.

The high number of nodes found creates an overly sparse count matrix, as most of the values present are 0. This can increase computation times when calculating different distance metrics among samples and introduce difficulty into the analysis of samples (147). To try to tackle this potential issue, I decided to investigate how much each node contributed to the total amount of reads. The results showed that there was a large portion of nodes that did not contribute greatly towards the total read count. A summary of this analysis is shown in Figure 38. The majority of the nodes, 4935 out of 5644, contribute between 0.00000194 and 0.0025% of the total reads (1282.82 reads). I decided to subtract these nodes from the file and test in parallel to the unfiltered version of the file to ensure that this node filtering step does not result in a loss of samples identified (due to the 1,656,900 reads that would be removed).

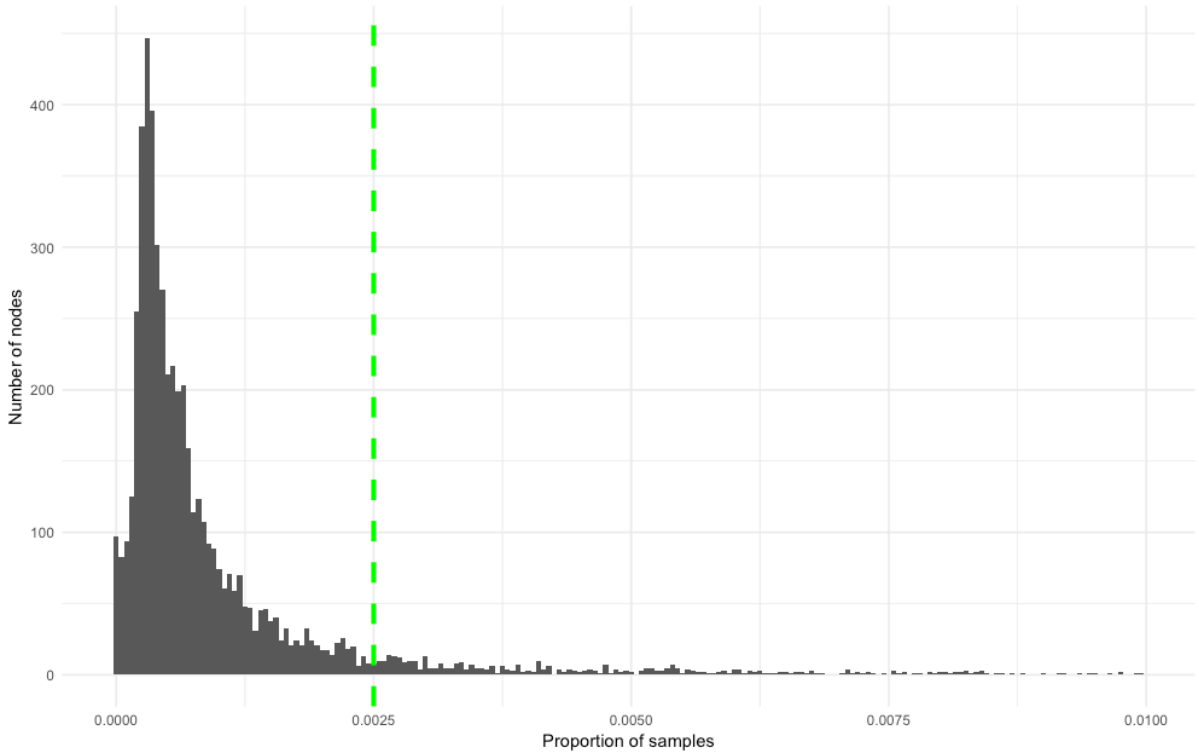


Figure 38 Node contribution towards total amount of reads. the Y axis indicates the number of nodes identified that contributed towards a percentage of reads (X axis). The green horizontal line indicates the threshold selected for node filtering

The final step taken for the data preparation steps of the analysis was to rarefy the count matrix. This step and the subsequent analyses of the data will be described in Chapter 6 (Section 6.4.1).

5.4.2. Data pre-processing for 454 sequencing reads

To process the files obtained after the 16S rRNA sequencing experiments performed on the 454 GS-FLX titanium, I used scripts included in QIIME. However, some difficulties appeared during the pre-processing steps taken. To explain the issues found in sufficient detail, I will explain, very briefly, the basics of 454 sequencing as applied in this project.

Similarly to the Illumina PCR steps described in earlier Chapters, a PCR step is performed to amplify the gene of interest, in this case the 16S rRNA gene, in particular the variable region 3 to 5 fragment. This process amplifies the gene while attaching an index sequence and adaptors necessary to make a 454-complatible library, in a process similar to the 1-step PCR described in Chapter 4 (Section 4.5.1). After the PCR, amplicons were gel-purified using 2% agarose E-Gels, quantified using the Quant-IT PicoGreen kit, and an equimolar pool was created by taking variable volumes of each sample so that 2 million amplicon copies per sample were mixed in the final library. The library was then quantified and mixed with resin beads that have an oligonucleotide sequence that binds the adapter sequence in the amplicons. This mixture was formulated at a ratio such that only 1 DNA molecule was bound to one bead. The DNA molecules attached to the beads are then amplified in an emulsion PCR that separates each bead from the others trapping each in a micelle, and the DNA sequence is amplified, which leads to beads covered with copies of the same DNA molecule. The beads are then placed in a pico-titre plate, which has wells that only allow 1 DNA-bound bead per well. In

addition, smaller beads with enzymes necessary for the sequencing process, such as DNA polymerase, ATP sulfurylase, luciferase and apyrase, are also added to the plate.

Once the sequencing steps start, solutions including only one nucleotide at a time are passed through the plate. If a nucleotide binds to the next free base in the DNA, it is added by the polymerase, releasing a pyrophosphate, which is converted to ATP by the ATP sulfurylase, which then acts as a substrate to the luciferase. The amount of ATP produced is proportional to the number of nucleotides bound, so the machine can differentiate a single nucleotide from a homopolymer based on the amount of light generated. The remainder ATP in the wells and the remaining nucleotides are then degraded by the apyrase, and another nucleotide is passed through the cell. There are two possible sequences in which nucleotides can be flown through the cell known as flow patterns A and B. Flow pattern A is a cyclic pattern in which the sequence TACG is repeated for 400 cycles to sequence the amplicons. Flow pattern B relies in this TACG pattern for four times, with the rest of the nucleotides added being added at random. Pattern B is said to increase read length and quality.

The samples sequenced as part of the Dormice study were sequenced in a total of eight sequencing experiments, three of these containing samples from the Oxford cohort. As the sequencing experiments were performed over the span of several years, some of the runs were sequenced using the flow pattern A, and two were processed with B. There was a setback during the data processing as the denoise_wrapper.py script in QIIME does not allow flow pattern B flowgrams. Thus, part of the data processing was done with FlowClus (95), as this application allows processing of output files generated based on either flow pattern. Due to the high error rate on 454 sequencers (148), steps are taken to remove spurious reads, *i.e.* to 'denoise' the data. Reads are clustered according to the intensities measured during the different sequencing cycles, and those that do not

follow the intensity patterns of the majority are excluded. The latter process is necessary to deal with one of the main problems found during 454 sequencing, *i.e.* the light intensity generated by homopolymer lengths does not always correlate with its length, and the signal identified by the sequencer could be erroneous, leading to insertion or deletion generated by misidentification of homopolymer tracts (149).

In summary the steps taken to process 454 amplicons was as follows:

- Step 1: Use FlowClus to denoise the files from each sequencing experiment individually, while filtering reads shorter than 200 bp or longer than 600 bp, sequences with ambiguities.
- Step 2: Identify chimeric sequences using the identify_chimeric_seqs.py, that relies on USEARCH (96). These are reads that align to two different reference sequences and are generated during the PCR amplification steps.
- Step 3: Filter chimeric reads from each denoised file using the filter_fasta.py script.
- Step 4: Combine all filtered files.
- Step 5: Proceed to select OTUs using the pick_open_reference_otus.py script, using sequences from the SILVA database as reference, keeping the standard 97% similarity to identify OTUs.
- Step 6: Align the reference sequences obtained for each OTU using MAFFT, and generate a midpoint rooted phylogeny.

5.5. Discussion

QIIME is one of the most used environments for analysis of 16S rRNA reads. There are multiple guidelines, some published by the QIIME development team, that explain how to approach the analysis of different sequence formats. Applying the scripts to the files obtained after sequencing 16S rRNA amplicons on the 454 GS-FLX Titanium platform, the only issue identified was the lack of support of flow pattern B sequenced reads. However, the use of other applications, like FlowClus (95), helped bypass this problem. Using both tools, I was able to process all the files and obtain OTU tables and phylogenetic trees which could be imported into the R environment as a phyloseq object.

An important point to consider was the type of tree generated based on the OTU reference sequences. The most widely used metrics used when comparing OTUs between pairs of samples are UniFrac distances, which summarize the phylogenetic distance among samples using by calculating the phylogenetic tree branches that are unique to a sample; 0 being two identical samples and 1 two completely different ones. There are two versions of the UniFrac metric; (1) the unweighted UniFrac (also referred to as just UniFrac), which is a qualitative version as it only considers presence or absence of OTUs; (2) the weighted UniFrac metric which considers the abundance of each OTU while calculating distance values. If one creates an unrooted tree, most algorithms take a random leaf of the tree to consider it as the root, and calculates all distances based on the newly rooted tree. Although this is not erroneous, each time the distances are calculated a different leaf will be used, providing different distance values for individual sample comparisons. If one generates a mid-point rooted tree, although arbitrarily rooted, it provides stable outputs if the distance is calculated more than once. It is also of relevance to note that the method used to generate the phylogenetic tree seems to not have a significant impact in UniFrac metric calculations (150).

It was possible to use QIIME scripts to process and filter reads obtained, as quality filtering and full amplicon assembly using both index and sequence reads was used. Once all the individual files were combined however, assigning a species to each of the reads proved a difficult task using the UCLUST approach. Developing a reliable pipeline that could provide reliable results to process *rplF* reads was a challenging endeavour. There are several different approaches and platforms available for the analysis of 16S rRNA reads that can be applied to any amplicon. However, it is necessary to use different approaches that can be tailored to the specific need of each amplicon sequence, and the source of variability that comes with it. For more varied amplicons such as 16S rRNA, tools like UCLUST perform as intended, and are widely used by the community. In contrast, amplicons where the sequence variability is more restricted, such as *rplF*, where the information comes from a few bases within the amplicon, may require different tools and/or approaches.

The oligotyping and MED algorithms provided robust and accurate results, being able to differentiate sequences that only differed in 1 base and doing so while analysing the reads using different parameters. The initial test performed with *rplF* reference sequences showed that there were bases that contributed towards the overall variation more than others, and this was especially apparent when analysing the clinical dataset. When trying different methods to process the data, it was difficult to identify a single parameter combination that was optimal.

I decided to use a slightly less-conservative method for the final dataset, and then proceed to do more filtering steps applicable to my dataset, rather than use default MED values. Most publications used MED rely on variations within 16S rRNA sequences to identify potential species (146,151-153), and it was difficult to translate the parameters used in these publications to my own, as the sources

of variation in the sequences were different. Thus, I decided to filter the nodes obtained in future steps to better understand outputs.

Based on the results obtained to that point, use of a low abundance filter (M=100) was sub-optimal, as the number of identified clusters was relatively high, and future analyses could be performed using higher values. However, the m (m=0.1) and d (d=5) parameters selected seemed to be sufficient. It was difficult to judge whether the values obtained were appropriate as there was no frame of reference as to how many different *rplF* sequence variants were expected, and it might be only possible to judge an estimated number after assessing the results of other studies using this approach. The first step taken, filtering nodes based on BLAST identity, showed that the vast majority of the nodes had a high similarity to the reference sequences. It was debatable where to set a threshold to include a node in the final matrix count table, but 97% seemed a reasonable level. Nonetheless, there is a clear separation between the sequences that were excluded but were above 90% identity. Although the sequences with a similarity below 97% may still belong to the *Neisseria* species assigned to them, it could also mean that these are really other *Neisseria* species, even novel *Neisseria* species, based on their difference in identity. It was, therefore, thought best to exclude them from the final dataset.

As mentioned in Chapter 3 (Section 3.4.5), the *rplF* and primer binding sequences of other *Neisseriaceae* are similar to those of *Neisseria* species. Here I identified several nodes containing sequences that belong to *Simonsiella muelleri,* and also as mentioned in Chapter 3, these sequences were sufficiently different from those of *Neisseria* to be able to differentiate them clearly. Finally, some human sequences were identified in our analysis, but these sequences were much more dissimilar from *Neisseria* sequences. It was never predicted that the *rplF* primer set could amplify

human sequences, however, it is possible that due to secondary structures in the primers, these could bind to human DNA in the samples. It is difficult to predict such binding, and it might not be possible to bypass this issue, nonetheless, the number of reads obtained that belong to non-bacterial DNA was low, and this could be disregarded.

The final steps taken to filter the nodes, filtering singletons, was a helpful step towards the ease of downstream analyses of downstream and resulted in removal of a small proportion of the nodes. At this point there were 5,492 nodes in our dataset, and I identified a great number of nodes in which the total read count of the nodes was less than 0.0025% (1282.82) of the total number of reads. Subtracting these nodes would be equivalent to applying the M filter in MED, however being able to visualise this provided a deeper understanding of how this parameter could be used in future analyses. The implications of using or not this filtering step will be discussed in following Chapters, as I will discuss the impact of this filtering in the sample recovery.

6. Chapter 6: Effects of Bexsero vaccination on the upper respiratory tract microbiome

6.1. Introduction

The development of the composition of the infant URTome has been studied in healthy individuals (107,154), and in certain diseases such as asthma (155,156), cystic fibrosis (157), and pneumonia (32). Several early life events such as mode of delivery (158), and initiation of breastfeeding (159), have been linked to changes in the microbiome. However, the effect of vaccination in the microbiome has not been studied in detail (160,161), and most of the focus has been on pneumococcal vaccines.

Although one could expect carbohydrate-based vaccination to have a low impact on microbiome composition as a whole, the use of protein-based vaccines, such as Bexsero®, could have a higher effect due to potential similarities in homolog proteins. In this Chapter I will describe the hypothetical effects associated with Bexsero® vaccination in infants, and compare a group of unvaccinated infants with two groups of Bexsero® vaccinated infants, one receiving doses at 2, 4 and 12 months and the second one at 6, 8 and 13 months. All samples will be compared at the 6 and 12-month time point.

6.2. Aim

- Identify differences in microbiome composition in infants associated with Bexsero® vaccination.

6.3. Summary of methods

6.3.1. DNA extraction method

Oropharyngeal throat swab samples were extracted according to manufacturer's especifications using the FastDNA SPIN kit for Soil (section 2.6). Purified DNA samples were stored at -80ºC until necessary.

6.3.2. Library preparation for 16S rRNA gene libraries

Copies of the 16S rRNA gene present in DNA extracted from oropharyngeal throat swab samples were amplified using Roche 454-compatible primers, described in (92). These primers amplify the V3 to V5 hypervariable regions of the 16S rRNA gene. After PCR amplification, amplicons were purified using AMPure XP beads, and quantified using Quant-iT PicoGreen, both as per manufacturer's specifications (sections 2.9.2. and 2.9.3). Finaly, amplicons were pooled to an equimolar concentration, and sent for sequencing.

6.3.3. *Data processing and statistical tests*

After the data were processed using QIIME (MacQIIME distribution V1.9.1, (93)), it was imported to the R environment (V3.4.4) and processed using functions in the phyloseq package (V1.22.3, (94)) and vegan (2.5-1, (162)). Statistical tests were also performed in this environment, using the function wilcox.test to calculate p-values. Differential abundance testing was calculated using the fitFeatureModel function within the metagenomeSeq package (v1.20.1, (163)). Linear discriminant analysis Effect Size (LEfSE) was calculated using the bioBakery (164) virtual machine provided by the developers (V1.7).

6.3.4. *Sample summary*

Samples used in this analysis included Dormice samples, collected in London, from infants at 6 weeks, 3, 4.5, 6, 9, 12, 18, 24 and 36-months of life. These samples were used as controls as none these infants were vaccinated with Bexsero®. The second set of samples were collected in Oxford. This set had two groups of samples, one with infants vaccinated at 2, 4, and 12 months (Group 1) and another group (Group 2) had infants vaccinated at 6, 8 and 13 months. All samples regardless of collection site had oropharyngeal throat swab samples collected at 6 and 12-months of life using the same swabs and collection protocols.

To establish the potential effect of Bexsero® on the upper respiratory tract microbiome, four variables were explored to compare the effect of vaccination. The first one, time point, codes the two time points for Dormice samples as 6 and 12 months (for the 6 and 12-month samples, respectively). The Oxford samples were coded as V06 and V07 for the earlier time points of Group 1 and 2 respectively, both of which were collected at the 6 months, followed by V12-G1 and V12-G2 for the

later time points, collected at 12 and 13 months, respectively. Secondly, sample origin separates Oxford and Dormice samples. Thirdly, vaccination status describes whether or not infants were vaccinated by the 6 and 12-month time points, regardless of the number of vaccine doses received. This variable is used in particular to identify changes due to vaccination at the 6-month time point, grouping Dormice and non-vaccinated Oxford infants. Finally, the sex of each infant is coded in the variable with that name. The number of samples per variable is shown in Table 20.

Table 20 Number of samples present per variable analysed. Samples are devided according to the different levels in each variable.

VARIABLE	EARLY TIME POINTS (6MO, V06, V07)	LATER TIME POINTS (12MO, V12-G1, V12-G2)
TIME POINT	6 mo (n=172)	12 mo (n=153)
	V06 (n=79)	V12-G1 (n=69)
	V07 (n=86)	V12-G2 (n=67)
SAMPLE ORIGIN	Dormice (n=172)	Dormice (n=153)
	Oxford (n=165)	Oxford (n=136)
VACCINATION STATUS	No (n=258)	No (n=153)
	Yes (n=79)	Yes (n=136)
SEX	Female (n=159)	Female (n=139)
	Male (n=178)	Male (n=150)

6.4. Results

6.4.1. Rarefraction of OTU table

Diversity comparisons in microbiome studies relies on quantifying the richness and evenness within individual samples (α-diversity), and the differences among samples (β-diversity). Different metrics can be used to calculate α and β diversity values. However, due to the varying number of reads samples can have, the diversity metrics calculated can show skewed values towards samples with a higher read count. In order to overcome this issue, most microbiome analyses have relied in a process denominated rarefaction, which consists of randomly selecting OTUs from a sample up to a given depth. To identify an adequate rarefraction value, I calculated the α-diversity values of each sample at different depth using four different metrics (Figure 39). As mentioned earlier, there are different metrics that can be used to assess diversity. One of them, counts the different number of OTUs observed for a given depth (Observed), the Shannon index (Shannon) considers the abundance and evenness of OTUs present, making rare OTUs more relevant, and the inverse Simpson index (InvSimpson) uses a weighted arithmetic mean, which puts more emphasis on the most abundant OTUs.

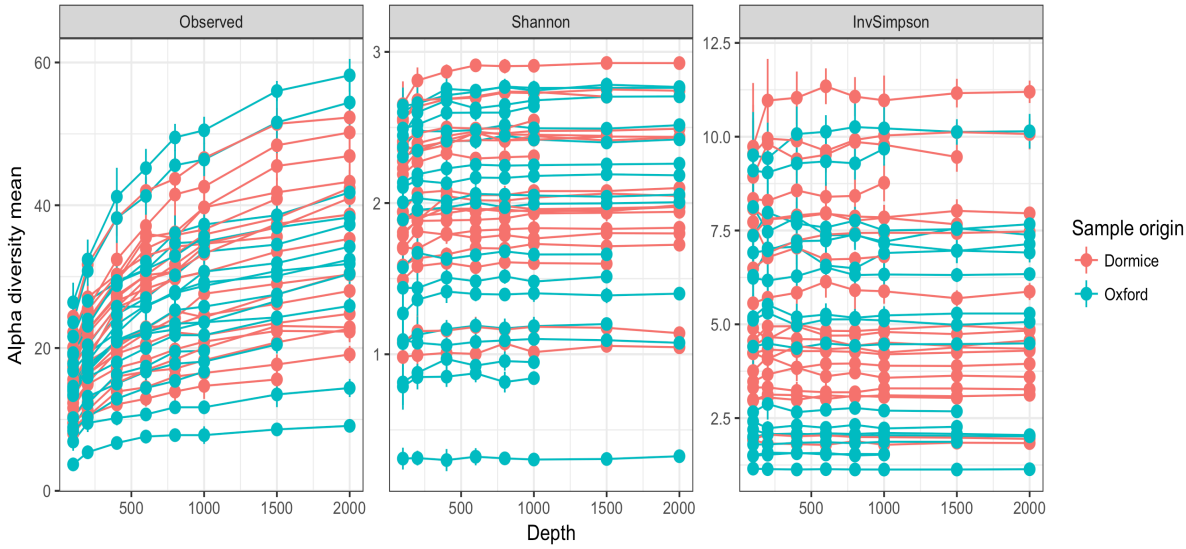


Figure 39 Alpha diversity mean values (Y axis) for Dormice (red) and Oxford (blue) samples at different sampling depths (X axis). The results presented here represent 100 randomly selected samples.

These metrics highlight important features in the data (Figure 39). The number of observed values increases with sequencing depth, which is expected, and does not converge by the 2000 read threshold. The other two metrics, although they focus on different aspects of diversity, seem to stabilize beyond the 500 read depth, suggesting that this could be a good threshold to capture the diversity in the samples. It is also important to consider the trade-off between the sequencing depth and samples that will not be included in the analysis, as the samples with a number of reads below a selected sequencing depth will be excluded from the analysis.

I calculated the number of samples that would be discarded when rarefying samples at a given depth. Figure 40 shows the number of samples lost (Y axis) at different sampling depths (X axis), when considering the Dormice and Oxford samples, plus the negative controls used during sequencing. The results pointed towards the range in between 600 and 750 to be the optimal sampling depth value, as it could maximize the number of reads obtained, minimize the number of samples lost, and still be above the 500-mark estimated from the α-rarefraction graphics.

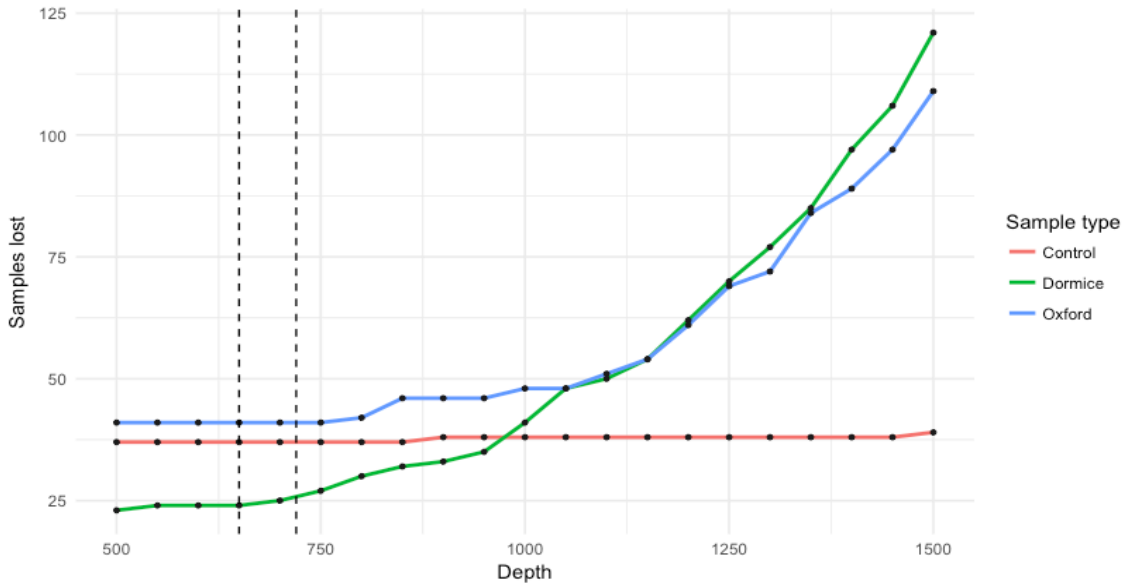


Figure 40 Number of samples lost (Y axis) at different sampling depth (X axis). The three different coloured lines indicate the numbers of samples lost for Oxford samples (blue), Dormice samples (green) or control samples (red). The two dashed vertical lines highlight the sampling depth selected for further inspection.

Looking closer into this range, I calculated the number of samples lost for depths in between 600 and 750 to see the changes in more detail. Figure 41 shows a more detailed version of Figure 40, showing the number of samples lost at sequencing depths in between 650 and 720 reads. This graph showed that, for this range, the Oxford and Control samples remained unchanged, with the loss of 42 and 37 samples, respectively. For Dormice, depths below 698 resulted in a loss of 24 samples, increasing up to 26 past the 701 depth. It was therefore considered that the optimal threshold to use was 698 reads, assuming the loss of 102 samples, out of 1862. All the negative control samples included during sequencing (unused swab heads) were rejected during this step.

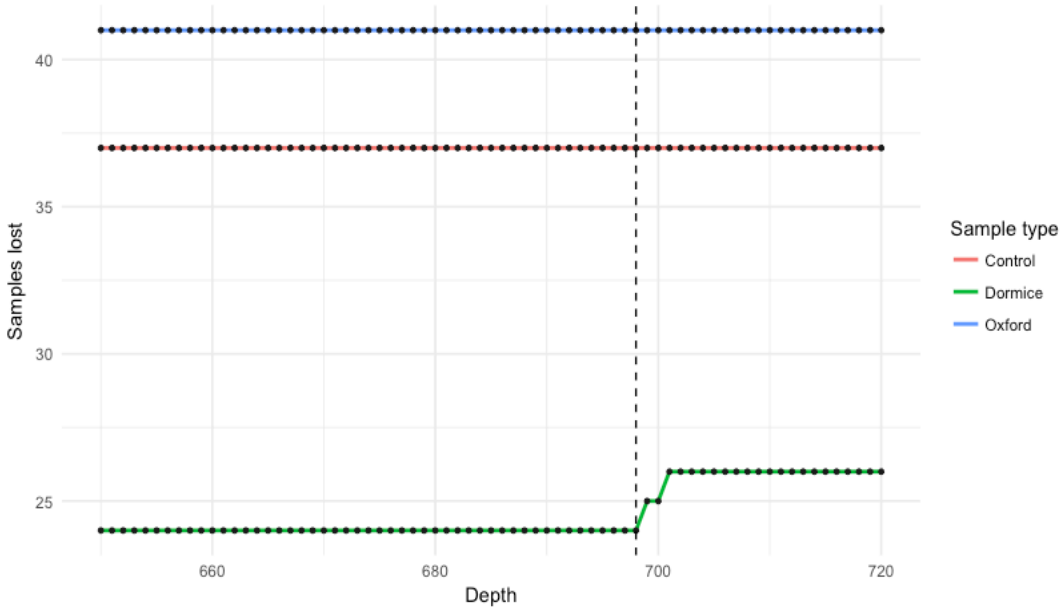


Figure 41 Number of samples lost at (Y axis) sampling depth (X axis) in between 650 and 720 reads. The three different coloured lines indicate the numbers of samples lost for Oxford samples (blue), Dormice samples (green) or control samples (red). The two dashed vertical lines highlight the sampling depth selected for further inspection.

6.4.2. Validation work: comparison of repeated samples

In order to assess potential inter-sequencing experiment bias, some samples were repeated in at least two different experiments to compare the reproducibility of the results. Figure 42 shows the results of 10 randomly selected samples repeated in different experiments. Each labelled column contains the repeats (2 to 3 per sample) for that sample, and each row shows the abundance, measured in number of reads, of the 10 most abundant genera. Overall, there was enough similarity among repeats to be confident that there was no experiment bias.

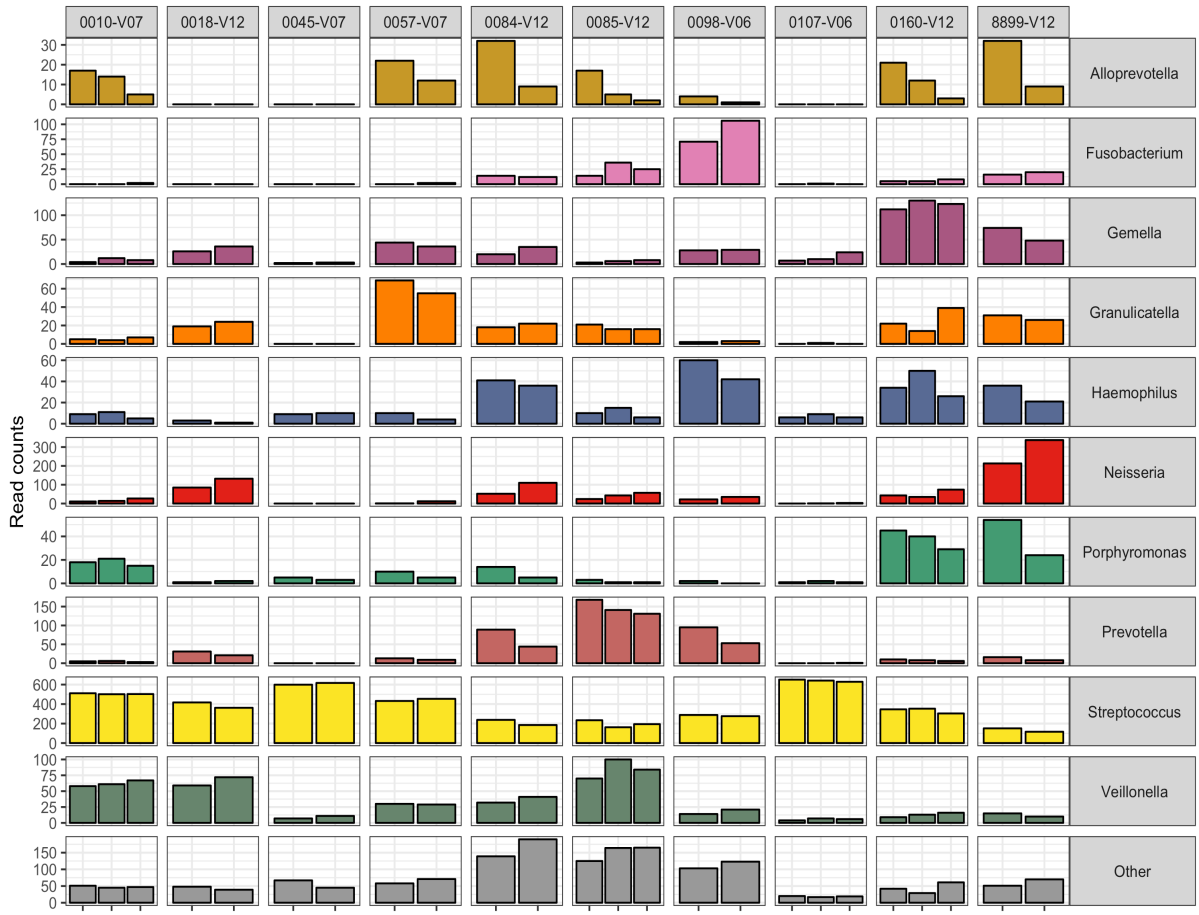


Figure 42 Read counts per genus and sample repeats identified for 10 repeated samples. Each column represents one sample, as per the label above, and each row represents the read counts per genus, showing one bar per repeat of a given sample.

6.4.3. Validation work: effect of different 454 GS FLX flow pattern in the data

As mentioned in Chapter 5 (Section 5.4.2), two libraries were sequenced using the flow pattern B in the GS FLX titanium instrument. Also, these two libraries contained only Oxford samples, so it is important to identify any potential source of variation that could arise from the change in flow pattern, as it might be a confounding factor in downstream analyses. An NMDS ordination using all the samples sequenced showed no identifiable differences that could arise from the differences in flow patter (Figure 43).

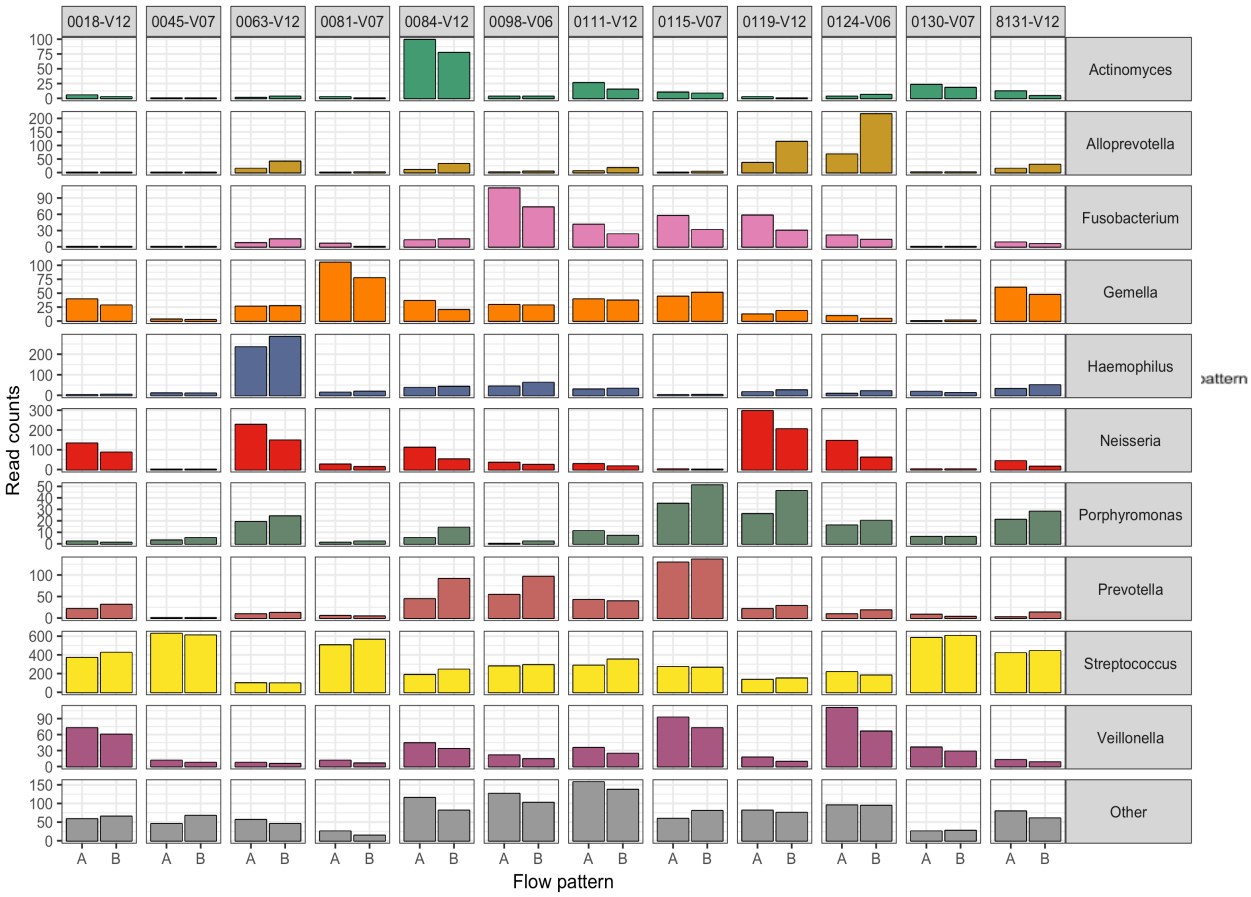


Figure 44 Abundance per genus in samples sequenced using flow pattern A and B. Each column represents one sample, as per the label above, and each row represents the read counts per genus. All sample was sequenced using flow patterns A and B.

However, to have a better representation of the potential effect of flow pattern, I selected all the samples with repeats that were sequenced using different flow patterns. Similar to the previous analysis focusing on repeated samples, no significant differences could be identified. Figure 44 shows the abundance per genera (showing the top 10 most abundant ones) per sample, sequenced using pattern A or B for 12 different samples. Finally, a paired Wilcoxon signed rank test showed no significant difference ($p = 0.9192$) when comparing genera abundances from all repeated samples (n=23, 54 repeats total). In addition, there were no significant differences identified when comparing individual genera.

6.4.4. Alpha diversity value comparisons

After rarefying samples to an equal sequencing depth, I wanted to compare the α-diversity values across time and identify any potential differences in the development of the Oxford samples compared to Dormice. Figure 46 shows the α-diversity values calculated by grouping all the samples available for each time point included in the Dormice cohort using different metrics. Each time point was compared with the prior and subsequent time point using the Wilcoxon signed rank test, correcting the values using the Holm-Bonferroni method. All three methods show a trend towards an increased α-diversity over time. The number of different OTUs observed increased significantly at each time point collected, while the other two metrics shows a significant increase in all time point comparisons except the 3, 4.5 month and 9 -12-month comparisons. Oxford samples showed a similar trend, being both of the 6-month time points (V06 and V07) significantly different from the 12-month samples (V12-G1, V12-G2). When comparing all early time points there were no significant differences. Later time points only showed a significant difference when comparing both Oxford groups using the Inverse Simpson metric (Figure 45).

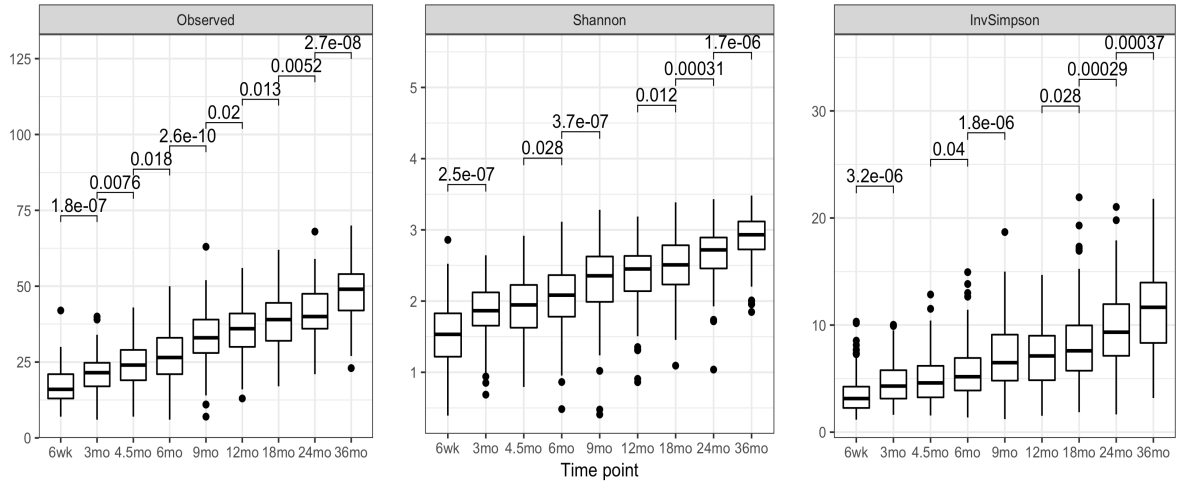


Figure 46 Boxplots describing the alpha diversity values calculated per time point of the Dormice study. Each time point is represented in the X axis, while the Y axis indicates the alpha diversity values calculated using the total number of OTUs (Observed), the Shannon index (Shannon) and the inverse Simpson metric (InvSimpson).

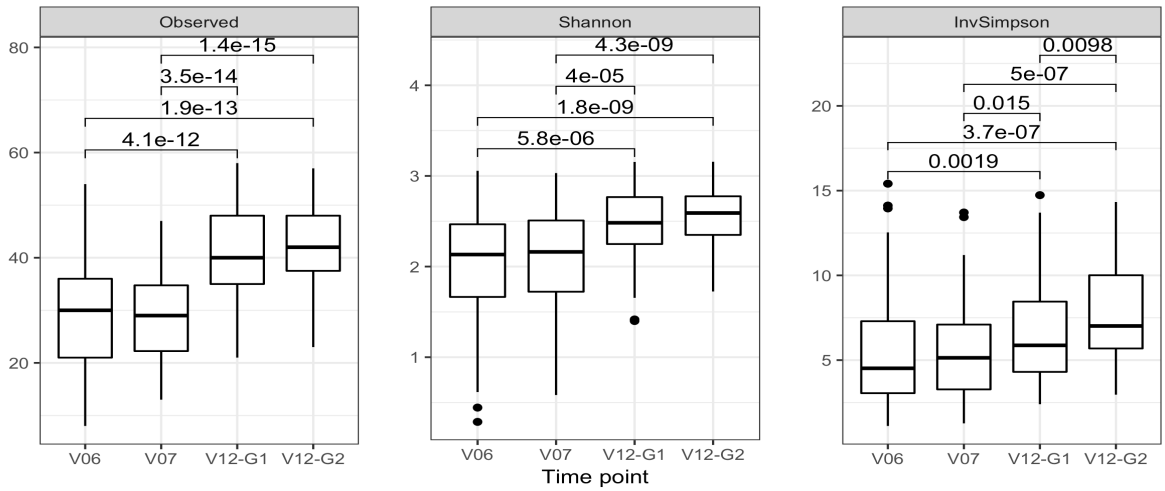


Figure 45 Alpha diversity values calculated per time sample collection group in the Oxford cohort. Each time point is represented in the X axis, while the Y axis indicates the alpha diversity values calculated using the total number of OTUs (Observed), the Shannon index (Shannon) and the inverse Simpson metric (InvSimpson).

In order to ensure more robust comparisons between Dormice and Oxford samples, I compared values across similar time points in both cohorts as follows: the 'earlier time points' 6 months for Dormice and the V06 and V07 for Oxford; and the 'later time points': 12 months for Dormice and V12-G1 and V12-G2 for Oxford, which can be seen in Figure 47. Earlier samples are shown in blue, while later samples appear in red. The earlier time points did not show a significant change in any of the metrics used. Later time points were more variable, showing a significant difference

between the Dormice and both Oxford sample groups when measuring the number of OTUs identified, and in between the Dormice and the V12-G2 sample group using the Shannon index. However, no differences were identified when using the Inverse Simpson metric. This suggest that the differences seen arose from differences in low abundance OTUs.

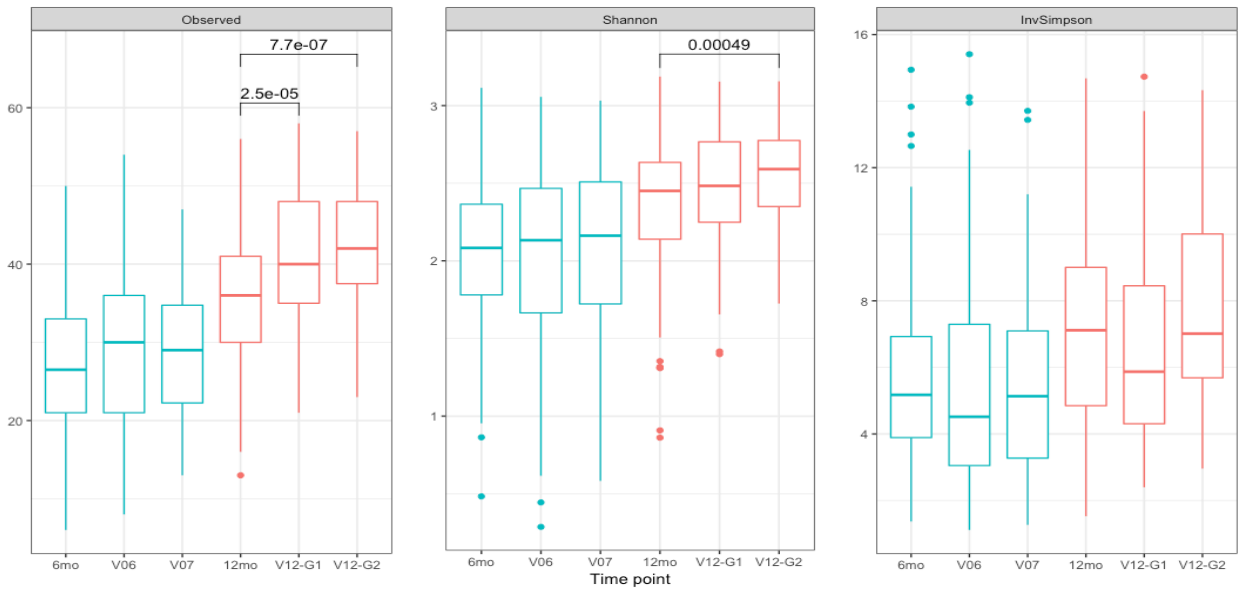


Figure 47 Alpha diversity values calculated for earlier time points (6mo for Dormice, V06 and V07 for Oxford. All shown in blue) and later time points (12mo for Dormice, V12-G1, V12-G2 for Oxford. All shown in red) in both cohorts. Each time point is represented in the X axis, while the Y axis indicates the alpha diversity values calculated using the total number of OTUs (Observed), the Shannon index (Shannon) and the inverse Simpson metric (InvSimpson).

6.4.5. Comparison of the microbiome of the Dormice and Oxford cohorts

To assess any possible disruptions caused by Bexsero® vaccination on the URTome, it is important to first define what constitutes the microbiome of healthy infants. Figure 48 shows the top 20 genera identified for all time points, including both cohorts, in chronological order, each time point normalized depending on the number of samples in it. Some trends can be identified:

- There is a development of the infant URTome over the first three years of life that leads to changes in the relative proportion of different genera over time; *Streptococcus* was the most abundant genus identified, although its relative abundance tends to decline over time.

- The samples from Oxford do not seem to differ from those of Dormice in terms of microbiome progression.

- Regarding *Neisseria* reads, the relative proportion of this genus appears to increase over time.

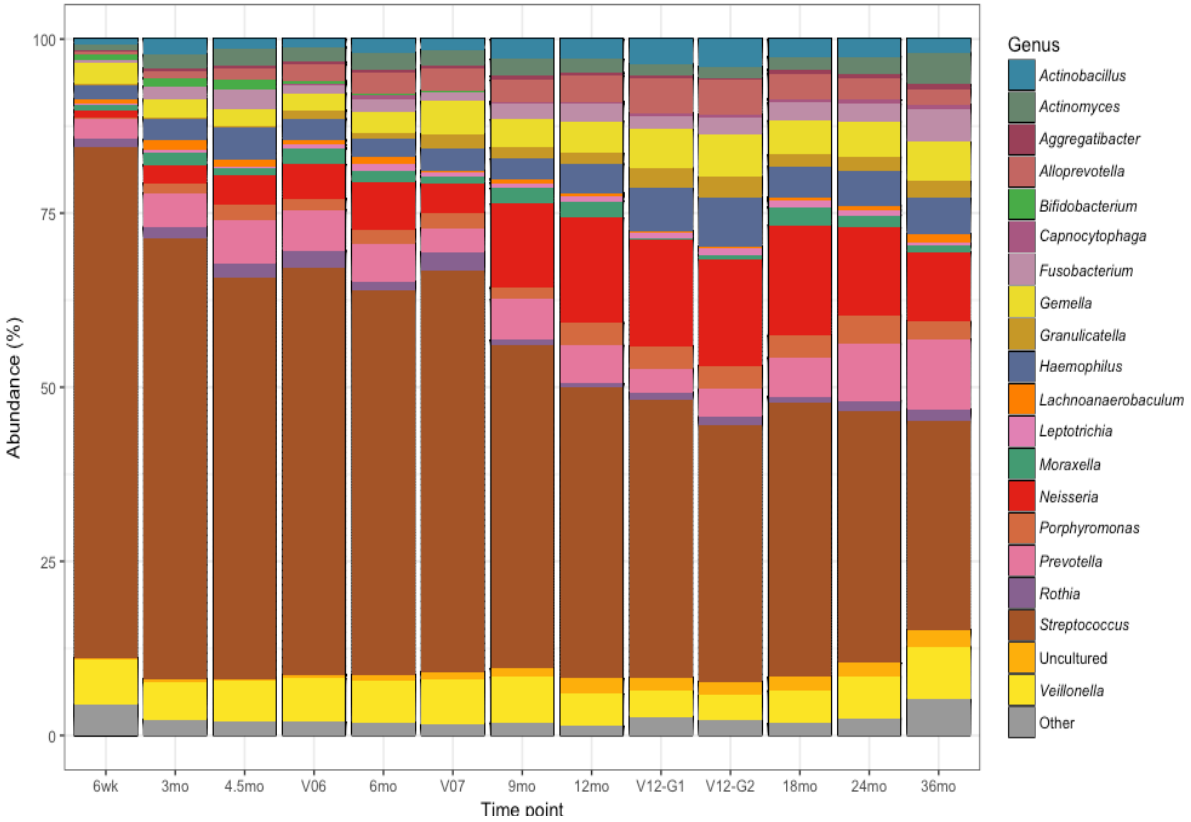


Figure 48 Relative proportion of the average reads (Y axis) for the top 20 genera identified for each time point included in the Dormice and Oxford cohorts (X axis). Each of the 20 most abundant genera are coloured according to the legend on the right. All other readsbelonging to other genera were grouped in the 'Other' category.

Looking closer at comparable time points between the two data sets, it is easier to appreciate differences. Figure 49 shows the average relative abundance of the top 20 genera identified when only selecting the 6 and 12-month time point samples. The genera identified using these samples was

the same as when selecting all the samples, with the exception of *Bifidobacterium,* which was replaced with *Bergeyella* in the top genera list. Some differences were observed, such as the decrease in *Streptococcus* in Oxford samples at later time points, or the presence of *Bergeyella* exclusively in Oxford samples. To further characterize the differences among these data sets, I explored the differences in microbiome composition using ordination and differential abundance methods.

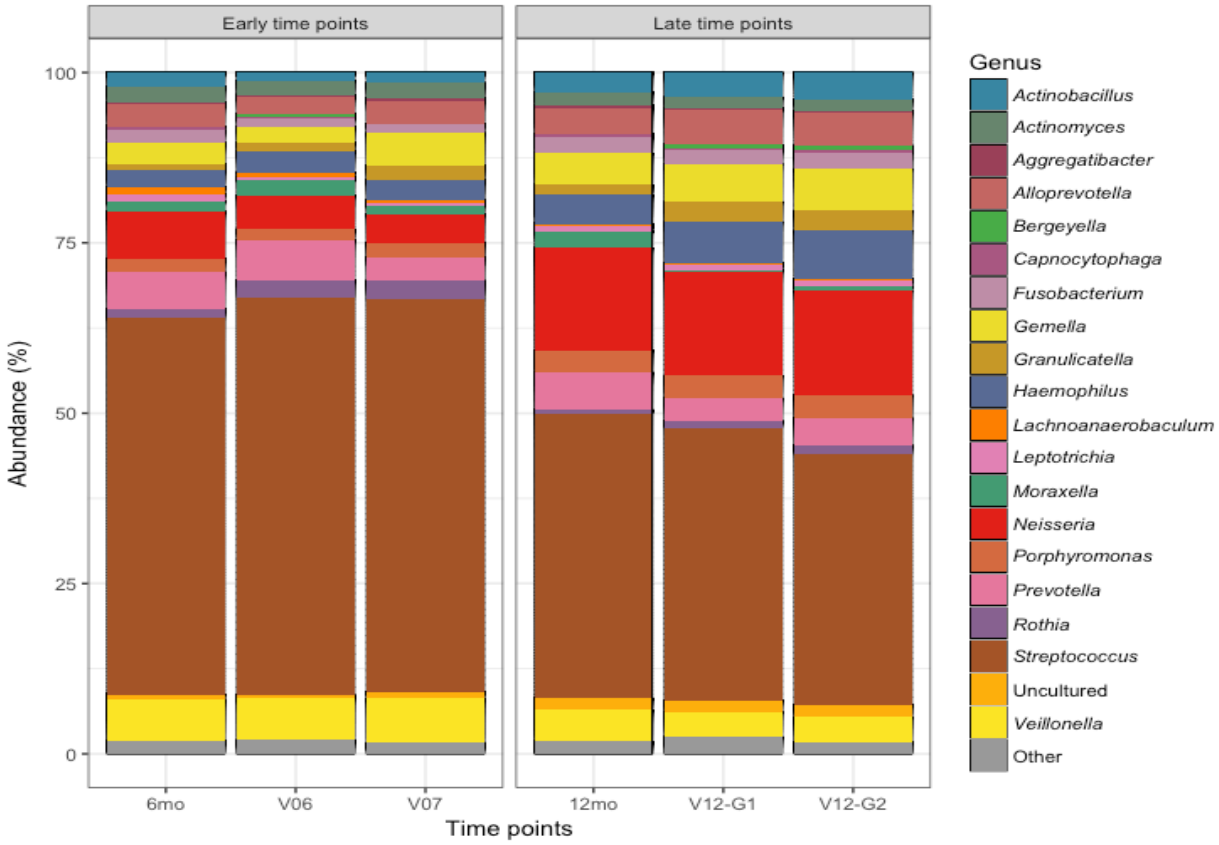


Figure 49 Comparison of sample groups in earlier and later time points for Dormice and Oxford samples. Early (6mo, V06 and V07) and late (12mo, V12-G1, V12-G2) time points are shown in the X axis, and the average proportion of reads assigned to each genus on the Y axis. Each of the 20 most abundant genera are coloured according to the legend on the right. All other readsbelonging to other genera were grouped in the 'Other' category.

6.4.6. Assessment of Neisseria carriage and abundance using 16S rRNA reads

To have a better understanding of the average carriage of *Neisseria,* I selected only *Neisseria* reads from each time point. Figure 50 shows the average percentage of *Neisseria* reads identified per time point in Dormice and Oxford samples. There seems to be a shift in between the 6 and 9-month time points, as the average percentage increases from 6.87% in time points before and up to 6 months, to 12% in all time points after 9-months. The abundance then increases and peaks at the 18-month time point, reaching 15.8% of the microbiome. *Neisseria* abundance then decreases until the 3-year time point.

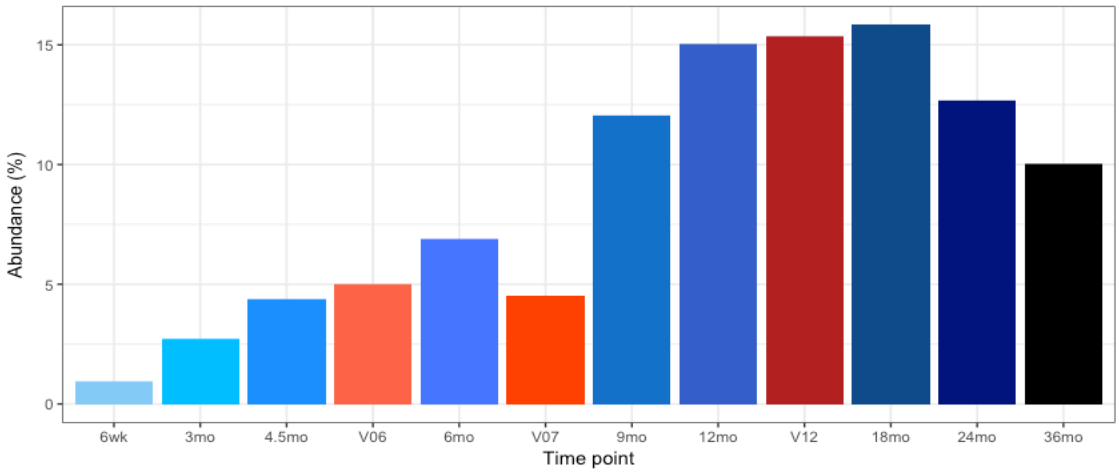


Figure 50 Average percentage of reads belonging to *Neisseria* per time point for Dormice (blue) and Oxford (red) samples. Time points are shown chronologically, with lighter colours indicating earlier time points.

Meanwhile, the proportion of infants colonized by *Neisseria* steadily increases over time, as shown in Figure 51. The percentage of *Neisseria* carriage plateaus around the 1-year time point, reaching almost 100% of carriers after this time point.

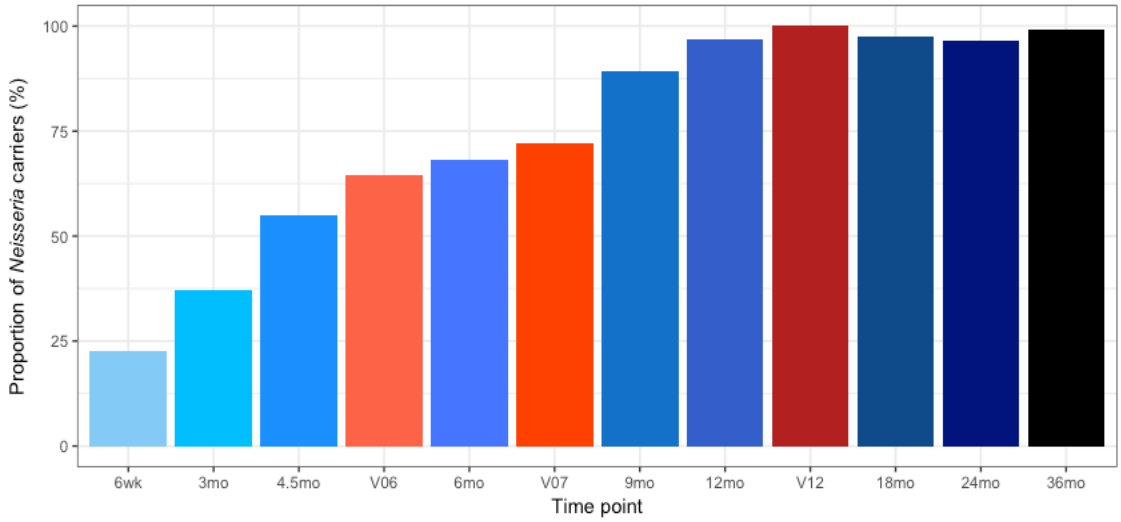


Figure 51 Percentage of infants with *Neisseria* reads per time point. All time points from Dormice and Oxford cohorts are shown chronologically. Dormice time points are shown in blue, and Oxford time points in red, with lighter colours indicating earlier time points.

Looking more closely at individual *Neisseria* OTUs, there are 47 different ones, but only three of these, (*Neisseria*_4, *Neisseria*_9 and *Neisseria*_27) are responsible for 99.86% of the total *Neisseria* reads (Figure 52). A closer look at the distribution of these three OTUs is shown in Figure 53. The abundance of these three OTUs seem to show a temporal effect in colonization (Figure 53), *Neisseria*_4 appearing earlier on in life, then being proceeded by *Neisseria*_27, and finally *Neisseria*_9.

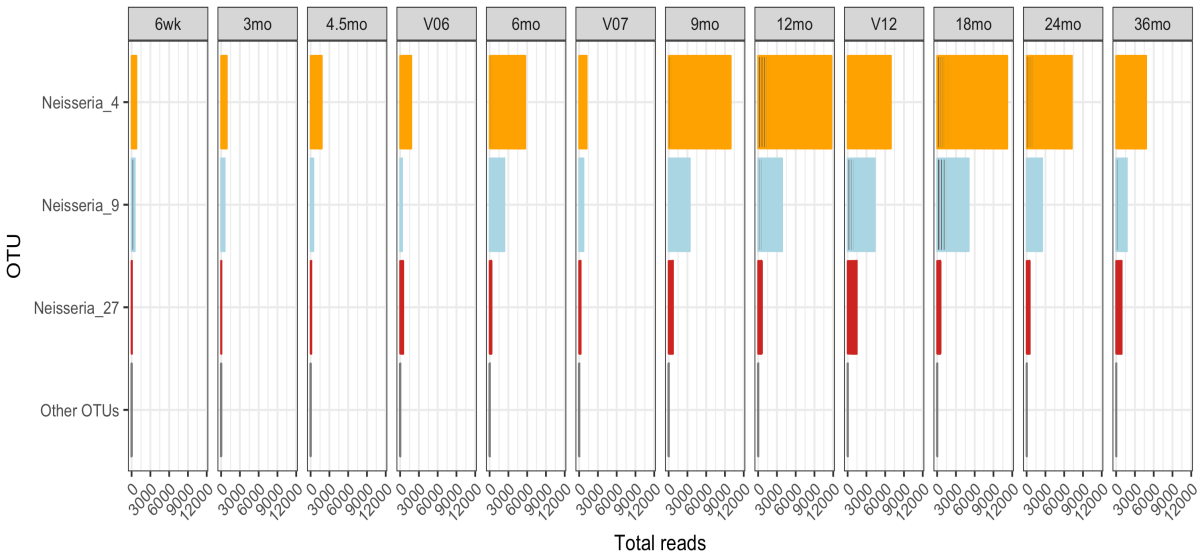


Figure 52 Number of total reads per *Neisseria* OTU and time pont. The three most abundant OTUs (Neisseria_4, Neisseria_9 and Neisseria_27 appear) in yellow, blue and red, respectively. All other *Neisseria* OTUs were grouped into the 'Other' category.

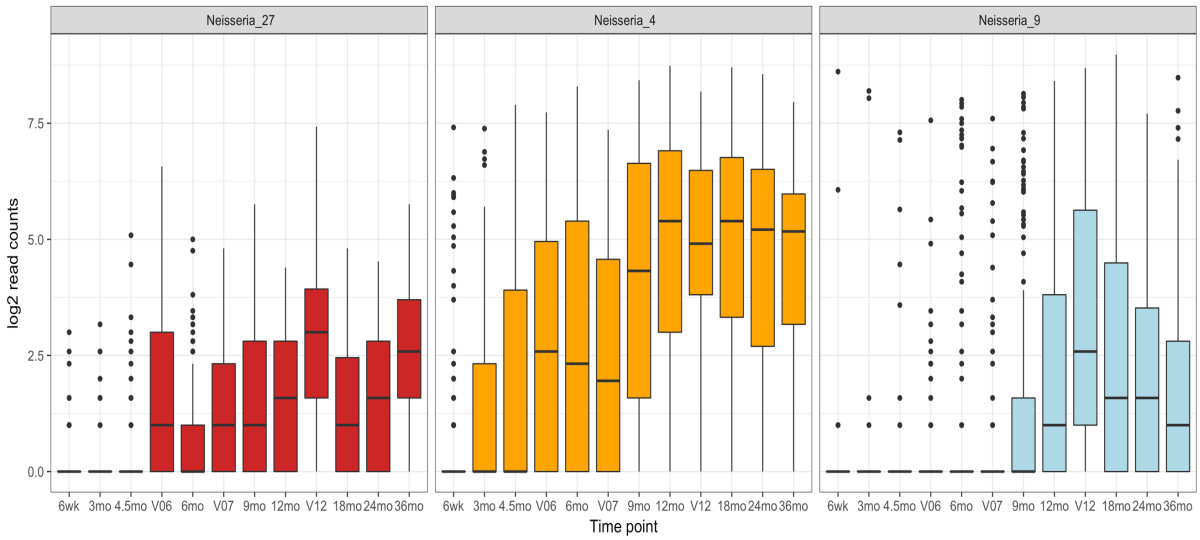


Figure 53 Log2 values of the top three most abundan *Neisseria* OTUs per time point. For each OTU, the read counts are summarized using a boxplot per time point.

6.4.7. Analysis of the microbiome data using ordination methods

Analysis and visualization of highly-dimensional datasets, such as those processed when analysing microbiomes, is difficult, as each different taxonomic level represent a different dimension of the data. To overcome these problems, ordination methods are used to preserve as much information as possible from high-dimensional data into 2 or 3 axes of variation. There are multiple ordination techniques, although the two most commonly used in microbiome analyses are PCoA and NMDS.

The PCoA algorithm generates the gradients of variation by calculating linear combinations of variables that explain most of the variability in the data. This generates a number of different axes of variation (also known as eigenvectors) with different values (eigenvalues) that highlight the amount of variation explained by it. The 2 or 3 axes with highest values are selected and a 2 or 3-dimensional plot is created by placing the samples along the axes. Although this method allows capture of the

variability in the data in a simple and efficient way, the variability visualized using this method depends on the distance metric used and may be less than 50% of the total variability.

NMDS on the other hand tries to summarize the total variability of the dataset by first calculating the distance in between data points including all the dimensions in the data, and then rearranging the points in a lower-dimensional space while trying to maintain the distances obtained in the higher-dimensional space. This algorithm iterates the data point locations by moving points around to better fit the distances. The difference in between the high-dimensional and low-dimensional distance is calculated in a parameter denoted stress, a higher stress representing a bigger difference in between calculated and observed distances. Usually a stress below 0.15 is considered acceptable, and values above 0.2 are considered to not represent the data set properly in the lower-dimensional space. With each iteration the algorithm tries to minimize the stress value until it reaches a convergent point, where more iterations do not decrease stress. Although, in theory, this method allows visualisation of all the variability in a data set, it is much more computationally intensive than PCoA, the stress values after several iterations may not be low enough, or the algorithm may not converge into a single solution as it may reach a local rather than a global minimum.

To explore the data set, I used both PCoA and NMDS analysis, focusing on four different variables: The sample origin, the vaccination status, the time point and the sex of the participants (Table 20). Four different distance metrics were used to assess the differences in between samples: UniFrac distance, Weighted UniFrac distance, Jaccard Index, and Bray-Curtis dissimilarity.

6.4.7.1. NMDS analysis

Samples were divided into two different groups: earlier time points (V06, V07 and 6 months) and later time points (V12-G1, V12-G2 and 12 months). Using the four aforementioned metrics, distance matrices were calculated, comparing all the samples in each group. These distance matrices were then used to generate the NMDS ordinations. For each ordination, I created four different versions, showing different colours for the different values of each of the variables included in this analysis.

For the earlier time points (Figure 54), none of the variables showed a noticeable difference in any of the metrics used, as the confidence interval ellipses and mean value positions for each variable were almost completely overlapping each other. The later time points (Figure 55) showed a similar trend, but some subtle differences were apparent. Looking at the unweighted UniFrac ordinations (Figure 55, E-H), the Dormice samples start diverging slightly from the Oxford samples, while the Oxford samples tend to cluster more closely, when comparing the time points individually and the vaccination status (the sample origin in the later time points represents the same samples as the vaccination status, as all Oxford samples come from vaccinated infants at this point). The same trend was observed when processing the samples using the Jaccard index and Bray-Curtis dissimilarity, but to a lesser extent. Very similar results, for earlier and later time points, were observed when the same data was processed using the PCoA method.

However, for the NMDS ordinations, only those generated using the Weighted UniFrac distance had acceptable stress levels, and R^2 values in the Shepard plots higher than 0.9. Shepard plots correlate the observed dissimilarity (in high-dimensional space) and the ordination distance; in an

ideal NMDS there is a perfect correlation in between these two variables (R^2 = 1). These results are shown in Table 21.

Table 21 Stress values, linear R^2 and non-linear R^2 values calculated for the different NMDS ordinations generated, the values obtained when using the Weighted UniFrac metric are highlighted in red.

Group	Metric	Stress	R^2 linear	R^2 non-linear
Earlier time points	Weighted UniFrac	0.1448	0.919	0.979
	UniFrac	0.1916	0.882	0.963
	Jaccard	0.2016	0.82	0.959
	Bray-Curtis	0.2013	0.821	0.959
Later time points	Weighted UniFrac	0.1300	0.934	0.983
	UniFrac	0.2249	0.832	0.949
	Jaccard	0.2231	0.785	0.95
	Bray-Curtis	0.2232	0.785	0.95

So far, the data suggested that there were no substantial qualitative differences between the two data sets, although some differences may be present between later time points. However, although these methods allow the visualization of the data, they do not provide a quantitative approach to measure differences. To further characterize the potential impact of vaccination, I used the PERMANOVA, distance-based redundancy analysis (dbRDA) and canonical correspondence analysis (CCA) methods, which can be used to estimate the variability explained by a set of variables.

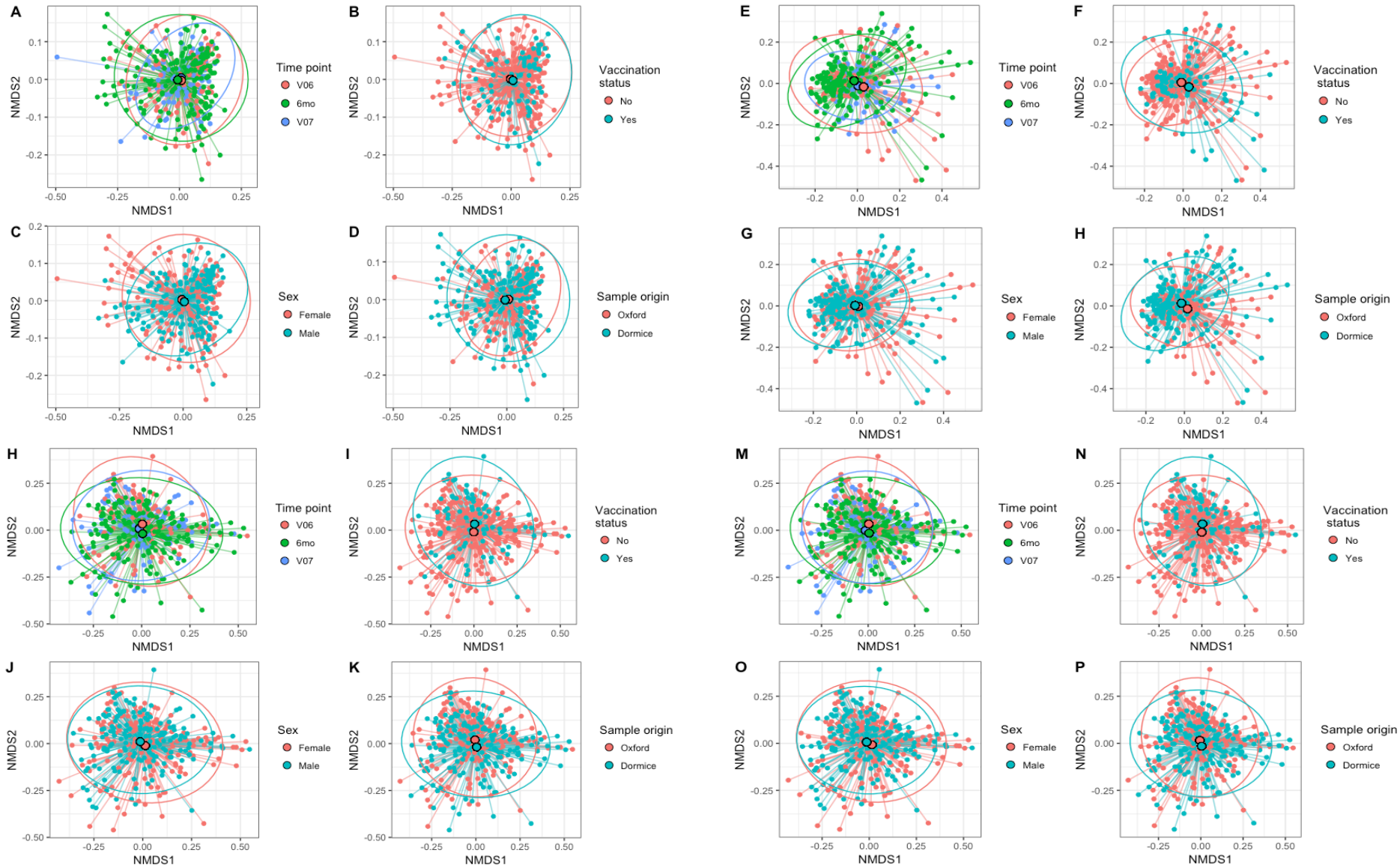


Figure 54 NMDS ordinations generated for the earlier time point samples (6mo, V06 and V07) using weighted UniFrac (A-D), UniFrac (E-H), Jaccard (H-K) and Bray-Curtis (M-P) metrics. Samples are coloured based on its time point, vaccination status, sex, and sample origin. Ellipses represent a 95% confidence interval for each group of samples, and the large circles marked in black represent the average position of each group.

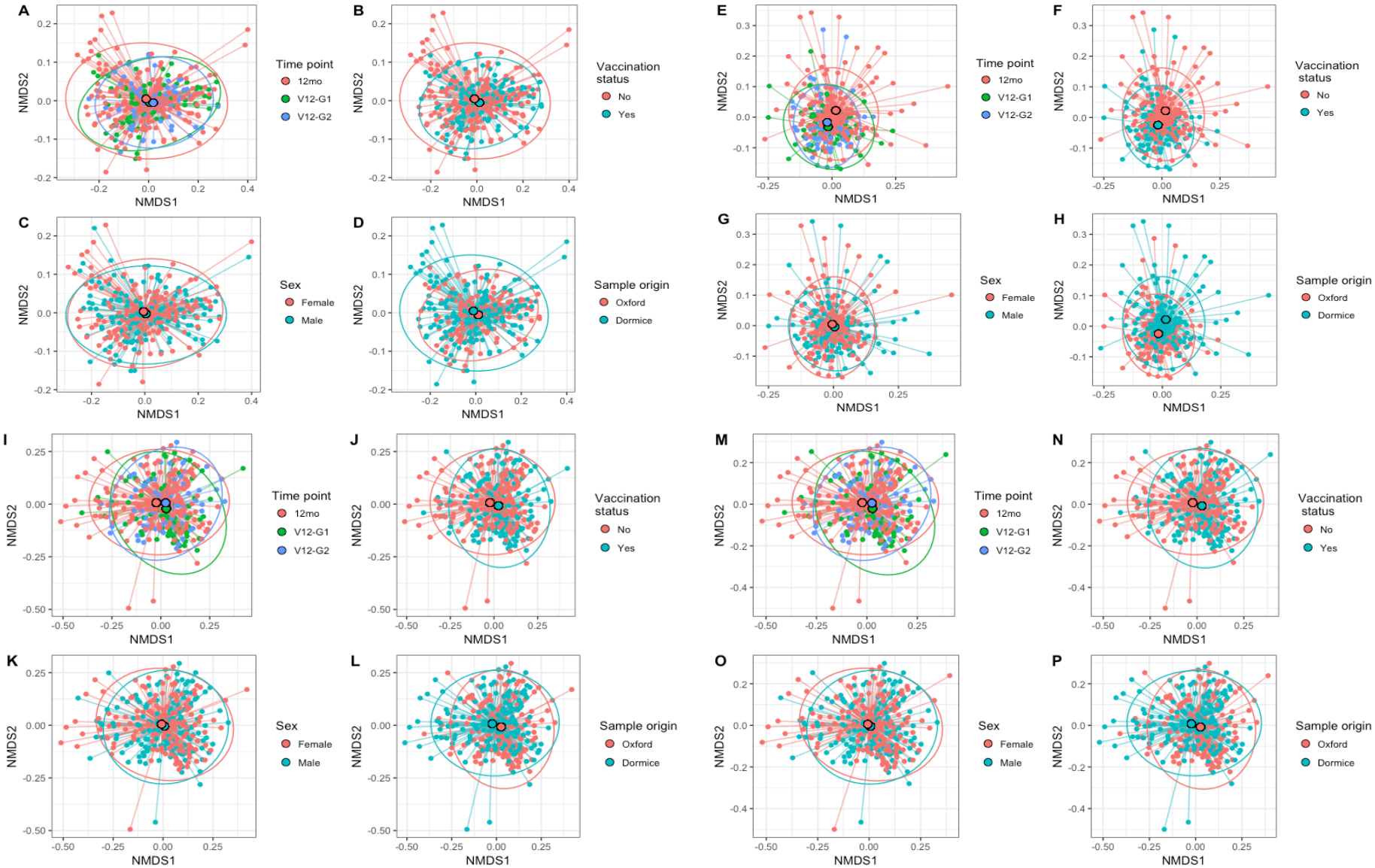


Figure 55 NMDS ordinations generated for the later time point samples (12mo, V12-G1 and V12-G2) using weighted UniFrac (A-D), UniFrac (E-H), Jaccard (H-K) and Bray-Curtis (M-P) metrics. Samples are coloured based on its time point, vaccination status, sex, and sample origin. Ellipses represent a 95% confidence interval for each group of samples, and the large circles marked in black represent the average position of each group.

6.4.7.2. Constrained ordination and PERMANOVA analysis

Constrained ordination methods are used to investigate the variability that can be explained by a particular variable, or set of variables, out of the total variability in a set. In the case of microbiome analyses, this approach can estimate the proportion of variation in the microbiome explained by the variables chosen. Different methods can be used to analyse the data, fitting it to different distributions. Two of these methods, dbRDA and CCA fit the data to a linear and unimodal distribution, respectively. Another method, PERMANOVA, is very similar to the dbRDA algorithm, but its output is more geared towards interpretation of data instead of a graphical representation and can handle negative eigenvalues that can arise during the analysis.

For the constrained ordination analysis, I used the same variables used earlier, and the same distance metrics. For the 'Time point' variable, I performed the analysis using all three groups together, followed by pairwise comparisons. For simplicity, only the results obtained from the Weighted UniFrac metric are shown, as this metric showed the best fit in NMDS analysis, and the two most prominent PCoA axes explained more variation that the other metrics. However, the same trends observed for the Weighted UniFrac results were observed when analysing the results from the other distance metrics used. Results are shown in

Table 22.

Group	Variable	Variance comparison	Inter-group variances	PERMANOVA R2	PERMANOVA p-value	dbRDA R2	dbRDA p-value	CCA R2	CCA p-value
Early time points	Time point	6mo-V06	0.4876	0.0124	0.0100	0.0144	0.0040	0.0049	0.00010
		6mo-V07	0.0122	0.0161	0.0040	0.0063	0.0300	0.0044	0.02000
		V06-V07	0.3255	0.0082	0.2200	0.0016	0.2570	0.020	0.1040
		All time points		0.0185	0.0500	0.0121	0.0010	0.050	0.00010
	Vaccination status	No - Yes	0.6668	0.0060	0.0050	0.0074	0.0020	0.022	0.00020
	Sex	Female - Male	0.1718	0.0019	0.6330	0.0121	0.1460	0.049	0.0920
	Sample origin	Dormice - Oxford	0.0178	0.0149	0.0010	0.0112	0.0020	0.046	0.01000
Later time points	Time point	12mo - V12-G1	0.0818	0.0201	0.0060	0.0259	0.0010	0.0112	0.00010
		12mo - V12-G2	0.0009	0.0262	0.0030	0.0101	0.0050	0.0099	0.00010
		V12-G1 - V12-G2	0.3916	0.0075	0.3720	0.0024	0.2300	0.012	0.7450
		All time points		0.0295	0.0020	0.0225	0.0010	0.110	0.00010
	Vaccination status	No - Yes	0.0006	0.0295	0.0010	0.0219	0.0010	0.0114	0.00010
	Sex	Female - Male	0.9126	0.0014	0.8550	0.0025	0.9810	0.012	0.9840
	Sample origin	Dormice - Oxford	0.0006	0.0266	0.0010	0.0219	0.0010	0.0114	0.00010

One important factor to consider when performing constrained ordinations is the potential effect of the variance in data, and its potential to confound in the analysis. Warton *et al*. (165) highlighted this problem, by showing two data sets with the same total abundance for certain taxa, but one more dispersed, are identified as significantly different using constrained ordination methods. To ensure that the differences identified do not arise from variability but from differences in abundance, the homogeneity of the variance in between sample groups was also compared. I

performed pair-wise comparisons using multivariate Levene's test to identify those groups in which

Group	Variable	Variance comparison	Inter-group variances	PERMANOVA R2	PERMANOVA p-value	dbRDA R2	dbRDA p-value	CCA R2	CCA p-value
Early time points	Time point	6mo-V06	0.4876	0.0124	0.0100	0.0144	0.0040	0.0049	0.00010
		6mo-V07	0.0122	0.0161	0.0040	0.0063	0.0300	0.0044	0.02000
		V06-V07	0.3255	0.0082	0.2200	0.0016	0.2570	0.0020	0.1040
		All time points		0.0185	0.0500	0.0121	0.0010	0.0050	0.00010
	Vaccination status	No - Yes	0.6668	0.0060	0.0050	0.0074	0.0020	0.0022	0.00020
	Sex	Female - Male	0.1718	0.0019	0.6330	0.0121	0.1460	0.0049	0.0920
	Sample origin	Dormice - Oxford	0.0178	0.0149	0.0010	0.0112	0.0020	0.0046	0.01000
Later time points	Time point	12mo - V12-G1	0.0818	0.0201	0.0060	0.0259	0.0010	0.0112	0.00010
		12mo - V12-G2	0.0009	0.0262	0.0030	0.0101	0.0050	0.0099	0.00010
		V12-G1 - V12-G2	0.3916	0.0075	0.3720	0.0024	0.2300	0.0012	0.7450
		All time points		0.0295	0.0020	0.0225	0.0010	0.0110	0.00010
	Vaccination status	No - Yes	0.0006	0.0295	0.0010	0.0219	0.0010	0.0114	0.00010
	Sex	Female - Male	0.9126	0.0014	0.8550	0.0025	0.9810	0.0012	0.9840
	Sample origin	Dormice - Oxford	0.0006	0.0266	0.0010	0.0219	0.0010	0.0114	0.00010

the variance was not similar to others. The results of this comparisons are shown under the column 'Variance comparison' in

Table 22. The results from this test showed a difference in between Dormice and Oxford samples in group 2 (V07 and V12-G2) that starts and earlier time points and appears until later time points.

When considering all four variables included in the analysis (time point, vaccination status, sample origin, and sex), all variables, with the exception of sex, showed a significant difference. Earlier time points, as seen in previous ordination results, showed less variability as shown by the R^2 values obtained for the PERMANOVA, dbRDA and CCA analysis, compared to later time points. As expected, PERMANOVA and dbRDA showed similar results since these two methods are similar, showing R^2 values in between 0.0016 and 0.0185, suggesting that these variables could explain in between 0.16 and 1.85% of the total variability of the microbiome, these being sex and time point, respectively. Ordination plots obtained from the dbRDA analysis for 6-month samples are shown in Figure 56. Later time points showed similar results to earlier time points in respect to the trends observed in earlier time points, but the differences identified increased, almost doubling in most cases for the R^2 values obtained. In this analysis, sex showed again the lowest R^2 value, and not significant p-values, while vaccination status showed the highest R^2, this variable explain almost 3% of the total variability. Ordination plots are shown in Figure 57.

Table 22 R^2 and p-values obtained for the constrained ordination analyses performed using the Weighted UniFrac using the PERMANOVA, dbRDA and CCA methods. P-values lower than 0.05 are highlighted in red.

Group	Variable	Variance comparison	Inter-group variances	PERMANOVA R^2	PERMANOVA p-value	dbRDA R^2	dbRDA p-value	CCA R^2	CCA p-value
Early time points	Time point	6mo-V06	0.4876	0.0124	0.0100	0.0144	0.0040	0.0049	0.0010
		6mo-V07	0.0122	0.0161	0.0040	0.0063	0.0300	0.0044	0.0200
		V06-V07	0.3255	0.0082	0.2200	0.0016	0.2570	0.0020	0.1040
		All time points		0.0185	0.0500	0.0121	0.0010	0.0050	0.0010
	Vaccination status	No - Yes	0.6668	0.0060	0.0050	0.0074	0.0020	0.0022	0.0020
	Sex	Female - Male	0.1718	0.0019	0.6330	0.0121	0.1460	0.0049	0.0920
	Sample origin	Dormice - Oxford	0.0178	0.0149	0.0010	0.0112	0.0020	0.0046	0.0100
Later time points	Time point	12mo - V12-G1	0.0818	0.0201	0.0060	0.0259	0.0010	0.0112	0.0010
		12mo - V12-G2	0.0009	0.0262	0.0030	0.0101	0.0050	0.0099	0.0010
		V12-G1 - V12-G2	0.3916	0.0075	0.3720	0.0024	0.2300	0.0012	0.7450
		All time points		0.0295	0.0020	0.0225	0.0010	0.0110	0.0010
	Vaccination status	No - Yes	0.0006	0.0295	0.0010	0.0219	0.0010	0.0114	0.0010
	Sex	Female - Male	0.9126	0.0014	0.8550	0.0025	0.9810	0.0012	0.9840
	Sample origin	Dormice - Oxford	0.0006	0.0266	0.0010	0.0219	0.0010	0.0114	0.0010

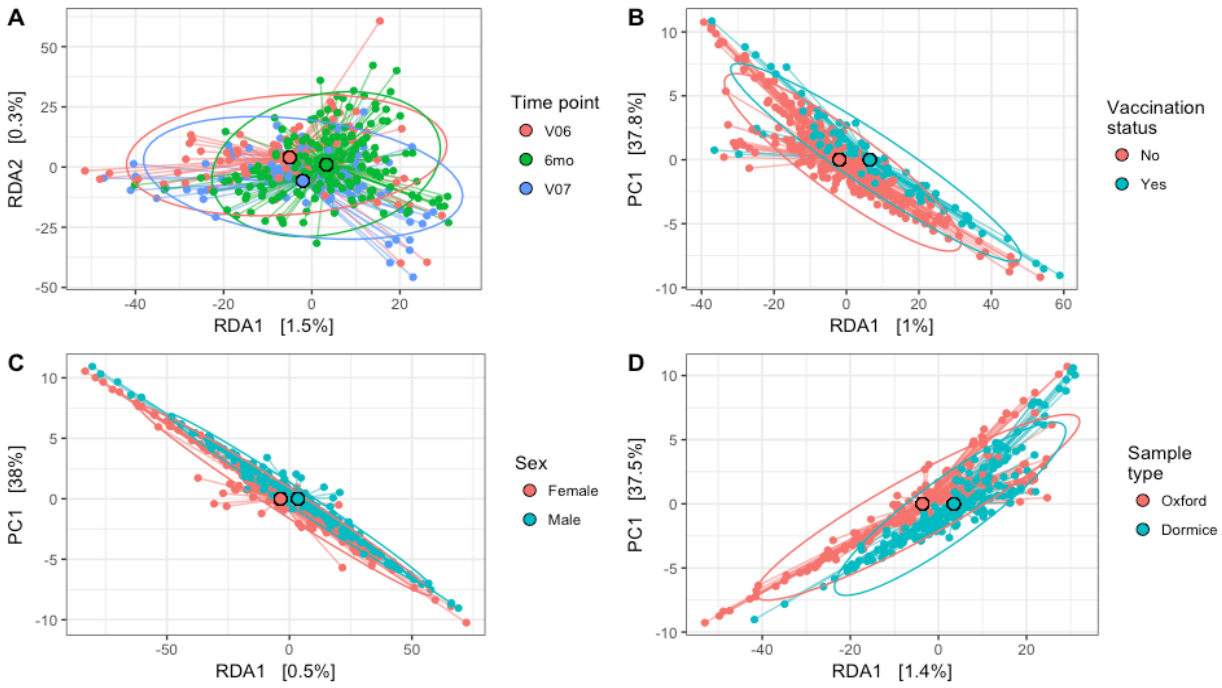


Figure 56 Constrained ordination plots obtained for the earlier time points using the Weighted UniFrac distance and the dbRDA method for the time point (A), Vaccination status (B), Sex (C) and Sample origin (D) variables.

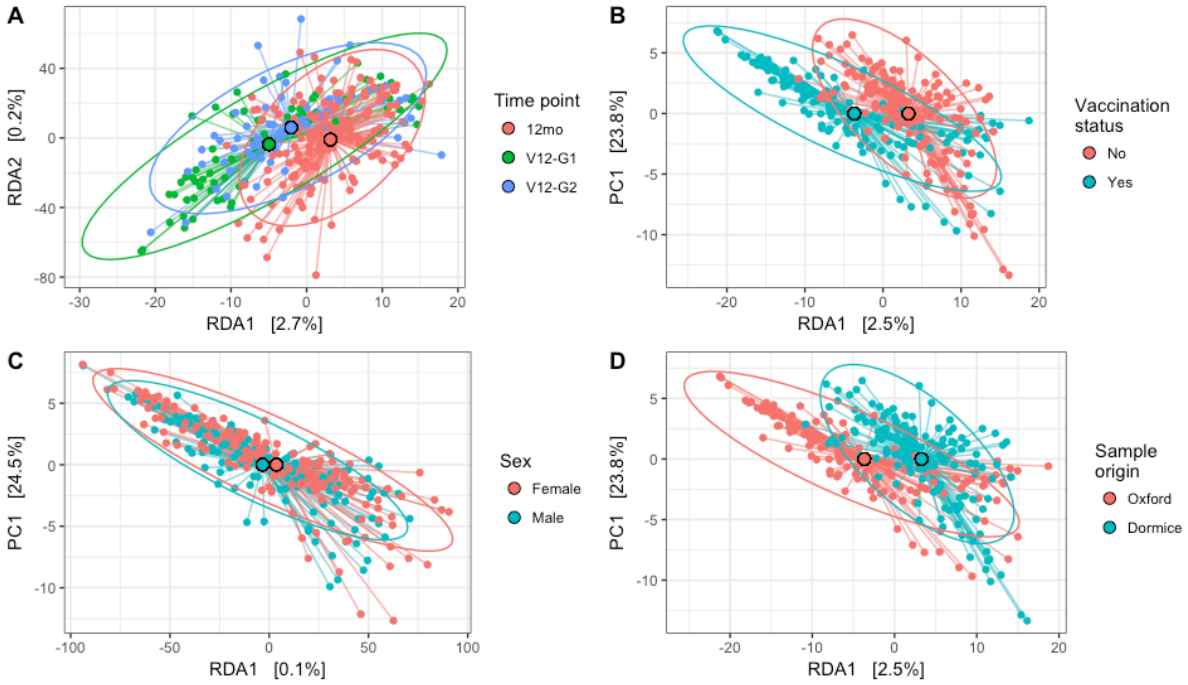


Figure 57 Constraind ordination plots obtained for the later time points using the Weighted UniFrac distance and the dbRDA method for the time point (A), Vaccination status (B), Sex (C) and Sample origin (D) variables.

6.4.8. Differential abundance comparisons

To further characterize the differences observed so far among Oxford and Dormice samples, I used two different packages, metagenomeSeq and LEfSe, to quantify the dissimilarities in microbiome composition. Differential abundance methods rely on normalization of the data to identify differences at certain taxonomic levels, instead of rarefied data, which has been used up to this point in the analysis. Microbiome data is usually characterized by its sparsity caused by the overabundance of 0 values in the OTU count matrix, which leads to non-normality of data if not transformed. Thus, normalization methods are used in order to be able to fit count data to the models used by these packages.

The approach used in LEfSe requires a total-sum scaling normalization of each taxonomic level, and then uses an algorithm similar to that of PCoA, but instead of identifying combinations of OTUs that explain the most variability in the dataset, it looks for arrangements that maximize the separation between the groups selected, identifying combinations of taxa that characterize each group when compared to the other groups included in the analysis. For each of the taxa associated with a group, an LDA score is calculated. Larger LDA score values indicate a stronger association with the group it belongs to. When comparing two groups, it displays one group with positive LDA values and the other with negative values to help emphasize differences. When comparing 3 or more groups, all are displayed with positive LDAs.

In contrast, metagenomeSeq uses a cumulative-sum scaling normalization that, instead of normalizing values by the total sum of reads, scales the values to a percentile of reads. For this analysis

I normalized the reads to the 3rd quantile. The algorithm then fits the normalized data to a zero-inflated log-normal model to assess significant differences between groups.

For the analysis, I will show the results obtained when comparing the variables mentioned in previous analyses, at both the earlier and later time points, with the exception of the sex variable, which showed no significant differences. When comparing the earlier time points, significant differences could be identified in the abundance at different taxonomic levels. Figure 58 shows the different taxonomic levels (from Domain, inner most circle, to OTU level, the outer most) when comparing earlier and later time points (A and B, respectively). The cladograms show differences among groups that arise not just at the OTU level, but at deeper levels such as the abundance of *Actinobacteria* and *Gammaproteobacteria*-related species in Oxford samples in earlier and later time points, respectively. Oxford samples had a higher number of taxonomic levels that were significantly associated with its respective time points compared to Dormice samples.

The differences between Dormice and Oxford, as seen in previous results, seem to increase over time. Figure 59 and Figure 60 are bar charts showing the LDA scores of the OTUs associate the most with each sample group. Regarding *Neisseria* specifically, only the V06 time point showed an association with *Neisseria,* in particular, with an increase of the OTU *Neisseria*_27. This effect persisted at later time points, where this OTU and *Neisseria*_9 were positively associated with the V12-G1 sample group, which are samples from the same infants as V06. Other OTUs identified correspond to commonly identified bacteria in the upper respiratory tract, such as *Streptococcus, Haemophilus, Rothia, Prevotella* or *Gamella*. Oxford samples had an over representation of unidentified OTUs, such as OTUs 110, 58, 82 and 62. Performing a BLAST search of sequences associated with these OTUs did not show a robust association with a classified taxonomy.

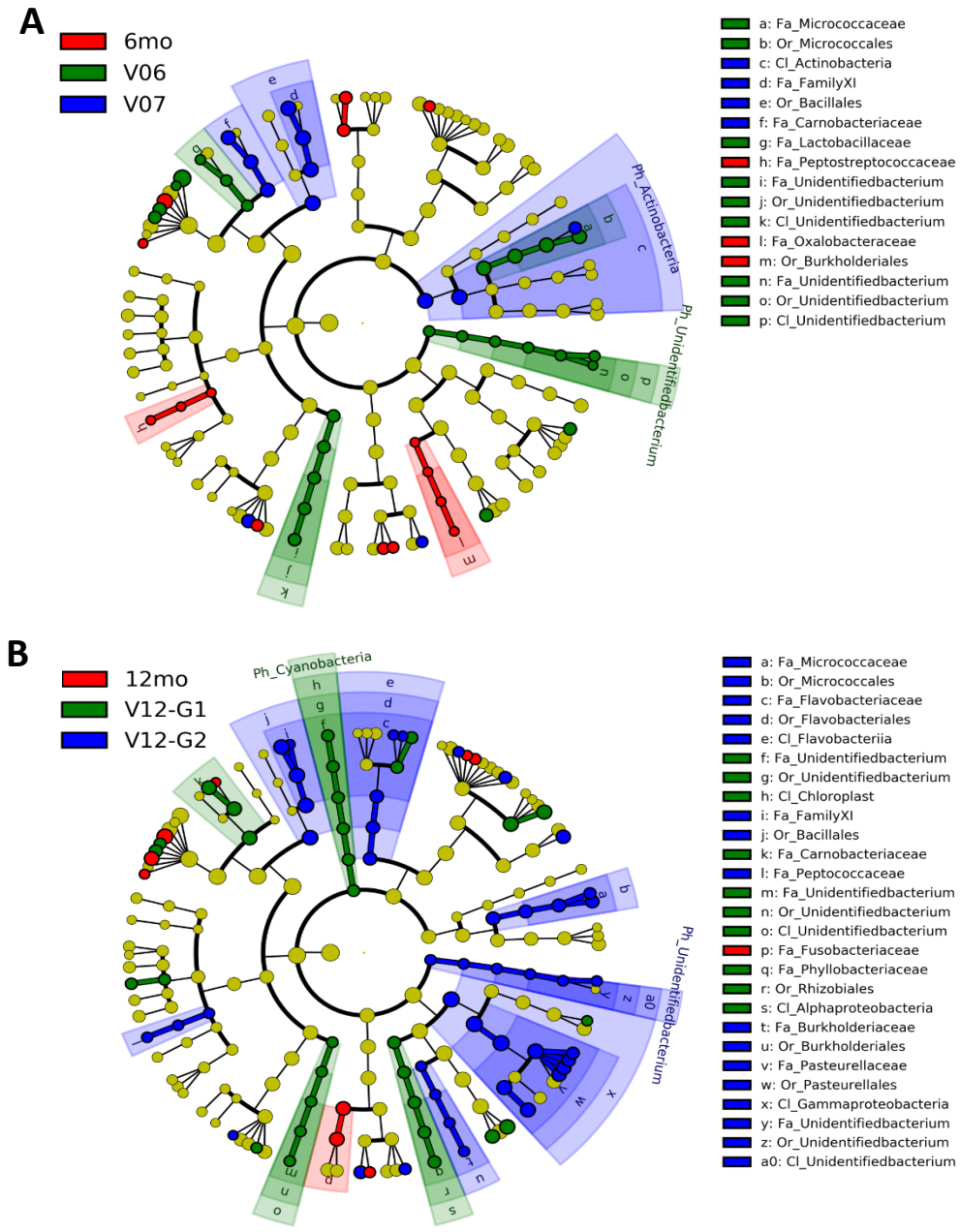


Figure 58 Cladogram generated using taxonomic levels in the OTU reads obtained. Coloured clades and nodes, according to the legend on the top left corner, highlight taxonomic levels where a significant difference was identified in a particular group at earlier (A) and later (B) time points. Taxonomic levels above genus, were marked with a letter, as shown in the legend in the right. A two-letter code was used to designate each taxonomic level (Fa: family, Or: order, Cl: class).

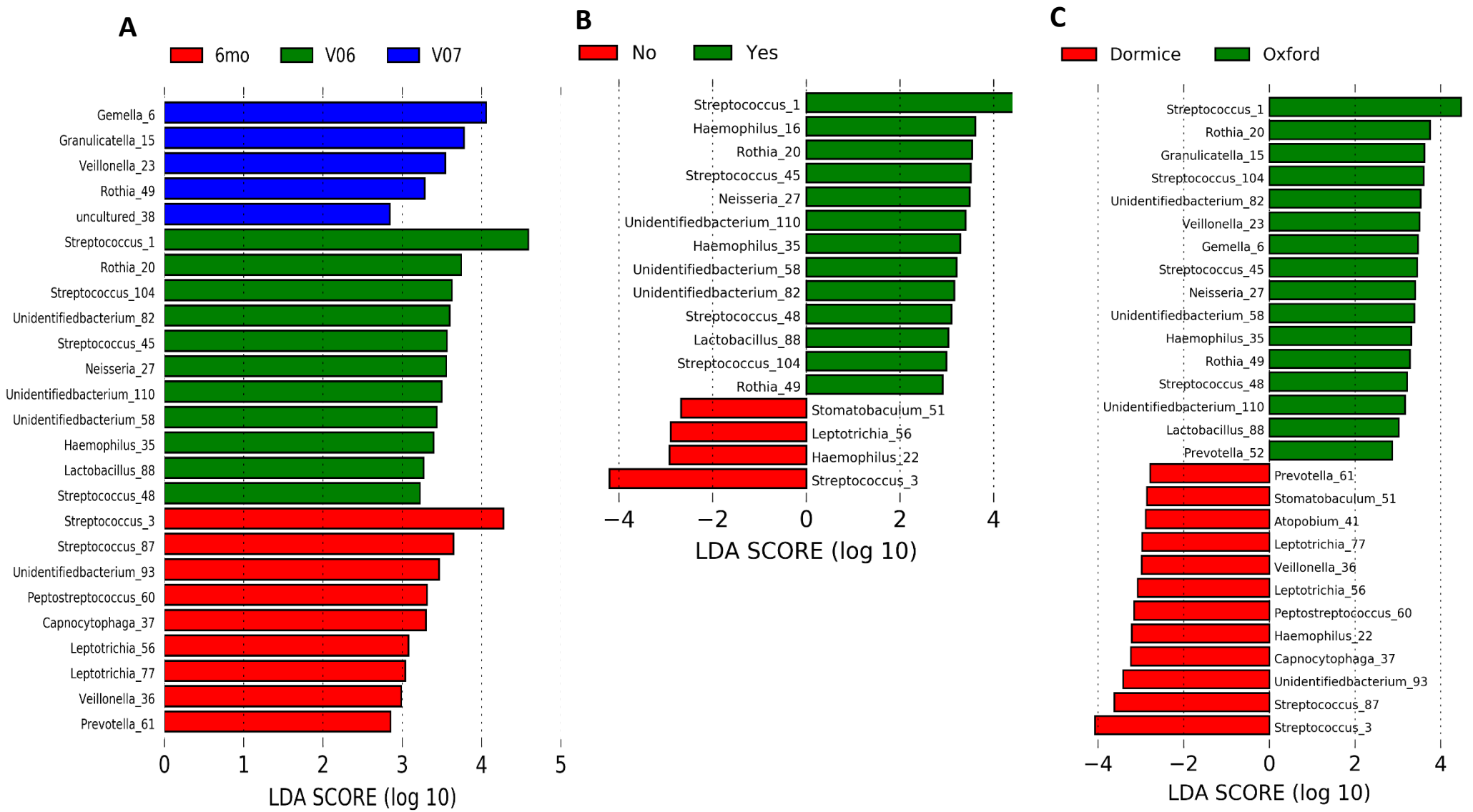


Figure 59 Bar charts showing LDA scores calculated for the OTUs associated with the time point (A), vaccination status (B) and sample origin (C) variables in earlier time points. Larger LDA score represent stronger associations in between an OTU and a group. For each plot, bars are coloured according to the legend at the top.

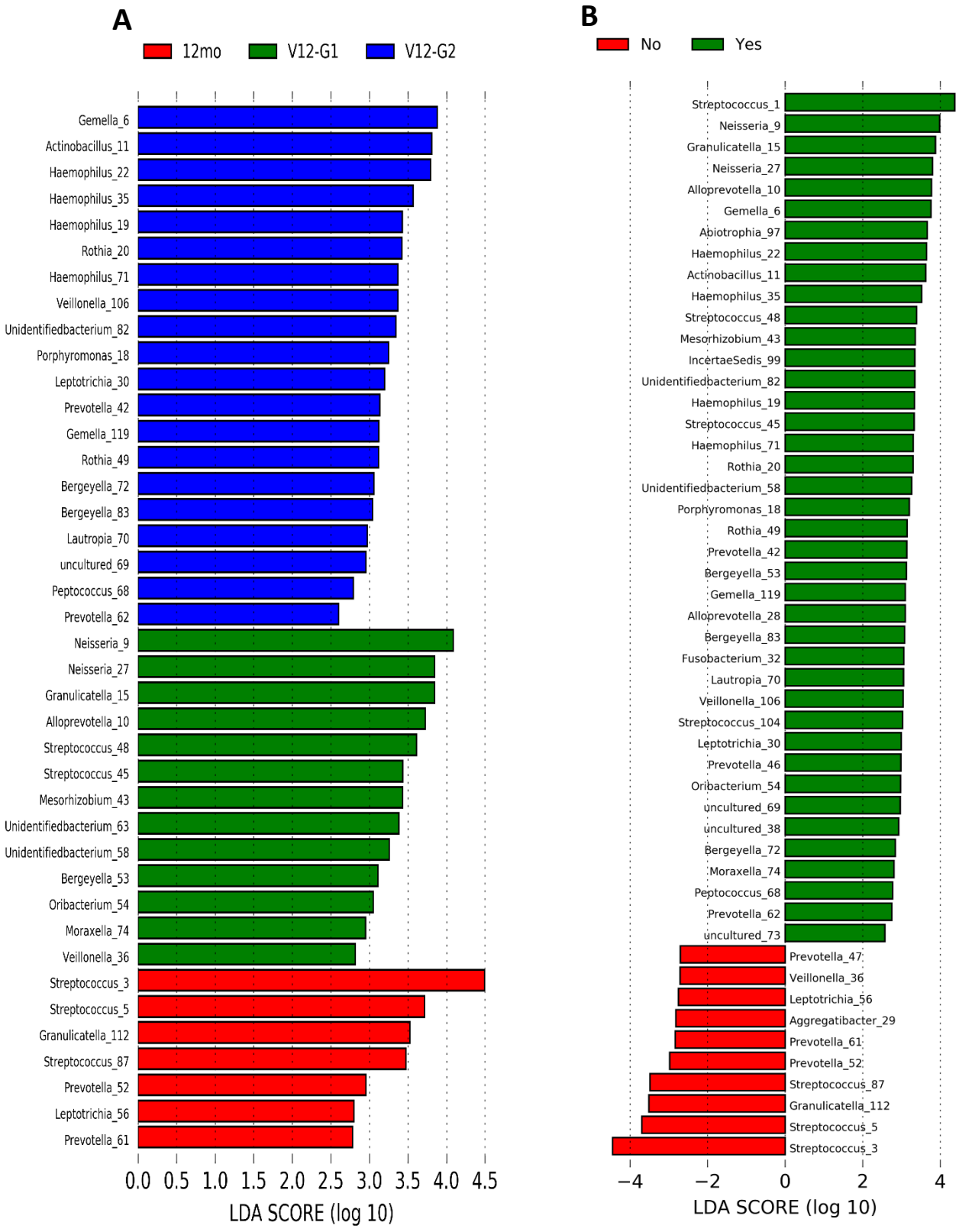


Figure 60 Bar charts showing LDA scores calculated for the OTUs associated with the time point (A) and vaccination status (B) variables in later time points. The sample origin variable is not shown as it is identical to the vaccination status at this point. Larger LDA score represent stronger associations in between an OTU and a group. For each plot, bars are coloured according to the legend at the top.

The results obtained from the metagenomeSeq analysis showed very similar results to those obtained from the LEfSe analysis (Table 23 and Table 24). To avoid repetition, I selected the top 10 OTUs that had an absolute log fold change value higher than 1.3 if one of the comparisons had more than 10 significant OTUs, otherwise all are included. Once again, the sex variable did not show any significant differences, as neither did the comparison of Oxford samples in different groups.

One of the main advantages of metagenomSeq is that it allows a more detailed identification of the underlying differences among groups compared to LEfSe. The aforementioned increase in *Neisseria* reads (*Neisseria*_27 OTU) arises from a 243 read difference in 11 samples that appears in V06 samples when compared to Dormice 6-month samples. In earlier time points the two OTUs with the highest log2 fold change were unidentified OTUs with a higher abundance in Oxford samples, followed by *Prevotella*_61 and *Capnocytophaga*_37 that appeared in Dormice samples. In later time points, the highest difference was a large difference in the *Streptococcus*_3 OTU, it being much more abundant in Dormice compared to Oxford samples.

Table 23 OTUs with a significant difference in abundance identified when using the zero-inflated log-normal distribution model in metagenomeSeq for earlier time points. The table shows the number of samples in each group that had reads for each OTU, the total read count in each group, the odds ratio, the log2 fold change plus the standard error, and the adjusted P-values for each OTU.

Comparison	*OTU*	*group 0 (n)*	*group 1 (n)*	*group 0 counts*	*group 1 counts*	*Odds Ratio*	*logFC*	*se*	*adjPvalues*
V06 vs 6mo *Group 0: V06* *Group 1: 6mo*	Unidentified bacterium_58	41	10	252	13	17.209	-2.454	0.527	0.000
	Unidentified bacterium_82	19	5	40	5	10.460	-2.292	0.581	0.001
	Haemophilus_35	45	33	270	101	5.529	-1.484	0.270	0.000
	Streptococcus_45	71	78	422	130	10.601	-1.368	0.175	0.000
	Rothia_49	41	49	150	87	2.697	-1.103	0.218	0.000
	Neisseria_27	42	53	499	256	2.539	-1.050	0.268	0.001
	Streptococcus_48	41	56	211	95	2.227	-1.045	0.234	0.000
	Rothia_20	65	131	1215	1439	1.451	-0.691	0.198	0.005
6mo vs V07 *Group 0: V07* *Group 1: 6mo*	Prevotella_61	19	2	128	3	5.192	-2.432	1.090	0.124
	Unidentified bacterium_58	10	41	13	249	0.069	2.407	0.501	0.000
	Capnocytophaga_37	28	5	519	13	3.138	-2.257	0.831	0.036
	Rothia_49	49	55	87	193	0.226	1.253	0.204	0.000
Sample Type *Group 0: Dormice* *Group 1: Oxford*	Unidentified bacterium_58	106	49	2264	89	4.488	-1.938	0.268	0.000
	Unidentified bacterium_82	61	21	288	26	4.199	-1.687	0.337	0.000
	Capnocytophaga_37	16	41	722	2256	0.344	1.673	0.541	0.023
	Veillonella_36	15	34	287	969	0.407	1.533	0.541	0.042
	Prevotella_61	16	31	99	474	0.489	1.498	0.498	0.029
Vaccination *Group 0: No* *Group 1: Yes*	uncultured_26	108	29	5076	503	1.241	-1.191	0.366	0.027
	Haemophilus_35	103	53	979	932	0.327	1.149	0.221	0.000
	Unidentified bacterium_58	105	50	1078	1275	0.399	1.146	0.267	0.001
	Streptococcus_45	228	77	1497	2073	0.198	1.095	0.146	0.000

Table 24 OTUs with a significant difference in abundance identified when using the zero-inflated log-normal distribution model in metagenomeSeq for later time points. The table shows the number of samples in each group that had reads for each OTU, the total read count in each group, the odds ratio, the log2 fold change plus the standard error, and the adjusted P-values for each OTU.

Comparison	*OTU*	*group 0 (n)*	*group 1 (n)*	*group 0 counts*	*group 1 counts*	*Odds Ratio*	*logFC*	*se*	*adjPvalues*
12mo vs V12-G1 *Group 0: V12-G1* *Group 1: 12mo*	Mesorhizobium_43	3	27	3	712	0.032	5.336	0.963	0.000
	Haemophilus_71	10	37	25	362	0.062	2.933	0.503	0.000
	Unidentified bacterium_82	17	30	18	129	0.164	1.986	0.357	0.000
	Streptococcus_3	153	67	44280	6767	Inf	-1.953	1.008	0.000
	Alloprevotella_95	17	4	122	4	2.026	-1.813	0.847	0.000
	Haemophilus_35	87	67	411	1036	0.040	1.637	0.181	0.000
	uncultured_69	16	20	245	463	0.288	1.548	0.535	0.000
	Actinomyces_153	12	2	16	2	2.840	-1.523	0.864	0.000
	Unidentified bacterium_63	8	9	30	555	0.370	1.485	0.815	0.000
12mo vs V12-G2 *Group 0: V12-G2* *Group 1: 12mo*	Mesorhizobium_43	3	22	3	430	0.042	4.844	0.889	0.000
	Haemophilus_71	10	43	25	598	0.040	3.367	0.499	0.000
	Streptococcus_3	153	67	44280	6797	0.000	-2.269	1.064	0.000
	Prevotella_92	15	2	59	2	3.517	-2.263	1.034	0.000
	Streptococcus_87	81	7	182	13	9.551	-2.208	0.467	0.000
	Prevotella_61	52	6	263	12	5.201	-2.046	0.607	0.000
	Unidentified bacterium_82	17	34	18	119	0.123	2.009	0.355	0.000
	Granulicatella_112	52	5	91	6	6.341	-1.992	0.549	0.000
	uncultured Fusobacterium sp._126	2	3	12	68	0.284	1.857	1.534	0.000
Vaccination *Group 0: No* *Group 1: Yes*	Mesorhizobium_43	3	49	3	1142	0.036	5.199	0.924	0.000
	Haemophilus_71	10	80	25	960	0.050	3.179	0.491	0.000
	Streptococcus_3	153	134	44280	13564	Inf	-2.114	1.147	0.240
	Unidentified bacterium_82	17	64	18	248	0.142	1.983	0.336	0.000
	Haemophilus_35	87	128	411	2210	0.083	1.613	0.159	0.000
	uncultured_69	16	36	245	850	0.326	1.586	0.461	0.005
	Prevotella_61	52	16	263	61	3.844	-1.546	0.398	0.001
	Bergeyella_140	7	3	22	4	2.120	-1.508	0.887	0.282
	Campylobacter_184	7	3	14	4	2.120	-1.500	0.837	0.253

6.5. Discussion

Analyses detailed in this Chapter suggest that vaccination with Bexsero® has an effect on the microbiome composition, although the change is relatively small. The two data sets included in this analysis, Dormice and Oxford, showed similar trends in terms of α-diversity values, both increasing over time. When looking more closely at comparable time points for each data set, the α-diversity values at earlier time points were most similar compared to later time points.The overall microbiome composition followed the same trends in both data sets, and seemed to be consistent with previous descriptions healthy of infant URTomes (166).

Focusing on *Neisseria* reads, it is important to highlight that by the one-year time point, almost all infants were colonized by this genus. Some publications suggest that the prevalence of *Neisseria*, not taking into account species, is below 40% (6,51,118), but the data suggests that many infants are colonized by *Neisseria*, and that culture-based techniques may not be the most efficient way of measuring the overall *Neisseria* carriage.

Assessing the effect of a vaccine using 16S rRNA sequencing reads can provide insight into unforeseen effects vaccination may have. Other studies focusing on the pneumococcal seven-valent conjugate vaccine showed a transient effect on the URTome in vaccinated compared to non-vaccinated infants (161). While it could be expected that changes on the URTome due to vaccination are temporary, it is still complicated to assess any potential changes caused by vaccination rather than microbiome maturation for the data sets analyzed.

Two major factors presented an added difficulty to the analyses. The Dormice control were collected in London, while the vaccinated infants were collected in Oxford. While one could expect similarities between the data sets, as there are, the factors at which infants are exposed in these two cities may results exposure to different bacteria that may be the underlying cause of variation measured and not necessarily a secondary effect of the vaccine. Also, while there is a plethora of medical data collected for the Dormice samples that we could try to consider when measuring the effect of Bexsero® on the microbiome, these data was not collected for Oxford infants.

Despite potential confounding factors as described above, there were some trends that suggested a location effect that could explain the differences seen rather than effects by Bexsero®. For example, the ordination results suggested that there were very few differences at earlier time points compared to later time points. The analyses performed using the NMDS approach showed almost complete overlap of the Dormice and Oxford samples regardless of the grouping variables used, especially for the 6-month comparisons. However, some differences appeared at later time points. Figure 55A and E show the ordination of all the time points using Weighted UniFrac and UniFrac distances, respectively. Figure 55E shows a separation of both Oxford time points compared to Dormice, while this difference is much more less pronounced in Figure 55A (made using UniFrac distances), which suggests that the changes seen are not due to different OTUs present in each group, rather than small differences in abundance.

While there were significant differences identified using all three constrained ordination methods, a very low percentage of variation was associated with vaccination, reaching 2.95% by one year. However, it is important to note that Oxford samples were never significantly different to each other. In earlier time points, V06 samples had received two doses of the vaccine while V07 samples

had received one the same day of the vaccination. We could speculate that the vaccine had no impact on the microbiome at earlier time points, as we could expect to see a larger difference at this stage if Bexsero® were to impact the microbiome, but no differences were identified. At the 1-year time point, Oxford samples still showed no difference between them. By this time point, V12-G1 samples had three doses administered, the third one being administered one month before sample collection, while V12-G2 infants had received the second dose 5 months prior to the collection of the sample. Despite this difference in vaccination timing, no significant difference was found. This again suggests the location of these samples is contributing to the variation found.

Looking at the differential abundance of OTUs, the LEfSe and metagenomeSeq algorithms showed similar results to each other. Two *Neisseria* OTUs were identified as different, *Neisseria*_27 in earlier time point and *Neisseria*_9 afterwards. This effect could also parallel what was shown by Biesbroek *et al*. (161), where infants that received a pneumococcal vaccine had a higher amount of streptococcal reads after vaccination.

The other OTUs found support the location hypothesis. Although there are several OTUs that differ in between sample groups, the majority of these are offset by an OTU of the same genus in another group. Examples of this are the *Streptococcus* OTUs 3 and 5 are more prominent in Dormice samples while 104 and 1 are so in Oxford samples, and *Prevotella* 61 and 52 are more abundant in Dormice and 62 and 46 in Oxford.

However, it is not impossible to disregard the possibility that these changes could arise from Bexsero®. There is a lack of research focused on the impact of vaccines on the microbiome, especially

of the upper respiratory tract. One of the components of Bexsero® is PorA, which is imbedded in the outer membrane and OMVs. Several studies have characterized the protein content of neisserial OMVs (64,86,167,168). The upper respiratory tract commensals could have homologous proteins to those found in the OMVs that could lead to antibodies binding to other commensals. For future analyses it would be ideal to have a control group of unvaccinated infants that were in the same location to avoid possible confounding effects. This is however not possible in all circumstances.

It was not possible to conclusively identify an effect of Bexsero® on the URTome. A change of less than 3% of the microbiome, as we observed, suggests that any effect may be minimal. It would be interesting in future analyses to have another time point such as one year after vaccination, similar to (161), to determine if any of the changes identified are permanent of just temporary.

7. Chapter 7: Effect of Bexsero® vaccination in the carriage of Neisseria species

7.1. Introduction

Since the introduction of meningococcal polysaccharide-protein conjugated vaccines for serotypes A, C, Y and W, the incidence of invasive meningococcal disease has fallen over the last decade, yet serogroup B is still the causative agent of most cases in European countries and the UK (169). Although the development of vaccines that can confer protection towards the B serogroup has been an arduous task, two meningococcal serogroup B specific vaccines, Bexsero® and Trumenba®, have been licensed in the last few years. In the UK, Bexsero® was implemented as part of the vaccination schedule as of summer of 2015, and it is given to infants at 8 and 16 weeks, with a third dose after one year of life.

Bexsero® has been shown to be an effective vaccine in infants and adolescents (82,170,171), and has been shown to protect against many strains, covering 73% of the isolates in the UK in the 2014-2015 period (80). However, the potential effect that this vaccine may have on meningococcal carriage is still a debated topic (85).

In this Chapter I will present the results of the comparison of neisserial populations between Bexsero®-vaccinated infants, at two different schedules, and a control group of unvaccinated infants.

I will compare the sequencing data obtained after processing *rplF* sequencing files as well as quantification of *Neisseria* present in the samples.

7.2. Aim

- Determine the effect of Bexsero® vaccination on the carriage of *Neisseria* species in infants.

7.3. Summary of methods

7.3.1. Samples used

The samples used in this analysis were the same as those used in Chapter 6.

7.3.2. Data processing and statistics

MiSeq data was processed using QIIME (MacQIIME distribution v1.9.1, (93)) and MED (99) . The data obtained were then imported into the R environment (V3.4.4) and processed with the phyloseq (v1.22.3, (94)) and vegan (2.5-1, (162)) packages. Statistical tests were performed in R. Differential abundance testing was calculated using the fitFeatureModel function within the metagenomeSeq package (v1.20.1, (163)). LEfSE was calculated using the bioBakery (164) virtual machine provided by the developers (v1.7).

7.3.3. Quantification of neisserial burden in samples

A qPCR-based *rplF* assay was used to quantify the number of *Neisseria* present in each sample. As each genome has only one copy of this gene it should provide a one to one ratio of quantification. All samples were quantified using the PowerUP SYBR Green master mix (Applied Biosystems) and *rplF* primers, using the modified Forward primer. All samples were diluted 1:4 in molecular grade water before quantification. Samples were tested in triplicate and the mean value was used as the final amount per sample. The limit of detection was 400 copies per microliter.

7.4. Results

7.4.1. Sequencing data rarefraction

To determine the optimal rarefraction level for the *rplF* sequencing data, I calculated the α-diversity using the same metrics used for the 16S rRNA analysis: Number of observed taxa, Shannon diversity and the Inverse Simpson (InvSimpson) indexes. Results are shown in Figure 61. The number of observed taxa increased with every sampling depth tested. The Shannon and InvSimpson values increased up to 1500 sequences per sample and then plateaued. This was true for both Oxford and Dormice samples. The disparity with the observed number of taxa could be due to a large number of very low abundance sequence variants present in the data sets.

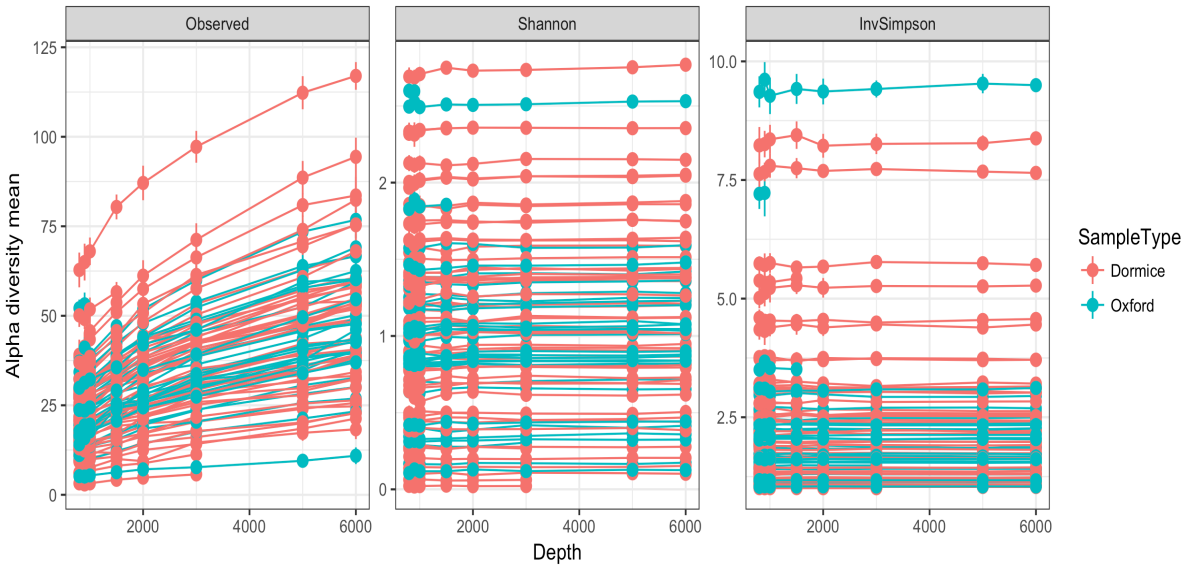


Figure 61 Alpha diversity mean values (Y axis) for Dormice (red) and Oxford (blue) samples at different sampling depths (X axis). The results presented here represent 100 randomly selected samples.

Looking more closely at the number of samples lost during rarefraction, if I had chosen 1500 as the rarefraction depth, 253 out of 1305 samples would have been lost (Figure 62). To ensure that I could capture as much diversity as possible, I decided to set the rarefraction depth to 2000, incurring a loss of 11 additional samples.

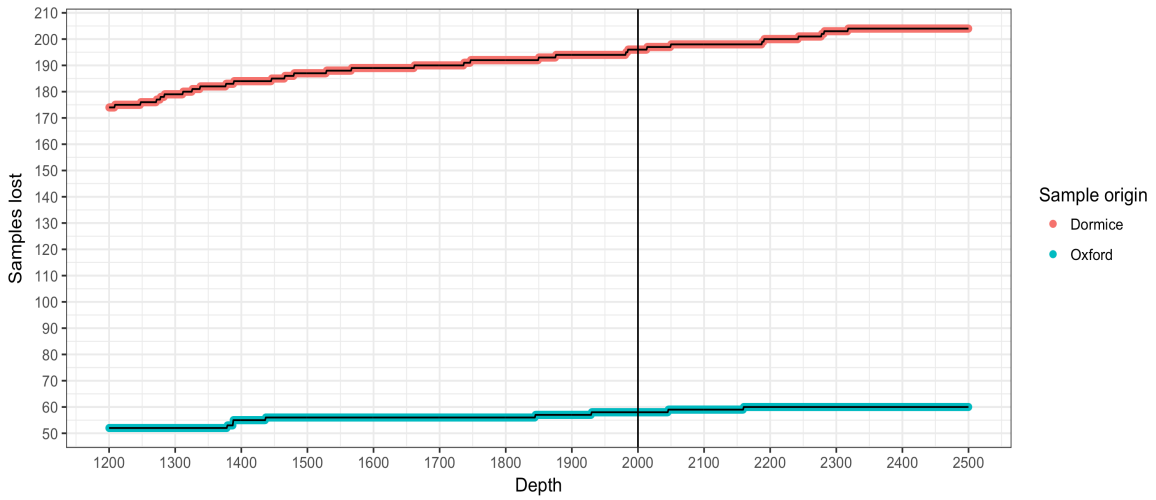


Figure 62 Samples lost (Y axis) per rarefraction depth (X axis) for Dormice and Oxford samples. Each coloured lines indicate the numbers of samples lost for Oxford samples (blue) or Dormice samples (red). The vertical line indicates the sampling depth used for rarefraction.

The data set was comprised of 5644 different taxa and 1305 samples. As mentioned in Chapter 5 (Section5.4.1.3), it was calculated that 4935 different taxa accounted for less 0.0025% of the total read count each. As the processing of the data using MED lacked stringency, the high number of taxa identified may be a result of noise in the data due to lack of filtering rather than true diversity. Two different versions of the dataset were processed, one including all the taxa identified, and another one with the aforementioned very low abundance taxa removed. Both of these were rarefied to the determined 2000 reads threshold. After rarefraction, both versions of the sequencing data had the same number of samples, 1100, so it was decided to keep the filtered data set.

Looking more closely at the samples lost due to rarefraction, the majority of the samples were from the 6-week time point from the Dormice study. The number of samples lost decreased in a chronological order (Figure 63). However, there was no association between the number of sample reads and time point for these discarded samples (p=0.24, as measured by a Kruskal-Wallis test). A breakdown of the total number of samples sequenced and the final number of samples that passed the quality control and rarefraction is shown in Table 25.

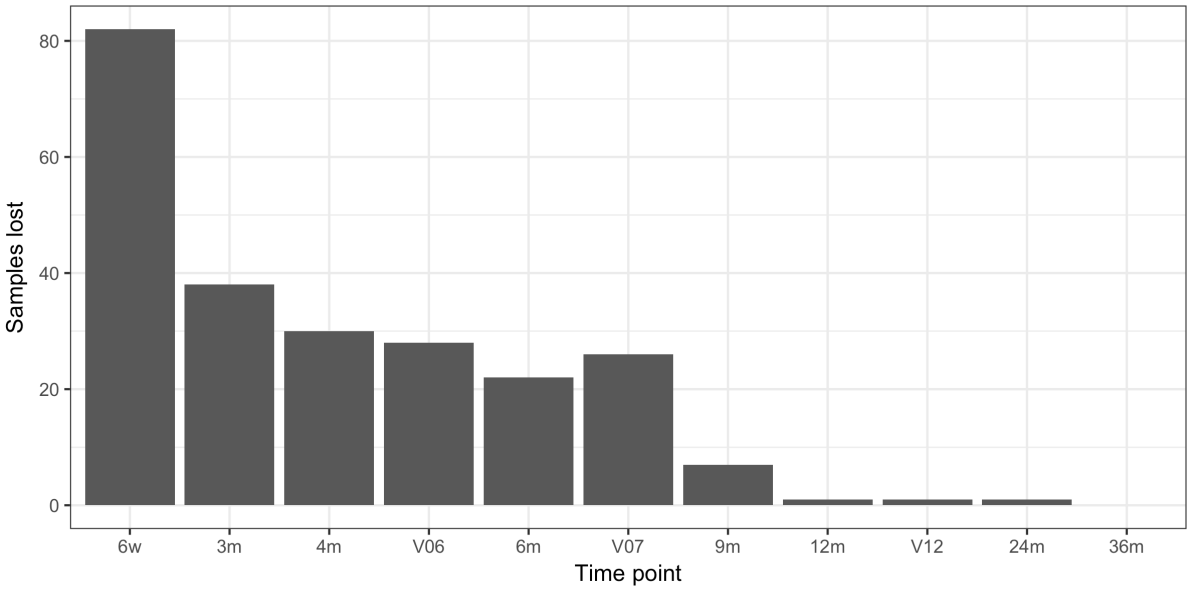


Figure 63 Number of samples lost due after rarefaction (Y axis) per time point of Dormice and Oxford samples (X axis).

Table 25 Number of samples per time point and cohort included in the analysis ("Samples included"), the proportion of samples obtained after sequencing, quality filtering of reads and rarefraction ("Samples after QC"), and the proportion of the total number of samples per time point.

COHORT	AGE	SAMPLES INCLUDED	SAMPLES	PROPORTION OF
Dormice	1.5	111	29	26.13
	3	80	42	52.5
	4.5	74	44	59.46
	6	119	96	81.35
	9	125	118	94.44
	12	117	116	99.14
	18	122	122	100
	24	117	116	99.14
	36	97	97	100
Oxford	6 (Group 1)	89	61	68.53
	6 (Group 2)	90	64	70.71
	12 (Group 1)	89	88	98.87
	13 (Group 2)	90	90	100

7.4.2. Validation of potential biases due to Forward primer binding

As mentioned in Chapter 3 (Section 3.4.4.1), due to one ambiguous base in the Forward rplF primer, two different versions of the primer, *i.e.* that published in (40) and the modified version (this thesis) which includes an ambiguous base at the 12th nucleotide, were tested to determine if this ambiguity could lead to an unbalanced amplification. Two sequencing experiments containing the same mock community samples, each one amplified with one of the two versions of the Forward primer, were performed, and the results were compared. The results of the comparison of four sets of mock communities are shown in Figure 64. Two samples, one prepared using a 99:1 ratio of *N. meningitidis* and *N. lactamica*, and another one with a ratio of 90:10 ratio of *N. polysaccharea* and *N. meningitidis,* did not amplify successfully.

No significant differences were identified when comparing the results obtained for each mock community group to the reference set of values when tested using the Wilcoxon sign rank paired test. However, although these differences were not significant, a small bias towards the amplification of *N. meningitidis* and *N. polysaccharea* was identified when amplifying the samples using the unmodified primers. Therefore, I decided to use the modified primer in subsequent experiments.

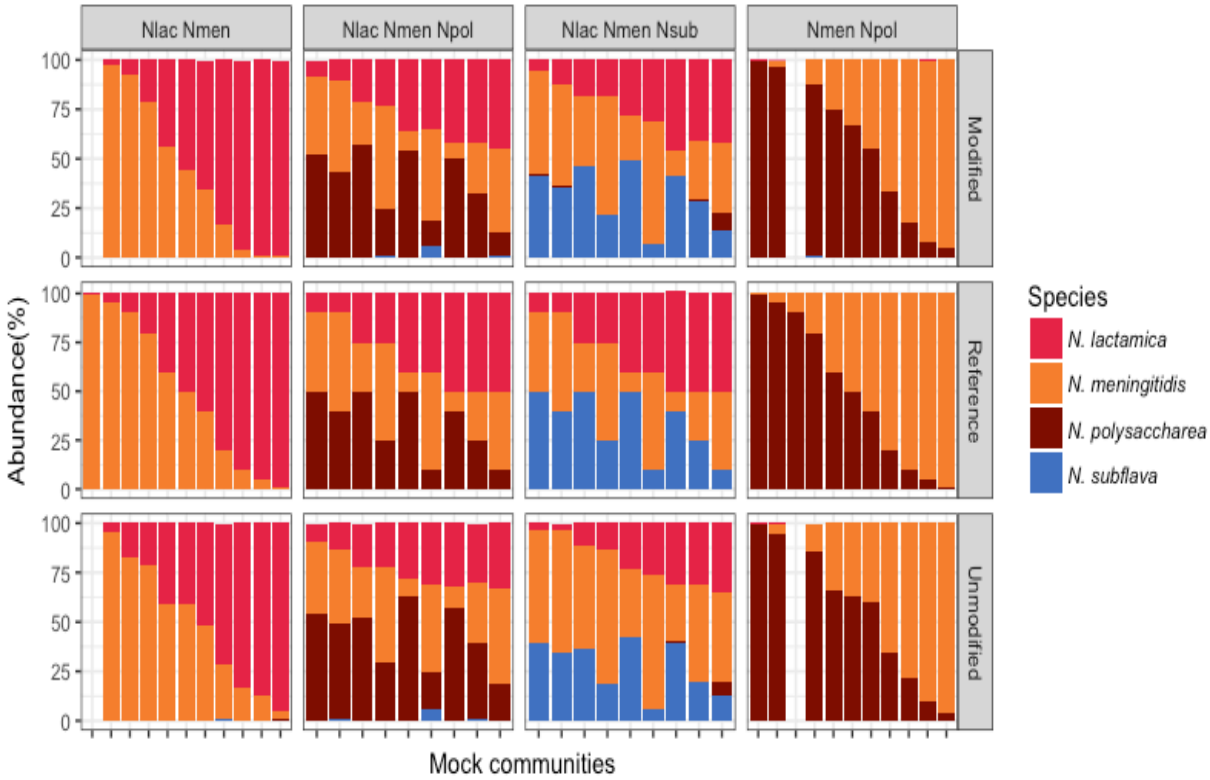


Figure 64 Relative abundance of species identified when sequencing the libraries prepared using the modified and unmodified Forward primer (Modified and Unmodified rows, respectively). A reference of the expected values is shown in the row with the same name. Each column shows the species used to create the mock community groups (Nlac = *N. lactamica,* Nmen = *N. meningitidis,* Nsub = *N. subflava*, Npol = *N. polysaccharea*).

7.4.3. Validation of inter-experiment variation

To identify any potential inter-experiment variation, samples were repeated across multiple libraries. The result of the comparison of 12 randomly selected sample pairs is shown in Figure 65, where each sample pair is shown in each column. The abundance of *Neisseria* species is divided by

rows. Very similar results were obtained when comparing repeated samples and it was determined that the method used was sufficiently consistent.

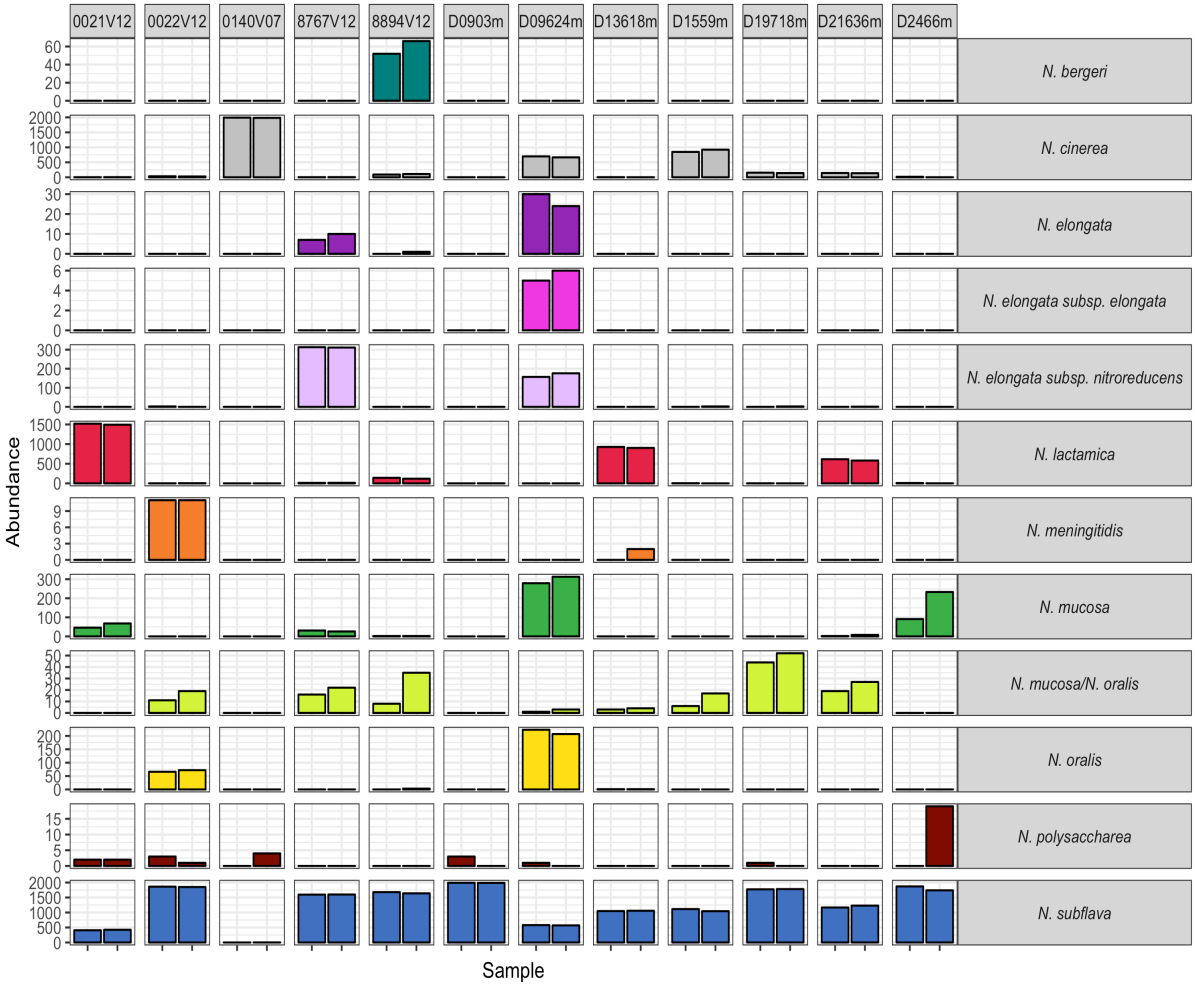


Figure 65 Comparison of 12 randomly selected sample pairs to identify potential sources of variation among sequencing experiments. Each column represents one sample, as per the label above, and each row represents the read counts per genus, showing one bar per repeat of a given sample.

7.4.4. Comparison of α-diversity values.

Alpha-diversity values were calculated per time point including all samples available using the α-diversity metrics used previously: Number of observed taxa, Shannon diversity and the inverse InvSimpson index (Figure 66). Pair-wise comparisons were performed using the Wilcoxon sign rank

test, as none of the diversity scores followed a normal distribution, as measured by a significant Shapiro-Wilk test. For ease of interpretability, Dormice and Oxford samples are displayed separately. Inter-time point comparisons for Dormice samples are shown in Table 26 (Observed taxa) and Table 27 (Shannon and InvSimpson). Dormice samples showed an increase in α-diversity over time, showing no significant difference among the 6 weeks, 3, 4.5 and 6-month time points using the Shannon and InvSimpson metrics, and up to 4.5 months for the observed taxa. Later time points showed an increase in α-diversity, although the 'Observed' metric showed no difference in time points from 12 months onwards, while the other two metrics showed significant differences with the exception of the 12 and 18-month, and 24 and 36-month time comparisons.

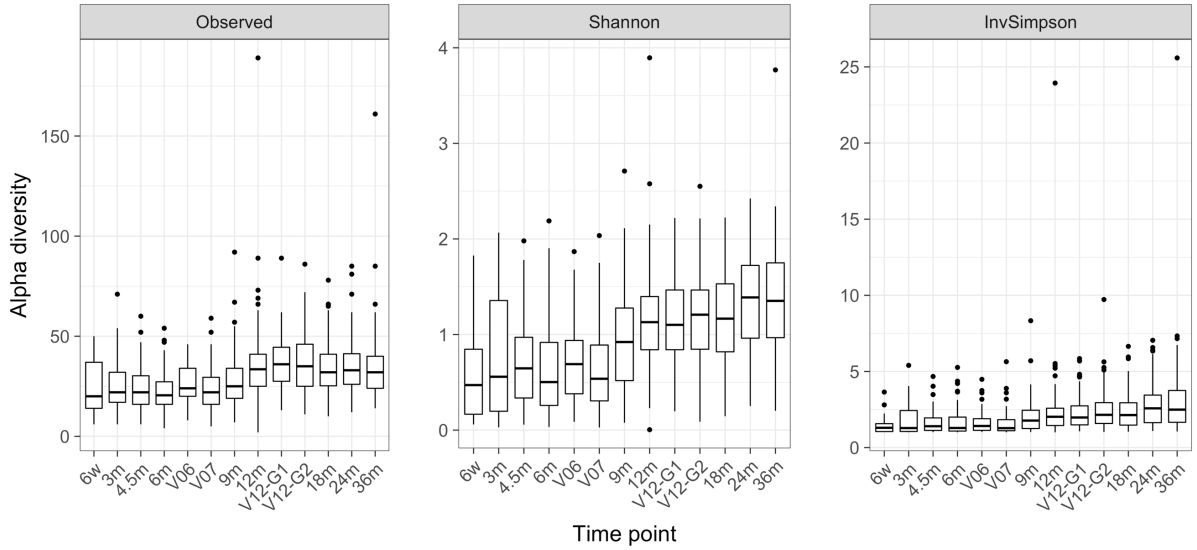


Figure 66 Boxplots describing the alpha diversity values calculated per time point of the Dormice study. Each time point is represented in the X axis, while the Y axis indicates the alpha diversity values calculated using the total number of OTUs (Observed), the Shannon index (Shannon) and the inverse Simpson metric (InvSimpson).

Oxford samples showed a similar increase in α-diversity over time (Figure 67), but no significant differences were observed when only comparing the two earlier or later time points. Finally, all 6-month and 12-month time points for both data sets were compared among themselves. With the only exception of the 6-month and V06 comparison for the number of observed taxa, all others were not significant (Figure 68).

Table 26 Pair-wise comparisons of alpha rarefaction values calculated for the number of observed taxa. Significant values are shown in red

	6W	3M	4.5M	6M	9M	12M	18M	24M
6W								
3M	0.349							
4.5M	0.651	0.6561						
6M	0.934	0.17024	0.33606					
9M	0.069	0.33886	0.08999	0.00033				
12M	0.00024	0.00043	4.20E-06	1.70E-12	1.50E-05			
18M	0.00025	0.00041	4.10E-06	1.00E-12	1.40E-05	0.91446		
24M	9.00E-05	8.60E-05	5.10E-07	3.80E-14	7.40E-07	0.55314	0.47744	
36M	0.00046	0.001	1.90E-05	4.50E-11	0.00014	0.62872	0.68006	0.2812

Table 27 Pair-wise comparisons of alpha rarefaction values calculated for the Shannon (lower triangle) and Inverse Simpson (upper triangle) metrics. Significant values are shown in red

	6W	3M	4.5M	6M	9M	12M	18M	24M	36M
6W		0.27924	0.11322	0.18916	0.00028	9.80E-07	4.20E-07	7.90E-09	1.40E-08
3M	0.32869		0.81221	0.96125	0.02487	0.00116	0.00023	8.80E-07	1.80E-06
4.5M	0.16385	0.94146		0.57934	0.01386	7.30E-05	1.50E-05	8.20E-08	9.70E-08
6M	0.31582	0.76889	0.49364		0.00019	2.40E-08	1.10E-09	2.40E-13	7.40E-13
9M	0.00061	0.04518	0.00942	3.70E-05		0.04053	0.00458	1.40E-06	3.00E-06
12M	8.20E-07	0.00033	2.10E-06	4.00E-11	0.00164		0.37753	0.00124	0.0025
18M	4.10E-07	8.70E-05	1.20E-06	8.10E-12	0.00036	0.5423		0.01927	0.01967
24M	6.20E-09	6.20E-07	2.00E-09	2.90E-16	1.40E-08	0.00144	0.01039		0.94216
36M	1.10E-08	1.80E-06	4.00E-09	5.80E-15	9.40E-08	0.00417	0.02047	0.88727	

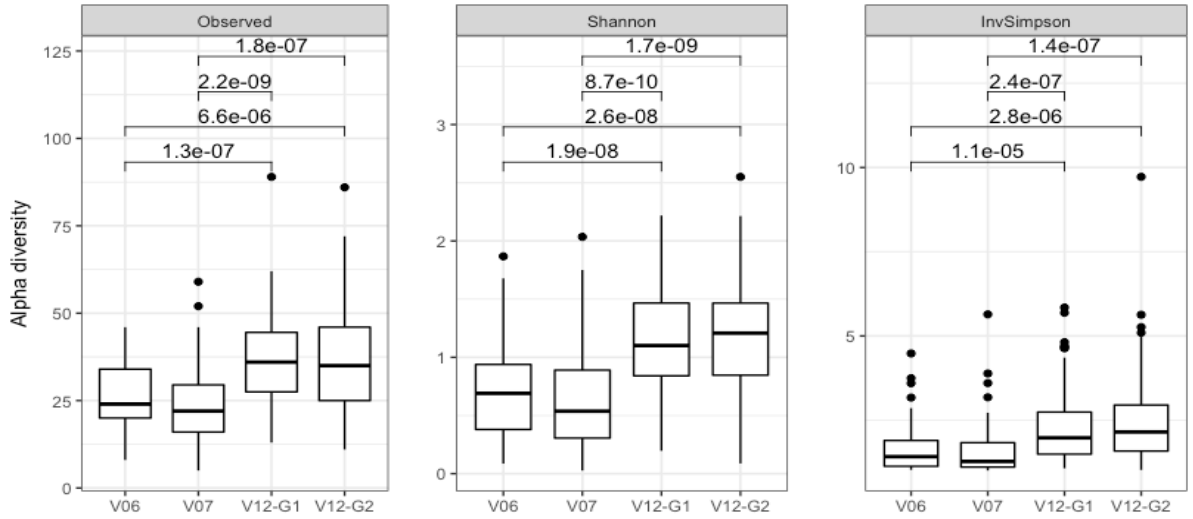


Figure 67 Alpha diversity values calculated per time sample collection group in the Oxford cohort. Each time point is represented in the X axis, while the Y axis indicates the alpha diversity values calculated using the total number of OTUs (Observed), the Shannon index (Shannon) and the inverse Simpson metric (InvSimpson).

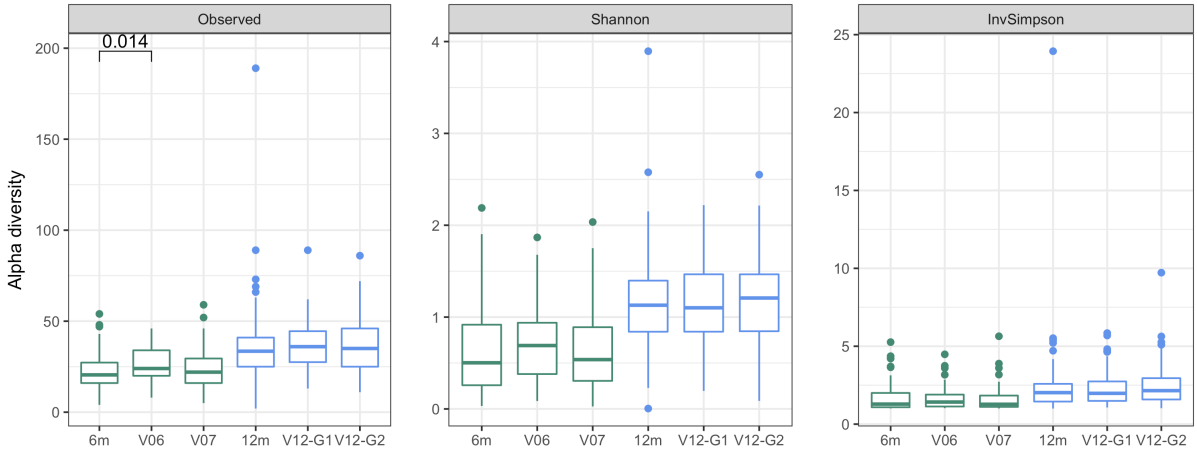


Figure 68 Alpha diversity values calculated for earlier time points (6mo for Dormice, V06 and V07 for Oxford. All shown in green) and later time points (12mo for Dormice, V12-G1, V12-G2 for Oxford. All shown in blue) in both cohorts. Each time point is represented in the X axis, while the Y axis indicates the alpha diversity values calculated using the total number of OTUs (Observed), the Shannon index (Shannon) and the inverse Simpson metric (InvSimpson).

7.4.5. Summary of Neisseria species identified per time point

To identify what proportion of species appeared at each time point, the average abundance of each species was computed per time point. Results are shown in Figure 69. There seems to be development of the relative proportion of *Neisseria* species identified over time. The most abundant

species was *N. cinerea,* it being responsible for 50.3% of all the reads present at the earliest time point, then staying at proportion larger than 40% of up to the 9-month time point. From this point onwards, *N. subflava* was the most prominent species, staying over the 40% mark until the latest time point. *N. lactamica* appeared in a relatively stable proportion throughout all time points, averaging 12% of the total reads across all time points, peaking at the 18-month time point at 17.2%.

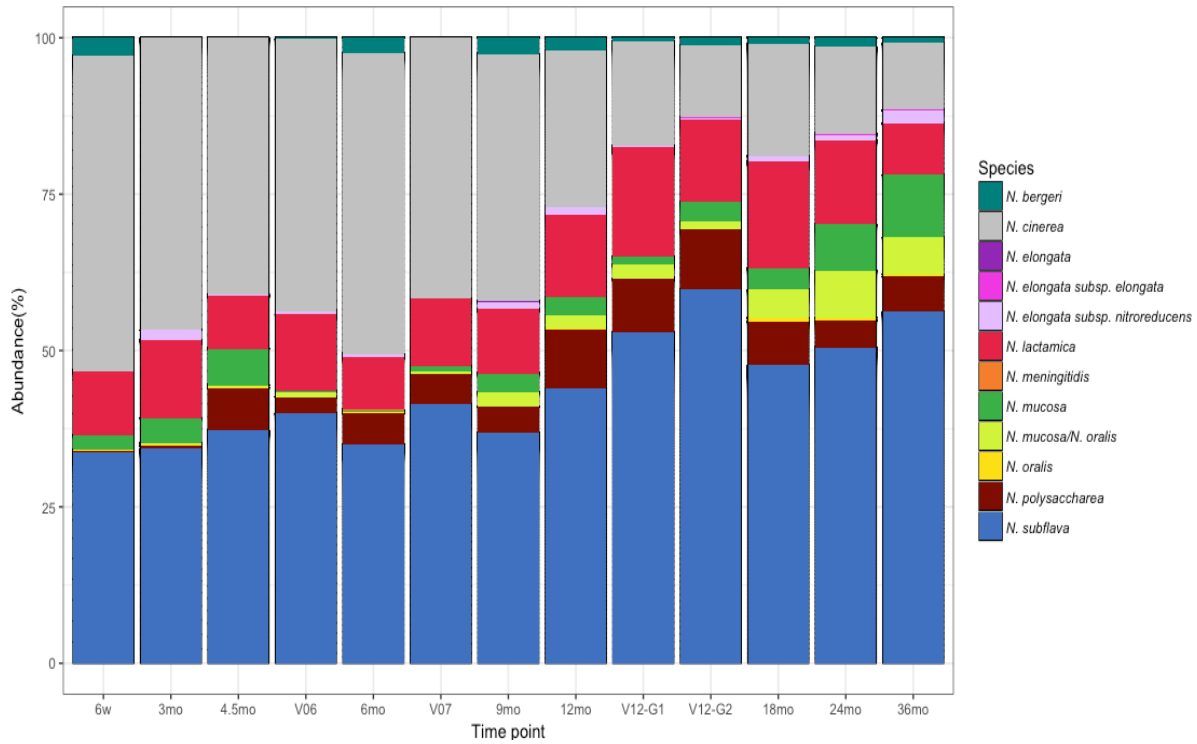


Figure 69 Proportion of *Neisseria* species (Y axis) counts identified per time point (X axis). Time points are displayed chronologically, including Dormice and Oxford samples.

Comparing the earlier and later time points in the two data sets, some differences are apparent. Most noticeably, Oxford samples had a higher proportion of *N. subflava* reads, while Dormice samples had a higher proportion of *N. cinerea* (Figure 70). Other smaller differences appear such as the higher proportion of *N. bergeri* and *N. elongata* in Dormice compared to Oxford samples. *N. lactamica* did not vary significantly between comparing similar time points. However, Oxford

samples in Group 1 (V06 and V12-G1) had a higher proportion of *N. lactamica* reads than the other groups of samples.

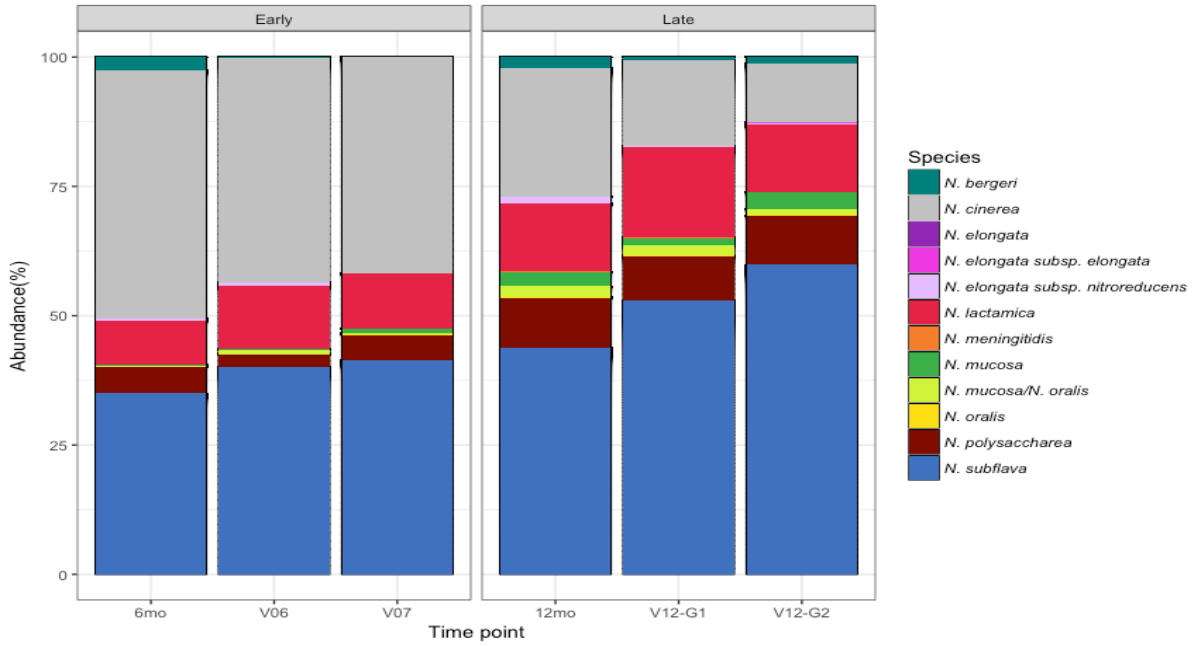


Figure 70 Relative abundances of *Neisseria* species identified in earlier (6mo, V06 and V07) and later (12mo, V12-G1, V12-G2) time points in the Dormice and Oxford cohorts.

For an alternative perspective on carriage of different species, the proportion of infants carrying each species, regardless of co-carriage, was plotted (Figure 71). Most infants carried both *N. cinerea* and *N. subflava* at all time points. *N. mucosa, N. oralis* and *N. elongata* species seemed to have a more pronounced development over time, showing a difference in proportion when comparing time points before and after 9 or 12-months. *N. lactamica, N. bergeri, N. polysaccharea* and *N. meningitidis* appeared to be more stable over time.

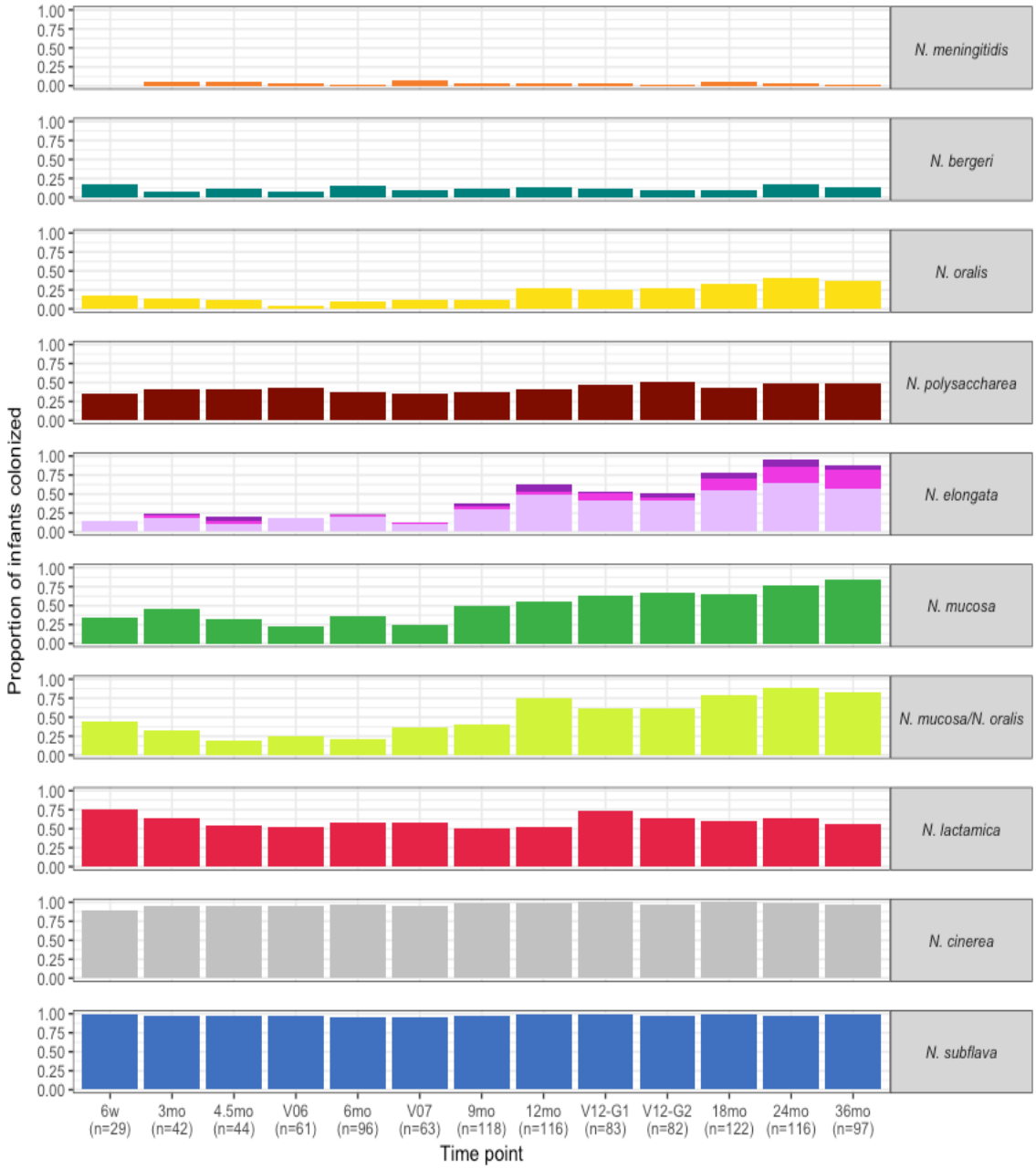


Figure 71 Proportion of infants colonized by each *Neisseria* species identified per time point. The number of samples per time point is indicated underneath each time point. Both *N. elongata* subspecies were merged with the remaining reads for that species. Species were sorted by the total sum of carriers. The number of samples per time point is shown underneath the time point labels.

7.4.6. Effects of Bexsero® vaccination of carriage of Neisseria species

In order to investigate potential effects of Bexsero® vaccination on the composition of *Neisseria* species, I used different ordination methods, as used in the evaluation of the effect of

vaccination on the microbiome (Chapter 6, Section 6.4.7) to visualize and quantify any effect. The variables included in these analyses were the same used in Chapter 6: Sample time point, vaccination status, sex and sample origin.

7.4.6.1. NMDS ordinations

The first step taken was to obtain an overview of the differences among our samples by applying NMDS ordinations using UniFrac, weighted UniFrac, Bray-Curtis and Jaccard metrics. Only the two UniFrac distances could be applied to the data, as the stress in the ordination when using the Bray-Curtis and Jaccard indexes was over 0.2 when testing samples at the earlier and later time points. For the UniFrac distances, the weighted distances provided a more stable result as the Shepard plots showed more linear results, with higher R^2 values (Figure 72). Whenever possible, results are presented from both metrics.

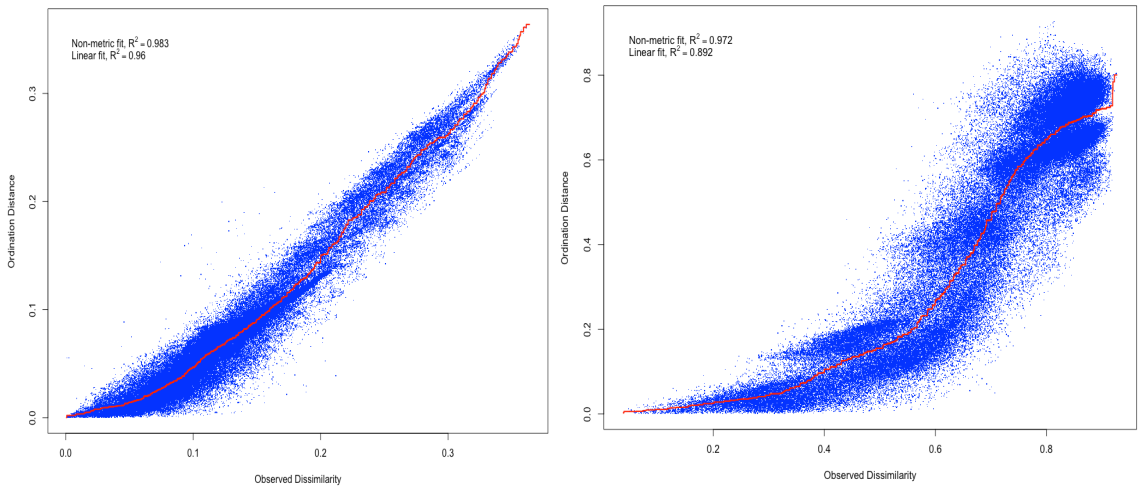


Figure 72 Shepard plots showing the correlation in between the Observed dissimilarities (X) and distances shown in the ordination (Y) for the weighted UniFrac (left) and UniFrac (right) distances. The linear and non-metric R^2 values are displayed on the top left corner of each plot. Blue dots represent the values obtained in each iteration and the red line the mean value obtained.

The results obtained for the NMDS analysis are shown in Figure 73 (earlier time points) and Figure 74 (later time points). In both Figures, plots A-D were made using weighted UniFrac, and E-H using UniFrac distances. For earlier time points no significant differences were identified when comparing variables in neither of the UniFrac metrics. Later time points showed a slight separation of Oxford samples with regards to the Dormice samples, especially when using the weighted UniFrac distance.

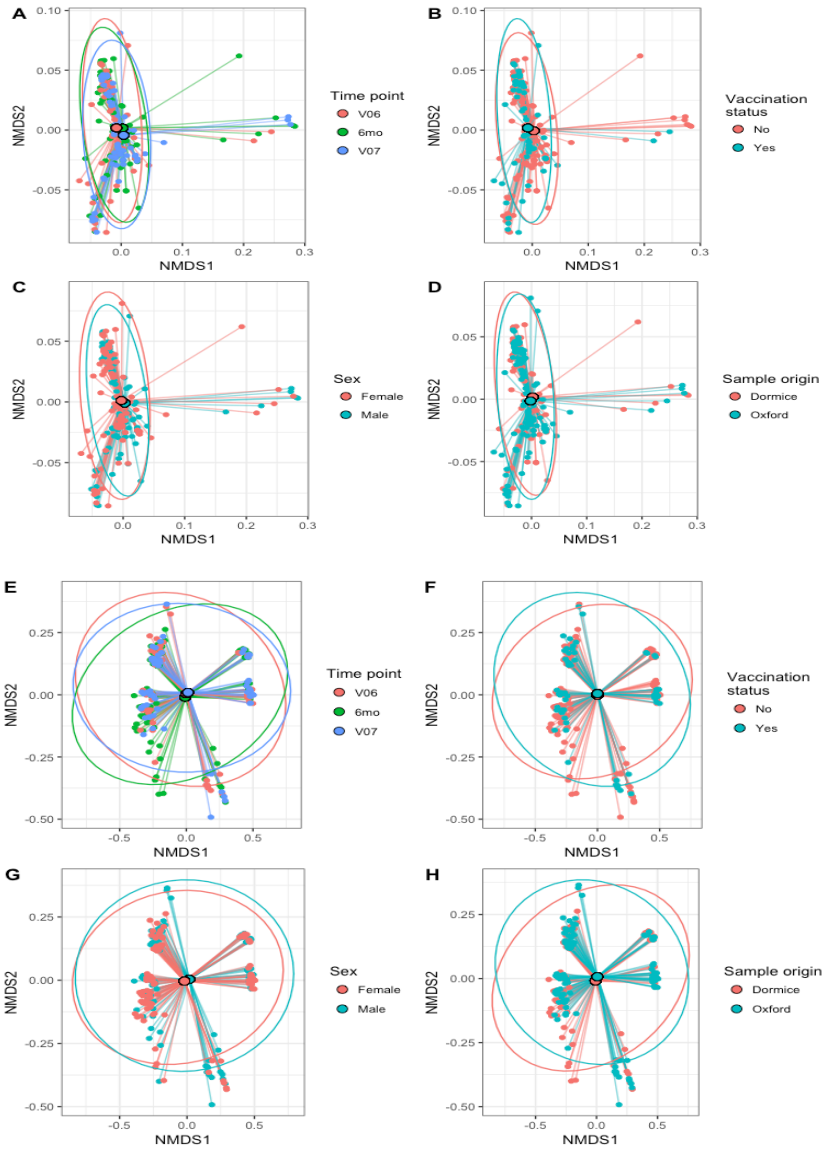


Figure 73 NMDS ordinations using earlier samples and created using the weighted UniFrac (A-D) and UniFrac (E-H) distances. Samples were coloured according to the sample's time point (A, E), vaccination status (B, F) sex (C,G) and origin (D,H).

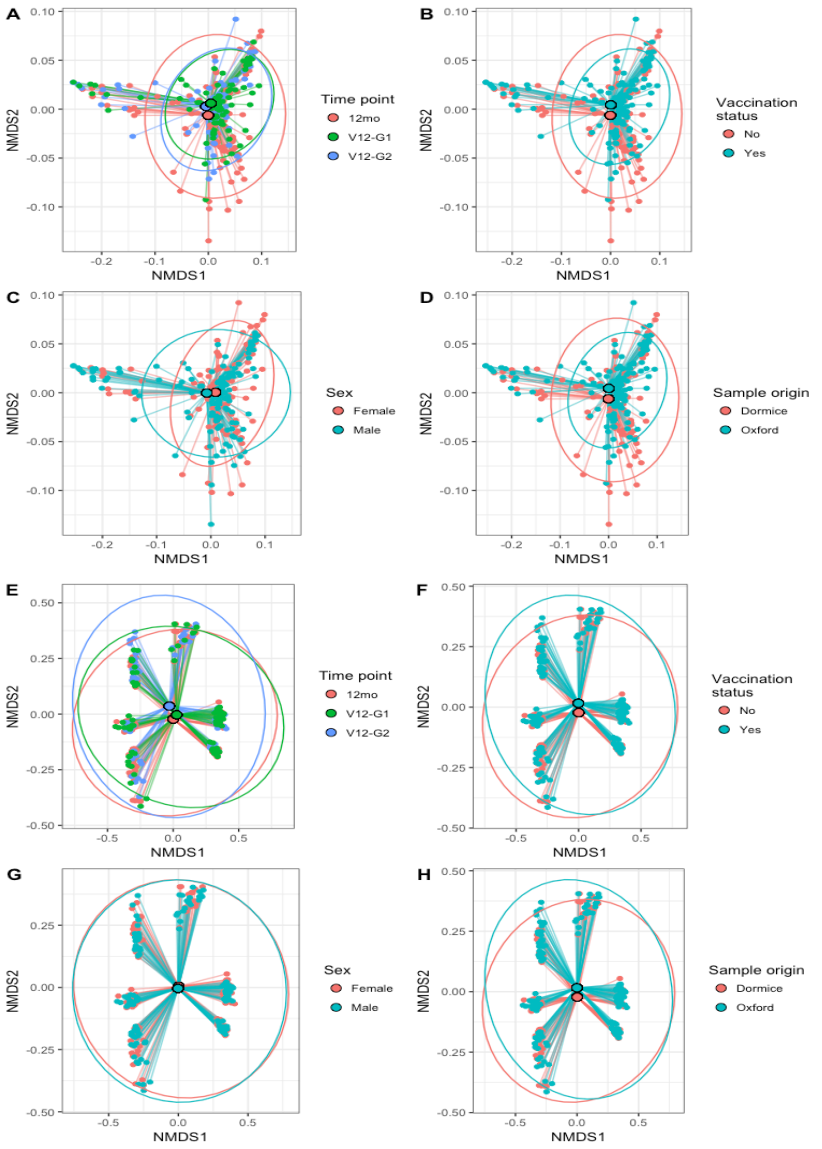


Figure 74 NMDS ordinations using later samples and created using the weighted UniFrac (A-D) and UniFrac (E-H) distances. Samples were coloured according to the sample's time point (A, E), vaccination status (B, F) sex (C, G) and origin (D, H).

7.4.6.2. PCoA ordinations

PCoA ordinations were also generated to identify potential effects of vaccination in neisserial populations. Similar to the NMDS ordinations, the four aforementioned metrics were tested initially. Although there is no stress metric in PCoA analyses that provides a measurement of the goodness of fit of the ordinations, one can plot the proportion of variation explained by each axis in the ordination. The higher proportion of variation that is explained by the first two or three axes, the more the ordination will reflect the variation in the dataset.

The proportion of variation for PCoA ordinations created using UniFrac, Jaccard and Bray-Curtis metrics are shown in Figure 75. In these plots, each bar represents one axis on variation in the dataset, and its height the proportion of variation it explains, out of a total variation of one. The two major axes of variation in the ordinations calculated using the Bray-Curtis and Jaccard indexes accounted for less than 0.082 of the variability, while the two most prominent axes in the UniFrac distances accounted for 0.4 and 0.31 in the earlier time points. Similar results were obtained for the ordinations obtained using samples from later time points. Therefore, only the UniFrac metrics were used in this analysis.

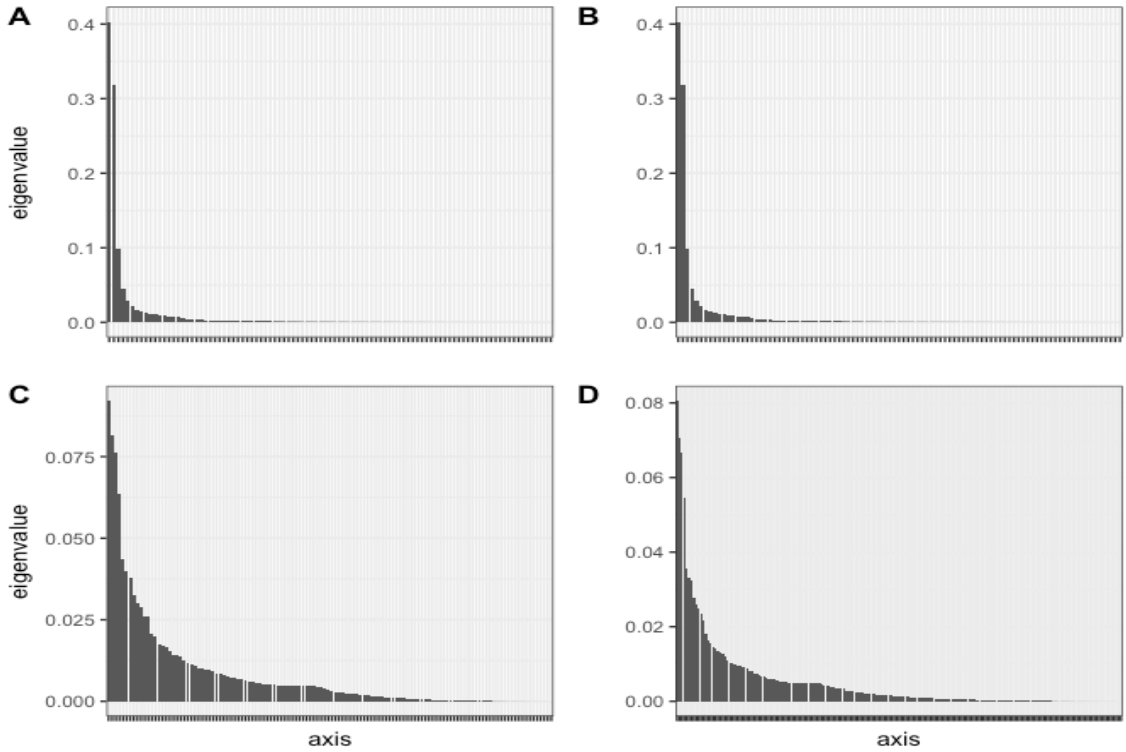


Figure 75 Proportion of the total variation (eigenvalues) explained by each axis in the PCoA ordinations calculated using weighted UniFrac (A), UniFrac (B), Bray-Curtis (C), and Jaccard (D) metrics.

The results obtained using the PCoA analysis were similar to those obtained with the NMDS ordinations: almost no differences were observed in earlier time points while slight differences could be observed in later time points. These differences, although small, may be of great importance in relation to the *Neisseria* population. To further quantify these differences, I proceeded to analyse the data using constrained ordinations.

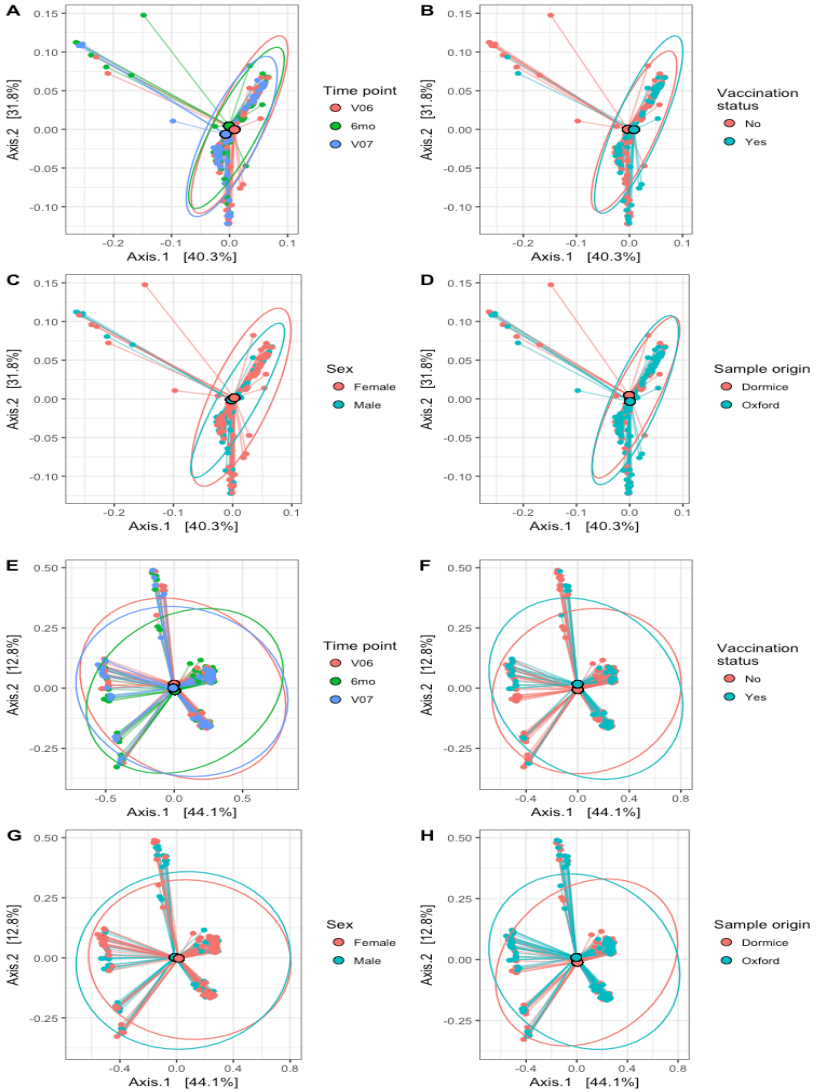


Figure 76 PCoA ordinations using earlier samples and created using the weighted UniFrac (A-D) and UniFrac (E-H) distances. Samples were coloured according to the sample's time point (A, E), vaccination status (B, F) sex (C,G) and origin (D,H).

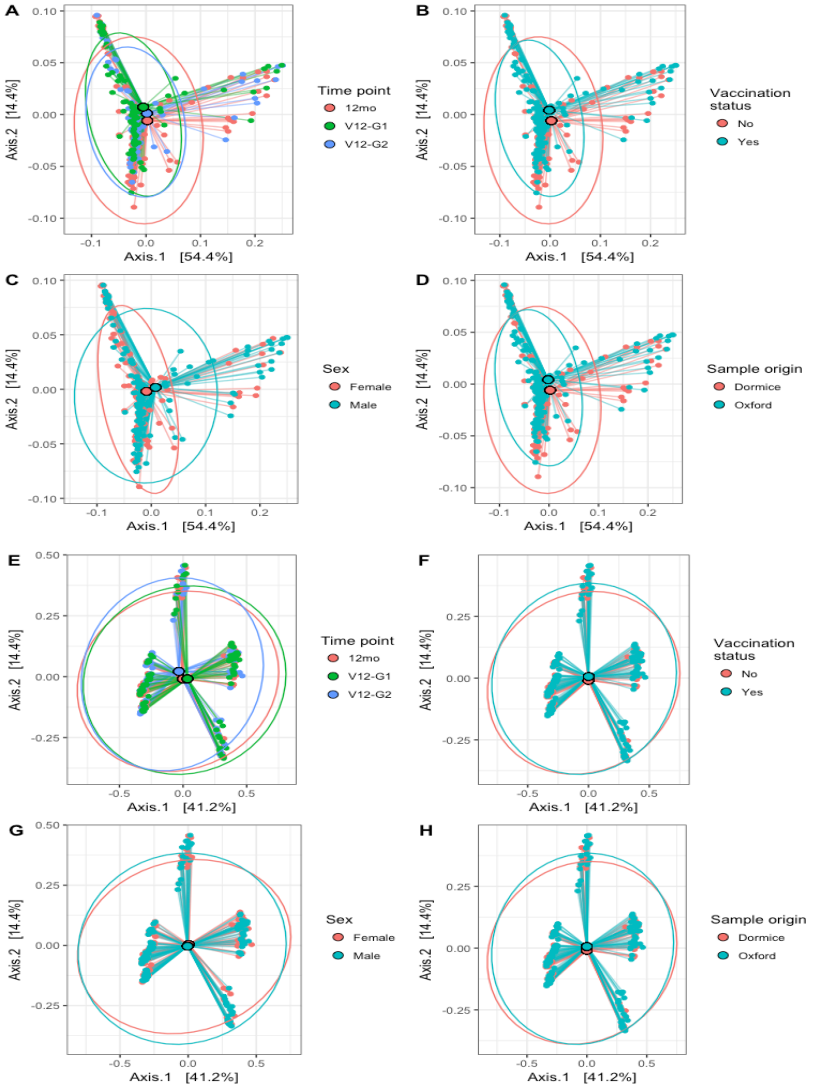


Figure 77 PCoA ordinations using earlier samples and created using the weighted UniFrac (A-D) and UniFrac (E-H) distances. Samples were coloured according to the sample's time point (A, E), vaccination status (B, F) sex (C,G) and origin (D,H).

7.4.6.3. Constrained ordinations

To quantify the effect of the variables being tested I used PERMANOVA, dbRDA and CCA methods. Brief explanations of these methods are given in Chapter 6 (Section 6.4.7). As with previous analyses, the differences were calculated using weighted and unweighted UniFrac metrics. Neither metric showed significant differences when comparing inter-group variances using the multivariate Levene's test. No significant differences were identified when using the unweighted UniFrac metrics in any of the variables. The weighted metric (Table 28) showed no significant differences in earlier time points. Twelve-month samples showed significant differences only when tested using a CCA approach when comparing the two Oxford sample time points to Dormice samples, and when comparing vaccinated and unvaccinated infants (which is identical to comparing samples by their origin at this time point, as all Oxford infants have received at least one dose of Bexsero®). However, all the significant differences accounted for a very low percentage of the variation in the data, with the highest value obtained, when comparing all the time points, accounting for 0.153% of the total variation. Factors used within PERMANOVA and dbRDA analyses, although not significant, accounted for slightly more of the variation. However, the largest amount, again achieved by the comparison of all time points, was 0.618%. The results of the dbRDA ordinations in when using the later samples is shown in Figure 78.

Table 28 P-values and R^2 values obtained for the PERMANOVA, dbRDA and CCA constrained ordinations using the weighted UniFrac distances. P-values below 0.05 are highlighted in red.

Group	*Variable*	*Variance*	*Inter-group variances*	*PERMANOVA R^2*	*PERMANOVA p-value*	*dbRDA R^2*	*dbRDA p-value*	*CCA R^2*	*CCA p-value*
Early time points *Dormice: 6mo* *Oxford: V06, V07*	Time point	6mo-V06	0.75006	0.00435	0.576	0.00102	0.241	0.00055	0.189
		6mo-V07	0.90959	0.00511	0.487	0.00069	0.649	0.00034	0.308
		V06-V07	0.96039	0.00932	0.335	0.00092	0.603	0.00023	0.406
		All time points		0.00781	0.53	0.00013	0.468	0.00027	0.341
	Vaccination status	No - Yes	0.47686	0.00394	0.446	0.00029	0.36	0.00019	0.633
	Sex	Female - Male	0.65439	0.00268	0.662	0.00033	0.578	0.00011	0.577
	Sample origin	Dormice - Oxford	0.57257	0.00276	0.633	0.00044	0.31	0.00057	0.11
Later time points *Dormice: 12mo* *Oxford: V12-G1, V12-G2)*	Time point	12mo - V12-G1	0.26996	0.00499	0.384	0.00015	0.41	0.00109	0.028
		12mo - V12-G2	0.47248	0.00451	0.427	0.00038	0.321	0.00153	0.005
		V12-G1 - V12-G2	0.93605	0.00413	0.558	0.00202	0.137	0.00102	0.082
		All time points		0.00618	0.511	0.00095	0.241	0.00153	0.002
	Vaccination status	No - Yes	0.47686	0.00394	0.446	0.00015	0.505	0.00120	0.002
	Sex	Female - Male	0.65439	0.00268	0.662	0.00021	0.52	0.00017	0.294
	Sample origin	Dormice - Oxford	0.57257	0.00276	0.633	0.00015	0.507	0.00119	0.002

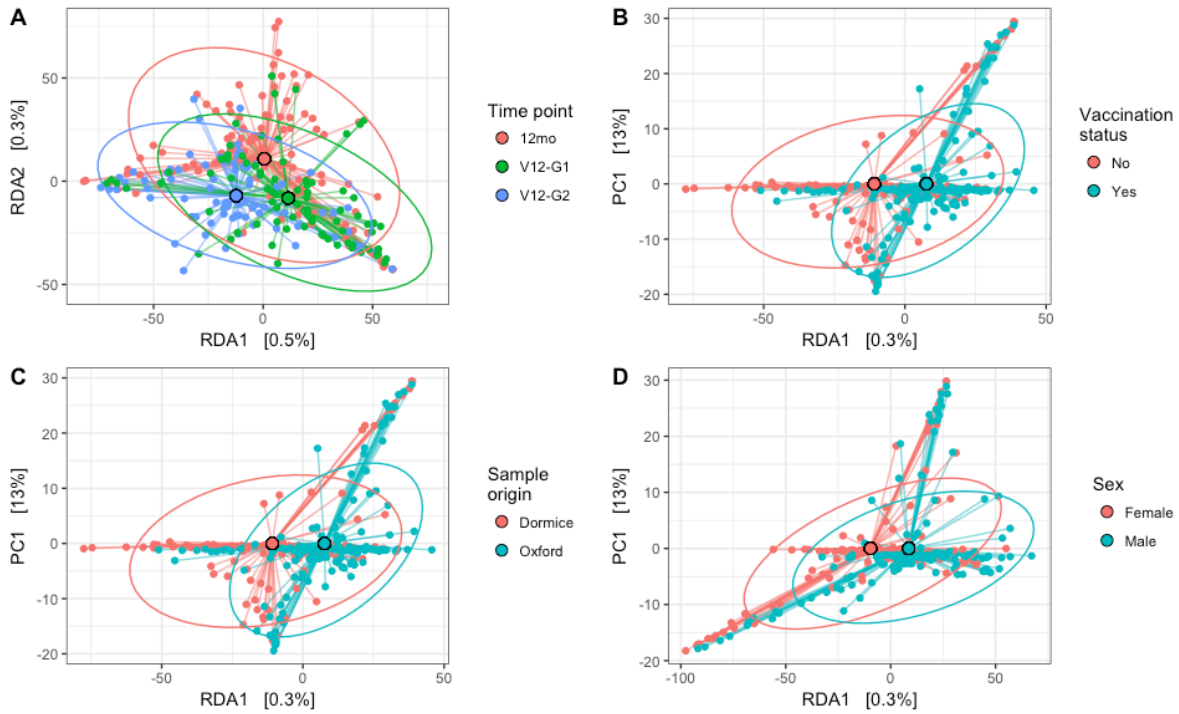


Figure 78 Constrained ordinations created using the dbRDA algorithm using the weighted UniFrac metric when comparing time points (A), vaccination status (B), Sample origin (C) and sex (D) variables.

7.4.6.4. Species-based ordination

Ordination methods can also be performed to identify relationships among taxa. While sample-based ordinations try to find samples that are similar in composition, species-based ordination aims to identify taxa that appear together across samples. Figure 79 shows the results of sample-based (A, C) and species-based (B, D) ordinations for earlier and later samples, respectively.

While no visible segregation was apparent when considering samples, some trends are apparent when *Neisseria* species are ordinated. A central cluster formed by the majority of species can be observed, as well as two other clusters, mostly formed by *N. lactamica* and *N. polysaccharea* SVs, and this separation appears to be more pronounced over time.

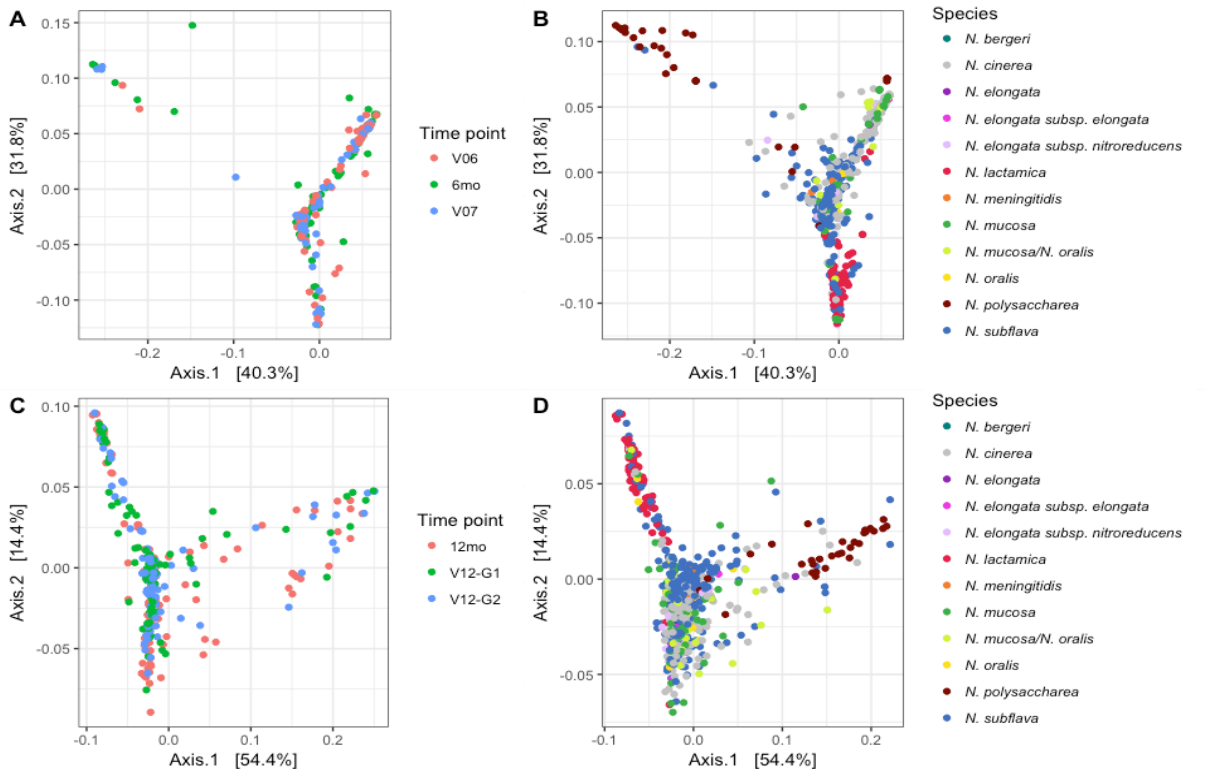


Figure 79 PCoA ordination obtained using 6-month (A, B) and later (C, D) samples, using the weighted UniFrac distance Plots A and C are sample-based plots while B and D focus on species.

7.4.6.5. Differential abundance testing

Although no significant changes in *Neisseria* species distribution were observed with the analyses described above, it was still considered that small differences may reflect crucial changes in the *Neisseria* population. As described in Chapter 6 (Section 6.4.8), metagenomeSeq (163) and LEfSe (164) were used to identify taxa that are significantly increased depending on sample groupings, as they provide different methods to assess these changes. One of the advantages of these methods is that, instead of relying on rarefied data, normalized raw data can be used as input increasing the number of samples included in the analysis, and increasing the number of reads per taxa. However, all the taxa with an abundance lower than 200 reads were excluded from the analysis. The four variables used until now were also included in this analysis.

7.4.6.5.1. Differential abundance testing using metagenomeSeq

The analysis carried out using the metagenomeSeq approach yielded very few significant differences, none of which appeared in earlier time point comparisons. The differences identified are summarized in Table 29. Only three SVs were identified as significantly different, one belonging to *N. mucosa* or *N. oralis* and another one to *N. subflava* when comparing vaccination status and sample origin in later time points. The third, another *N. mucosa* or *N. oralis* SVs, was identified when comparing both Oxford time points. No significant differences were found when comparing reads at the species level.

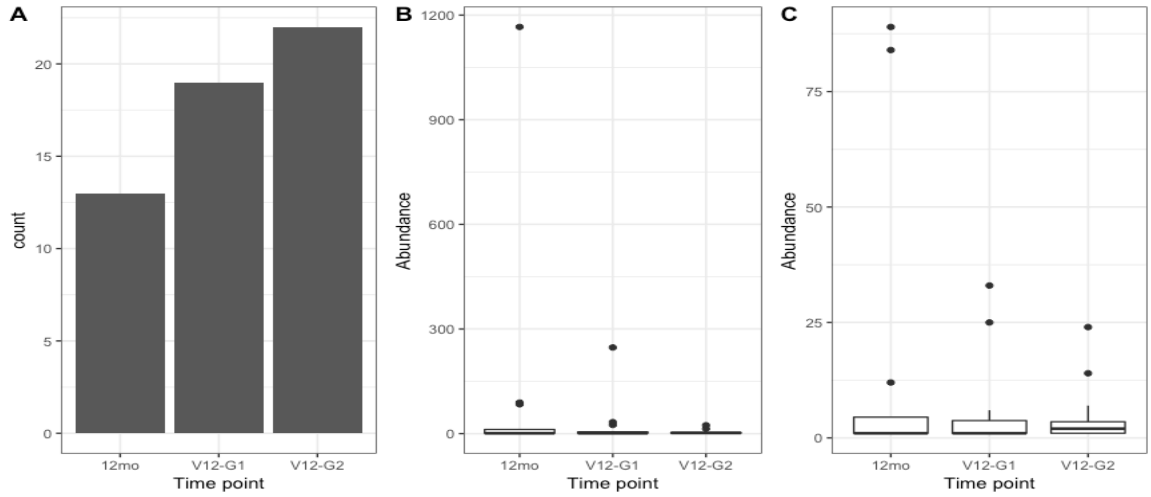


Figure 80 Carriage of *N. meningitidis* in later time point samples (Dormice: 12mo, Oxford: V12-G1, V12-G2). Figure A shows the total number (count) of infants colonized per time point. Figure B shows boxplots describing number reads identified for this bacterium, while C focuses on those samples with fewer than than 150 reads.

Table 29 Sequence variants identified as significantly different when compared to its corresponding group. The columns show, in order, the comparison in which the SV were found to be significant, the SV involved, the number of samples in each group carrying that SV, the total number of reads per group, the odds ratio, log2 fold change and standard error, and the adjusted p-values

Comparison	*SV*	*Group 0 (n)*	*Group 1 (n)*	*Group 0 (counts)*	*Group 1 (counts)*	*Odds ratio*	*Log FC*	*se*	*P-value*
Sample origin *Group 0: Dormice* *Group 1: Oxford*	*N. mucosa/ N. oralis-538*	21	20	17500	634	1.583	-3.539	0.781	0.004
	N. subflava-343	38	79	2759	42270	0.525	1.991	0.496	0.019
Vaccination status *Group 0: No* *Group 1: Yes*	*N. mucosa/ N. oralis-538*	21	20	17500	634	1.583	-3.539	0.781	0.004
	N. subflava-343	38	79	2759	42270	0.525	1.991	0.496	0.019
Time point *Group 0: V12-G1* *Group 1: V12-G2*	*N. mucosa/ N. oralis-535*	15	8	25	6855	2.032	3.680	0.903	0.027

7.4.6.5.1. Differential abundance testing using LEfSe

The LEfSe method identified several SVs that differed among the groups being compared, however, this method also identified differences at the species level. For ease of interpretation of the results, I will focus on those obtained when only considering changes at the species level.

When comparing groups using earlier samples only *N. elongata* subspecies *nitroreduscens* was significantly different, appearing exclusively in Dormice samples. However, later time points showed more diverse differences. Individually, Oxford samples were characterized by increased LDA scores in *N. lactamica* and *N. subflava* reads for the V12-G1 and V12-G2 groups, respectively. Conversely, Dormice samples were characterized by an increase in *N. elongata* subspecies *nitroreducens*, *N. cinerea* and *N. meningitidis.* These differences were also observed when comparing vaccination status

(and sample origin) for each group, plus an increase in *N. bergeri* and *N. mucosa* or *N. oralis* reads in Dormice (Figure 81). Seeing *N. meningitidis* appears as a differential factor between vaccinated and unvaccinated infants could be of great significance. To further characterize this difference, I selected only those infants that were colonized by this species (Figure 80). Although Dormice samples had a lower number of infants colonized by this bacterium, some of the infants in this groups showed a higher number of *N. meningitidis* reads. Although this result could be relevant, it is important to notice the small number of samples and larger variation in read counts, makes this a difficult comparison.

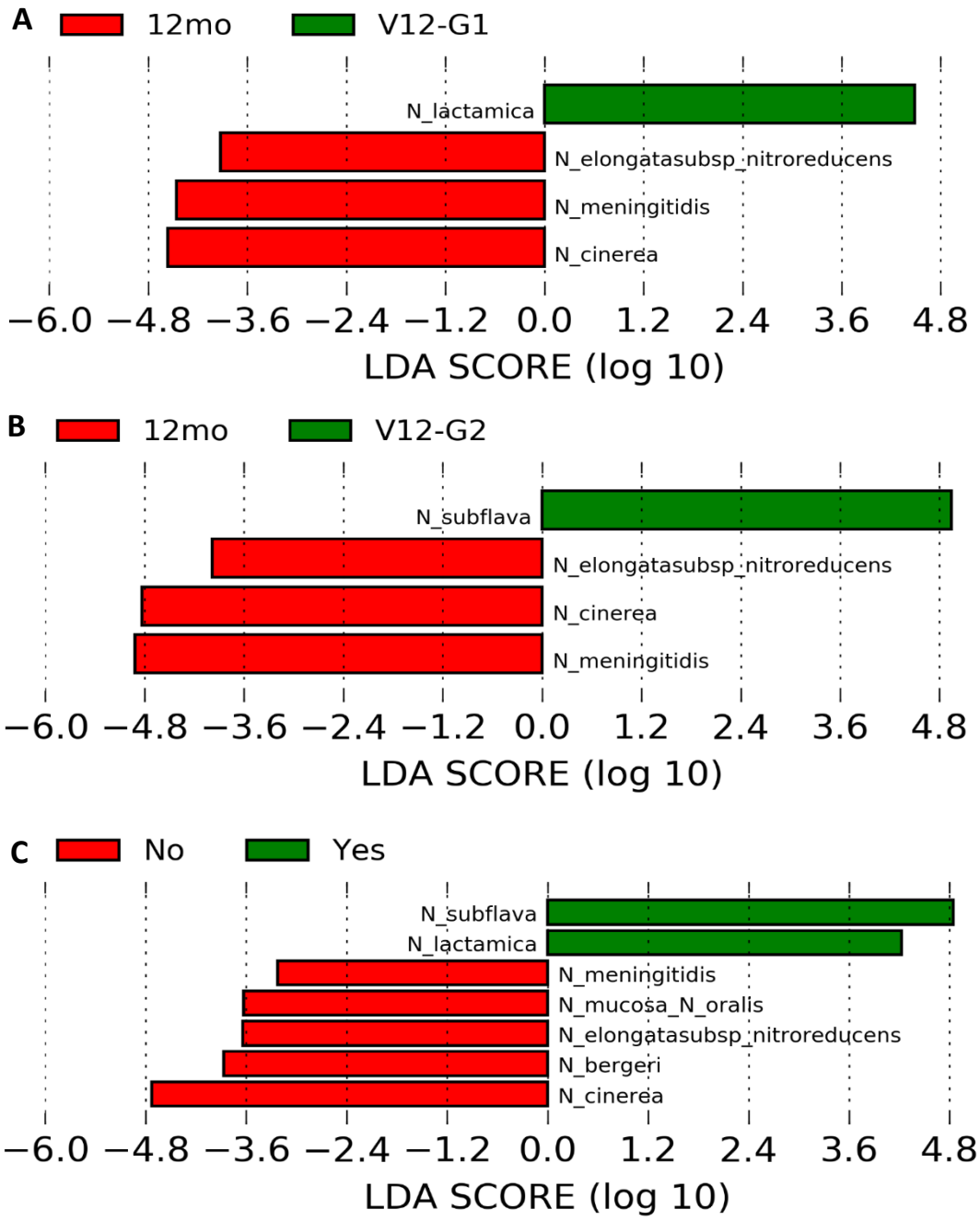


Figure 81 Results obtained when using the LEfSe algorithm when using samples collected at the 12-month time point. The results show the comparison, at the species level, of Dormice samples and the two Oxford sample time points (A, B), samples compared by vaccination status. Larger LDA score values represent a stronger association of the species with a group, according the legend. For each plot, bars are coloured according to the legend at the top.

7.4.6.5.2. Changes observed within sample pairs

The analyses performed so far had relied on comparing groups as a whole, rather than focusing on individuals and comparing earlier and later samples. For this analysis, I used rarefied data, to ensure comparisons among equally sized samples. To obtain a clearer result as to any effect of vaccination comparison of abundance levels of *N. meningitidis* was determined. However, it should be acknowledged that due to the low number of infants colonized by this bacterium such comparisons may be indicative and not completely robust. *N. cinerea, N. polysaccharea* and *N. lactamica* will be used as proxies for *N. meningitidis* as Muzzi *et al*. (45) showed that these species shared a higher degree of antigen conservation in the NHBA antigen, *N. cinerea* also showing similarities in the fHbp variant included in Bexsero®. I will also include *N. subflava* as a control, as its abundance is not expected to change with vaccination. The results of this analysis are shown in Figure 82.

None of the abundance distributions of the species analysed followed a normal distribution, so Wilcoxson signed-rank test, Fisher's exact test, and Chi-squared test were used as appropriate. *N. cinerea* showed a significant difference in abundance when comparing values in Dormice and Oxford samples, however there was no significant difference in terms of a higher proportion of samples increasing or decreasing in abundance over time. *N. polysaccharea* showed no significant difference, neither when considering its abundance or changes over time. *N. subflava* showed a significant difference in terms of abundance in earlier time points, but also showed no significant differences in terms of change. Finally, *N. lactamica* had a slight increase in abundance in later time points, although not significant. This species also showed no significant differences in variation over time.

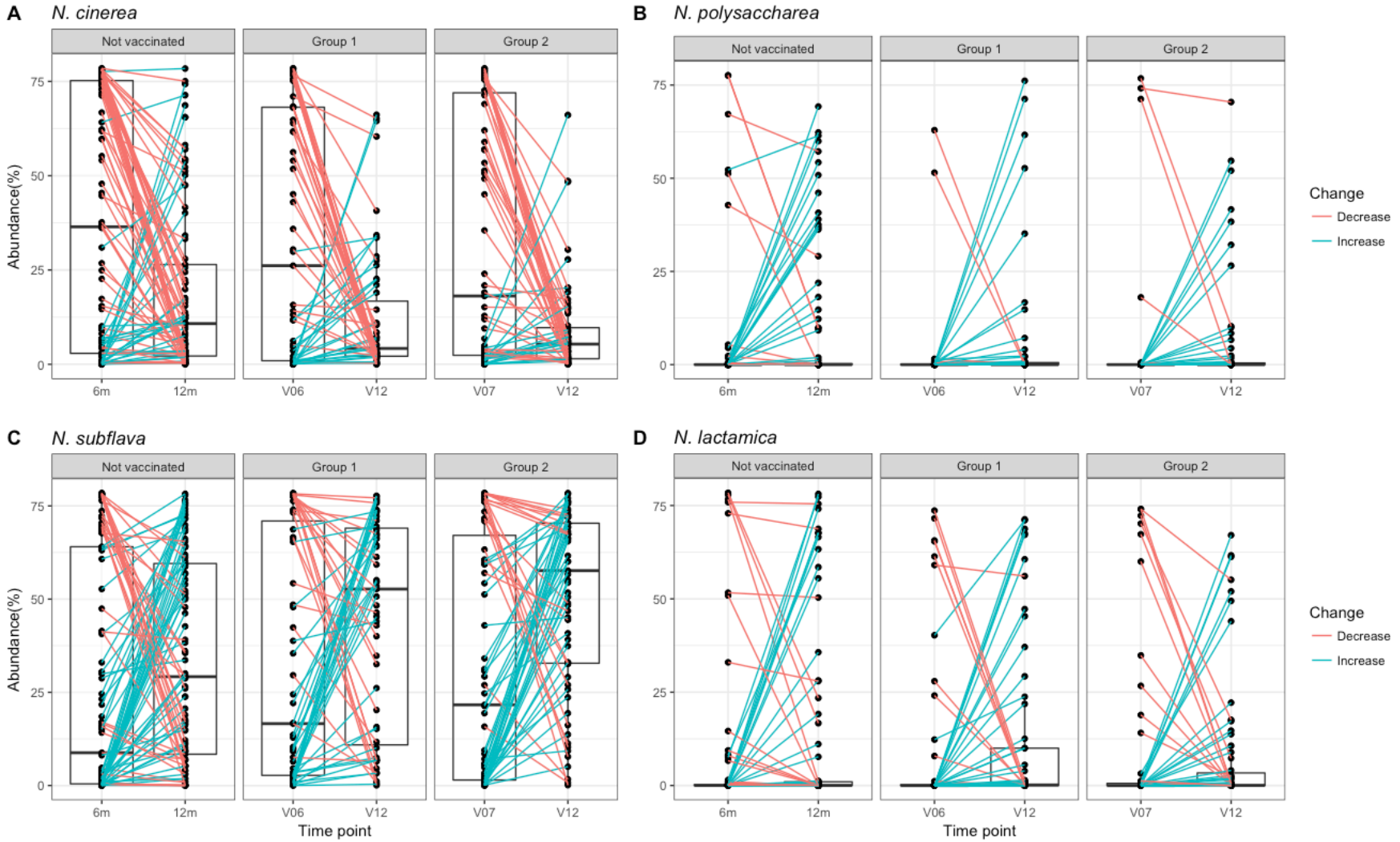


Figure 82 Variation within sample pairs observed for Dormice samples (Not vaccinated) and the two Oxford sample groups (Group 1, Group 2) when comparing *N. cinerea* (A), *N polysaccharea* (B), *N. subflava* (C), and *N. lactamica* (D) abundances. For each plot, dots represent the abundance in each sample, and samples from the same infant are connected by a coloured line (red if decreases, blue If it increases).

7.4.6.6. Quantification of the neisserial burden in infants

To quantify the total amount of *Neisseria* in the throat swab samples, a qPCR assay was designed for this purpose, using the *rplF* gene as a target. As each neisserial genome only has one copy of the *rplF* gene it can be used as a proxy for the total number of *Neisseria* present in samples. The results displayed on Figure 83 show boxplots of the quantity of *Neisseria* calculated, using a log10 of the total amount of *rplF* copies estimated, for all throat swab taken at each time point. No significant changes were identified over time; however, the measurements showed a large variance.

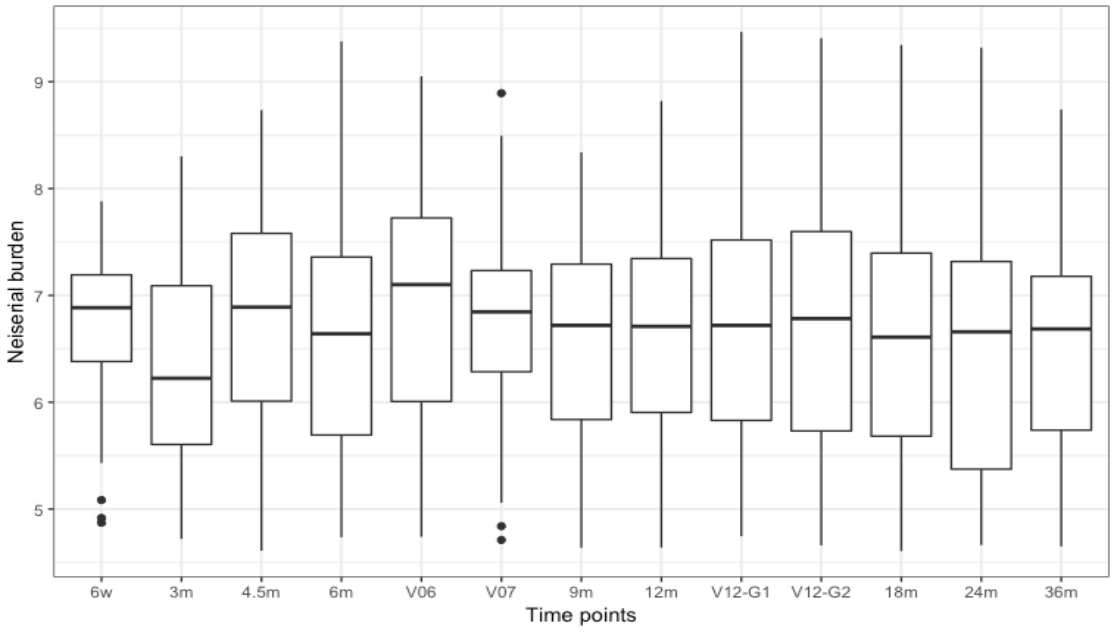


Figure 83 Total neisserial burden measured per time point using a qPCR approach described using bloxplots. Neisserial burden is measured as the log10 of the estimated total number of copies of the *rplF* gene per swab.

Coupling the neisserial burden data with the proportion of each species obtained from the sequencing data, we can obtain the total amount of species present in the dataset. With this new data, I compared the mean absolute values of each species per time point in the 6-month and 12-

month time points for Dormice and Oxford, while also considering the difference in number of samples per time point. The results are shown in Figure 84.

Earlier Dormice samples showed a higher average burden of *N. cinerea* compared to Oxford samples, while other species such as *N. subflava* remained more comparable levels. *N. lactamica* had a noticeable increase in the V06 Oxford time point, coupled with almost complete disappearance of *N. polysaccharea.* Both Oxford later time points showed an increase in mean neisserial burden, V12-G1 being characterized by a surge in *N. subflava,* and a *N. lactamica* expansion in V12-G2, similar to the one seen in V06 previously. Furthermore, Dormice samples had a decrease in total *Neisseria* counts.

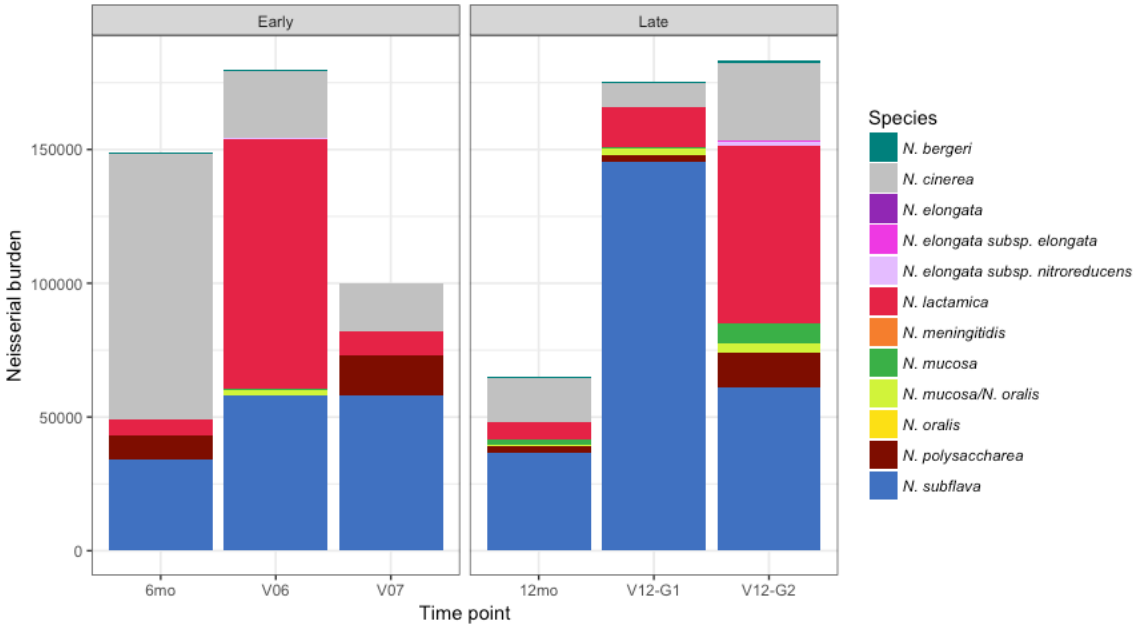


Figure 84 Mean absolute species abundance per time point in earlier (6mo, V06, V07) and later (12mo, V12-G1, V12-G2) time points. The Y axis shows the mean abundance of all the samples in each time point in terms of number of *Neisseria* per swab. Reads were normalized to account for the different number of samples in each time point

7.5. Discussion

The analysis of *rplF* sequencing data comparing vaccinated and unvaccinated infants at similar time points revealed very few differences between the groups, most of these appearing around the 12-month time point. While comparisons using ordination methods yielded minute differences, differential abundance methods, especially LEfSe, hinted at an irregular species distribution that could be caused by vaccination. Pairwise comparisons of samples, considering the change of particular species over time, showed no significant differences. Finally, qPCR data coupled with the sequencing data showed a potential side-effect of vaccination in infants.

The required quality control analyses showed acceptable results. The comparison of the sequencing data using the modified and unmodified Forward primers revealed no significant difference between the data obtained and the expected values, but visual inspection of the results showed a better fit of the results when using the modified primer. A small number of samples used contained low amounts of reads of species that should have not been present in the mixture when tested using either primer. Although the quality of the sequencing experiment was satisfactory, it was lower than that of other libraries prepared using human samples, possibly due to the low sequence diversity in these runs. It could be possible that due to the low quality, some reads could have been assigned a species erroneously. However, given that the errors obtained when sequencing samples mirrors those of the results obtained with the standard primer, it is likely that these are technical errors. Sim *et al*. (92) showed that one base pair difference in primers, when amplifying the bacterial 16S rRNA gene from faecal samples, could lead to the loss of detection of *Bifidobacterium*. The results seen here do not show such a drastic outcome, as all species were identified, but a slight over-representation of *N. meningitidis* and *N. polysaccharea* could be observed. This was expected, as these

two species bind the unmodified Forward primer perfectly. The modified primer overcame this effect, showing results more similar to the ones expected.

The α-diversity values observed seemed to follow the trend observed with the 16S rRNA data analysis. Although an increase in diversity over time was identified, the changes did not seem to be gradual as seen during the microbiome analysis Chapter 6 (Section 6.4.4). The diversity metrics formed clusters of time points, one formed by those up to, and including, the 6-month time point and the other one formed by all the other time points. This could suggest that events in between the 6 and 9 months of life could drive an increase in the diversity of *Neisseria* species in the infant upper respiratory tract.

Looking at the relative abundance of species per time point provides an overview of the development of *Neisseria* colonization over time. While most of the efforts made to characterize the neisserial colonization in infants have been focused on *N. meningitidis* and *N. lactamica,* very little is known about other neisserial commensals. The data shown here shows that a large proportion of the *Neisseria* population in the upper respiratory tract in infants is formed of other colonizers, such as *N. cinerea* and *N. subflava*, and to a lesser extent *N. mucosa, N. oralis, N. elongata* and *N. bergeri.* Several publications (9,172) have suggested that *N. lactamica* can protect against *N. meningitidis* colonization. *N. cinerea,* which was found to be more abundant earlier on in life, shares antigenic similarities with *N. meningitidis* (52) and can express antigens present in Bexsero® (45). Future research could focus on this species as another source of anti-meningococcal immunity.

Looking at the proportion of infants colonized by individual species, the two most prominent were *N. subflava* and *N. cinerea,* these being present in almost all infants. For the remaining species, there seemed to be two patterns of development. *N. lactamica, N. polysaccharea, N. bergeri* and *N. meningitidis* were relatively stable over time, while *N. mucosa, N. oralis* and *N. elongata* were at a relatively stable level earlier in life, increasing past the 9-month time point, following the trend observed when measuring α-diversity values. However, *N. meningitidis* was only identified in very few samples, so it is difficult to identify robust correlations regarding this species. Considering *N. lactamica* carriage, most studies focusing on its carriage in infants have estimated a proportion in between 20 and 40% of infants being colonized by this bacterium (118,173), while our data suggests a slightly higher value, ranging from 50 to almost 75% of infants being colonized depending on the time point. It is possible that carriage of this species has been underestimated in previous studies which have relied on culture-based analysis.

Ordination methods did not show as many differences as those seen when comparing 16S rRNA data. However, a similar trend was identified as most differences were more pronounced in later time points. Both NMDS and PCoA analyses showed very similar results, and the use of constrained ordinations did not provide a sizable correlation in between vaccination and changes in the *Neisseria* population. The only significant results obtained were those from CCA analysis in later time points, but only accounted for less than 0.15% of the total variation.

Nevertheless, species-based ordinations showed a potential insight into *Neisseria* population dynamics. The central core of taxa was compromised by most species, while two clusters seemed to separate from it, with this separation being more pronounced over time. The two main species in these clusters were *N. lactamica* and *N. polysaccharea*. These results could mirror those obtained

from live inoculation challenges (9,10) where *N. lactamica* carriage inhibited that of *N. meningitidis.* Although purely speculative, these two species could create a less welcoming environment for other *Neisseria*, be it by cross-reactive antibodies generated by the host while colonized, or by overtaking the *Neisseria* niche temporarily, impeding another *Neisseria* species colonization of the upper respiratory tract.

Differential abundance testing showed some potentially relevant changes, particularly for the analysis carried out using LEfSe. The data showed an increase in *N. lactamica* in Oxford samples, and an increase of *N. meningitidis* in Dormice samples. Although, as mentioned earlier, the comparison of *N. meningitidis* is difficult due to the low number of samples and the large variance in read counts, it could point towards a decrease, not necessarily in colonization, but maybe in the amount of *N. meningitidis* colonizing the upper respiratory tract. However, it is difficult to assess this difference as being exclusively due to vaccination, as there are other factors that could contribute to meningococcal carriage (174).

When comparing variation relative abundance data at both time points, no significant differences were found. However, combining qPCR and sequencing data was potentially more informative. Regarding total burden, Dormice samples seemed to have a higher amount of *N. cinerea* at earlier time points compared to Oxford, which then decreased over time. Oxford Group 1 samples (V06 and V12-G1) started at the highest levels, and stayed reasonably constant over time, while Oxford Group 2 samples (V07 and V12-G2) started the lowest and overtook all other time points. A possible reason for the latter result is as follows: Vaccinated samples have a higher neisserial burden, but two of these time points, V06 and V12-G2, had a much larger proportion of *N. lactamica.* These two groups had received a vaccine relatively close to the sample collection date (2 months for V06

and 4 months to V12-G2), so it could be possible that this change in neisserial burden is a potentially beneficial side effect of the vaccination. A hypothesis could be that Bexsero® vaccination disrupts the neisserial population in such a way that it is able to proliferate more than normal, and somehow facilitates the growth of *N. lactamica*, that could also help with the prevention of *N. meningitidis* colonization. Biesbroek *et al.* (161) showed a similar effect in their study assessing the changes in the microbiome due to vaccination with *S. pneumoniae* PCV-7 vaccine, in which vaccinated infants saw an increase of overall bacterial burden in parallel to an increase in streptococcal burden. However, as stated before, it is difficult to determine if the changes in the neisserial microbiome observed are exclusively due to vaccination rather than a secondary effect due to other life events.

8. Chapter 8: Effect of early life events on *Neisseria* species

8.1. Introduction

Identification of factors affecting *Neisseria* carriage have been focused on two species: *N. lactamica* and *N. meningitidis.* However, several *Neisseria* species, such as *N. cinerea* and *N. polysaccharea* have been described, based on genome analyses, as being more similar to *N. meningitidis* than *N. lactamica* (52), and also contain genes encoding several of the antigens present in Bexsero® (45). It could therefore be hypothesized that being colonization by these species may have a protective effect against *N. meningitidis*, similar to that suggested for prior *N. lactamica* colonization (9).

Several carriage studies have investigated risk factors that alter the likelihood of carrying *N. meningitidis,* and the risk of suffering invasive meningococcal disease, or *N. lactamica*, as it has been proposed as a protective agent against meningococcal colonization (8). Although studies on these species is of great interest, very little is known about other *Neisseria* species, and their potential effects on human health and disease.

In this Chapter, I will present the results of the analysis of the effect on *Neisseria* species abundance and presence on one another. I will also discuss how several early life events may affect the carriage of certain *Neisseria* species.

8.2. Aims

- Determine possible interactions among *Neisseria* species.

- Determine if certain early life events affect the abundance of different *Neisseria* species.

8.3. Methods

8.3.1. Samples used

A subset of Dormice samples collected were used in this analysis. Depending on the analysis, a different number of samples were used. The total number of samples will be mentioned for each analysis.

8.3.2. Statistical analysis

Phyloseq (94) objects containing count data, metadata, and taxonomy tables were imported into metagenomSeq (163). The fitTimeSeries function was used to calculate intervals of time in which the abundance of a particular species varied when controlling for the variable specified. Fisher's exact test was used to calculate the odds ratios for presence and abundance of different species. All the p-values presented for the odds ratios were adjusted using the Bonferroni multiple testing correction method (100).

8.4. Results

8.4.1. Neisseria population in infants

The data presented in previous Chapters 6 and 7 have been focused on groups of infants at different time points. Although the averaged data for individual time points can provide insights into the dynamics of the neisserial population as a whole, it does not provide a clear picture of what could be expected to be found in an individual. To have a deeper understanding of the colonization in infants it is necessary to investigate how infants are colonized by different species, and if there are any driving forces behind the colonization of individual species.

To visualize the different patterns of neisserial colonization in infants, and how it changes over time, a subset of 70 randomly chosen infants was selected, and the proportion of reads per species and time point for each infant was plotted (Figure 85). There was a great diversity in the patterns of neisserial development among infants through the first three years of life. For most infants there was a dominating species at each time point that was accompanied by a number of other species in a lower abundance. As seen previously, almost all infants were colonized by *N. subflava* and *N. cinerea* at every time point, although the proportion of these two species was variable. In the previous analysis using ordination methods to investigate the relationships among *Neisseria* sequence variants, two clusters (of the three identified) with majorities of *N. lactamica* or *N. polysaccharea* were found. The data shows, when present, that these two species tend to become the most prevalent amongst *Neisseria.*

Despite the variation in *Neisseria* carriage and its variation among infants, I wanted to identify potential trends and influences that species may have over one another. To do this I focused on two different questions: Are some species more, or less, likely to appear in a given time point if an infant is also colonized by another? Secondly, do the odds of colonization for one species change based on the presence of other species on the previous time point? To identify significant differences, I used Fisher's exact test to calculate the odds ratio in between groups, which in this case will be infants either colonized or not by a given species. Odds ratios were calculated considering two factors: first the abundance of each species, and then whether or not the species was present.

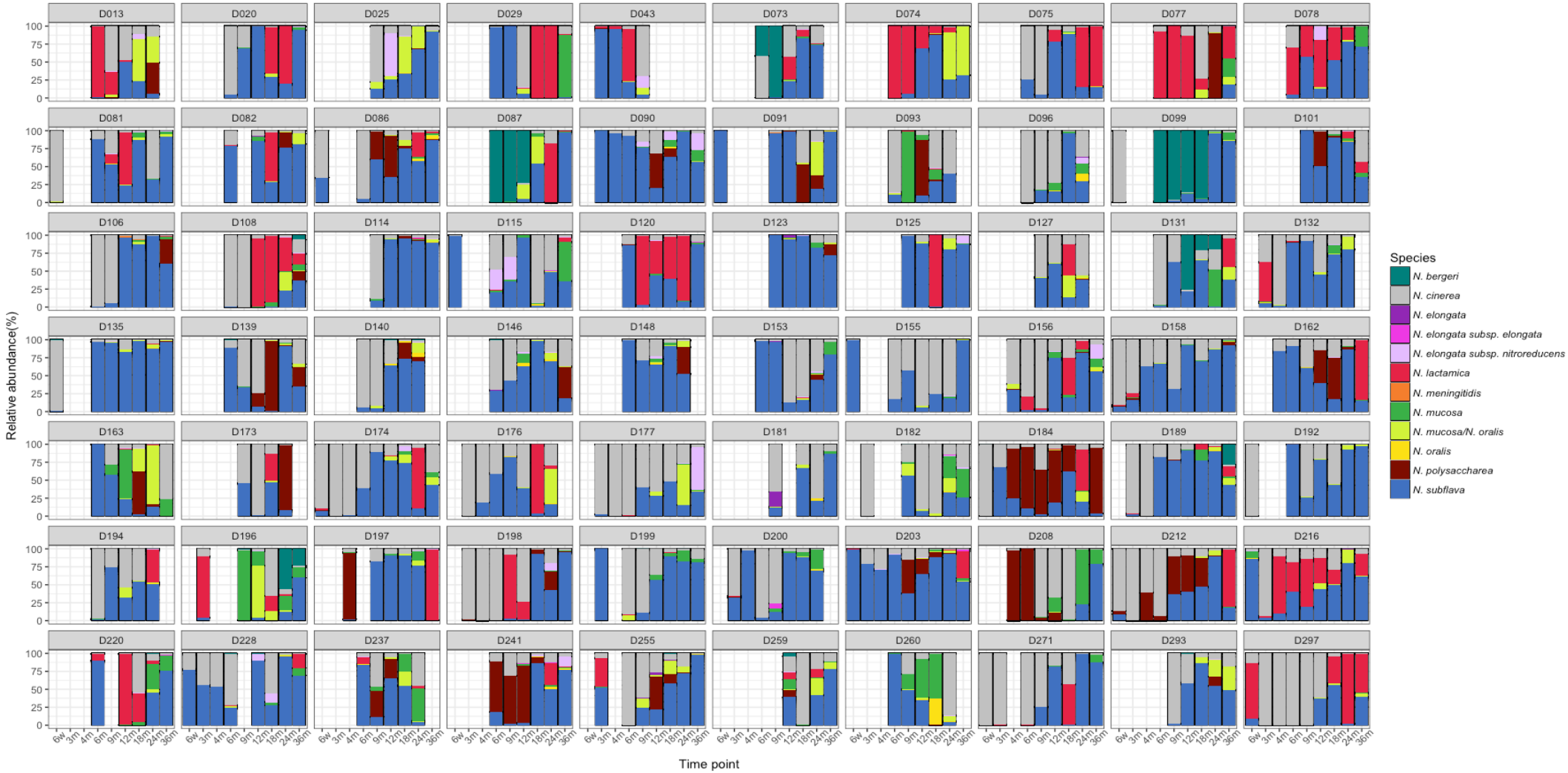


Figure 85 Relative abundances of each *Neisseria* species identified for 70 randomly selected infants at each sample collection time point.All available time points are displayed for each set of samples from a given infant. Empty columns represent samples that either could not be collected for a particular time point, or a sample that was collected but either no neisserial DNA could be identified or it did not meet the rarefaction threshold of 2000 reads.

For the following graphs (Figure 86, Figure 87), I plotted the odds ratio calculated for each species (as marked by the label in the X axis) when infants were colonized or not by the species being considered (denoted by the labels in the Y axis). Only those species that showed a significant increase or decrease are shown. The odds ratios of each species were calculated based on the presence or absence of other species, considering the group of infants not colonized as the reference group.

First, I calculated the odd ratio values for species when considering current co-colonization in samples (Figure 86). Similar trends were obtained when comparing the odds ratios of species based on their abundance and their presence, although more species were affected in terms of abundance. *N. mucosa, N. oralis* and *N. elongata* (and the *N. mucosa/N. oralis* group) had a positive effect on each other, increasing their abundance when one of the others was present. *N. cinerea* and *N. subflava* had their abundance and presence and abundance odds diminished by almost all other species, which is expected as these form the majority of the *Neisseria* population. A similar trend was observed with *N. lactamica*, whose abundance was also reduced by almost all other species. *N. meningitidis* had a higher odds ratio when *N. lactamica* and *N. polysaccharea* were present, and the presence of the meningococcus also seemed to increase the odds of *N. polysaccharea.* Also, the odds ratio of presence of *N. meningitidis* was reduced in those infants where *N. mucosa* and *N. oralis* were in higher abundance. *N. subflava* increased the likelihood of carrying *N. mucosa,* while decreasing that of *N. cinerea* and *N. bergeri*.

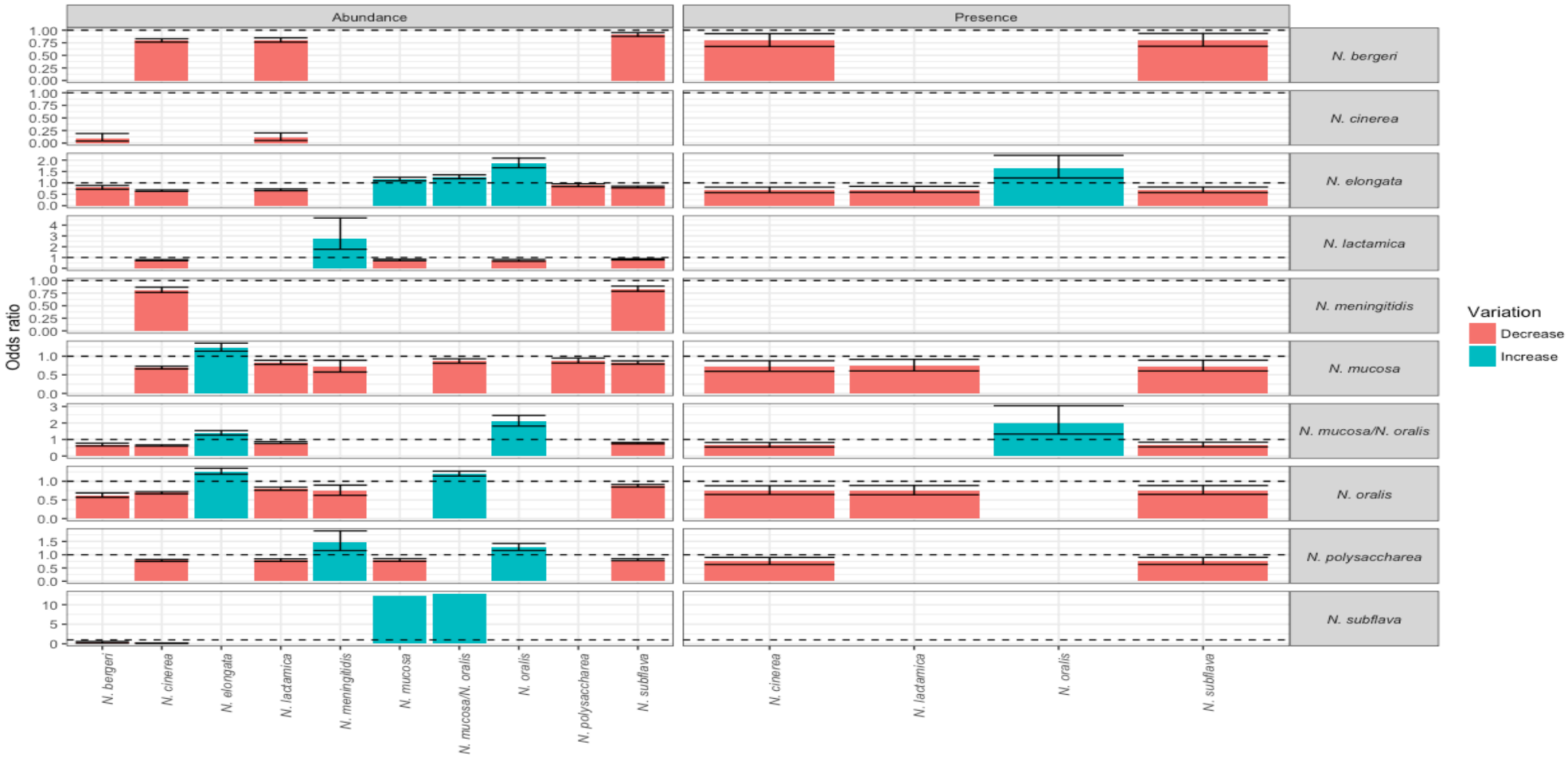


Figure 86 Odd ratios calculated for each species, marked in the X axis, when considering their abundance or presence, based on the presence the species marked in the Y axis labels. The reference group considered when calculating the odds was those colonized, hence, odd ratios above 1 indicate an increase in odds of having a higher abundance or presence when the species in the Y label is present. Comparisons of the effect of one species onto themselves were ignored for this analysis as they will always be correlated. The confidence interval bars are not shown for the *N. mucosa* and *N. mucosa/N. oralis* odds when considering the presence of *N. subflava* as they were too varied and obfuscated the visualization. The upper and lower confidence intervals were 2.258 and 512 for *N. mucosa,* 2.146 and 486 for *N. mucosa/N. oralis*.

Using the same data, I calculated the odds of each species based on the appearance on other species at the previous time point for each infant. As not all infants had sequencing data for the previous time point, fewer samples could be used for this analysis, as only those that that appeared in one time point and the previous one could be considered, for example, only 3-month samples that also had a positive 6-week sample were included, 4.5-month samples that had a positive 3-month sample etc. A total of 581 samples out of 780 were used for this analysis. None of the species showed significant differences in terms of odds based on presence of other *Neisseria* species.

When considering the data on previous time points (Figure 87), *N. meningitidis* was more likely to be less abundant in those infants that were colonized by *N. lactamica, N. polysaccharea* or *N. bergeri*, the biggest inhibition being that of *N. lactamica*, being 77% less likely to carry *N. meningitidis.* At the same time, *N. lactamica* had decreased odds of appearing when *N. oralis* (or those in the *N. mucosa/N. oralis* group) were present. *N. polysaccharea* was more likely to appear if *N. oralis* (and *N. mucosa/N. oralis*), *N. lactamica* or *N. elongata* were present in previous time points, all these increasing the odds in between 12 and 16%. The presence of *N. meningitidis* only had an effect in *N. mucosa*, increasing its odds.

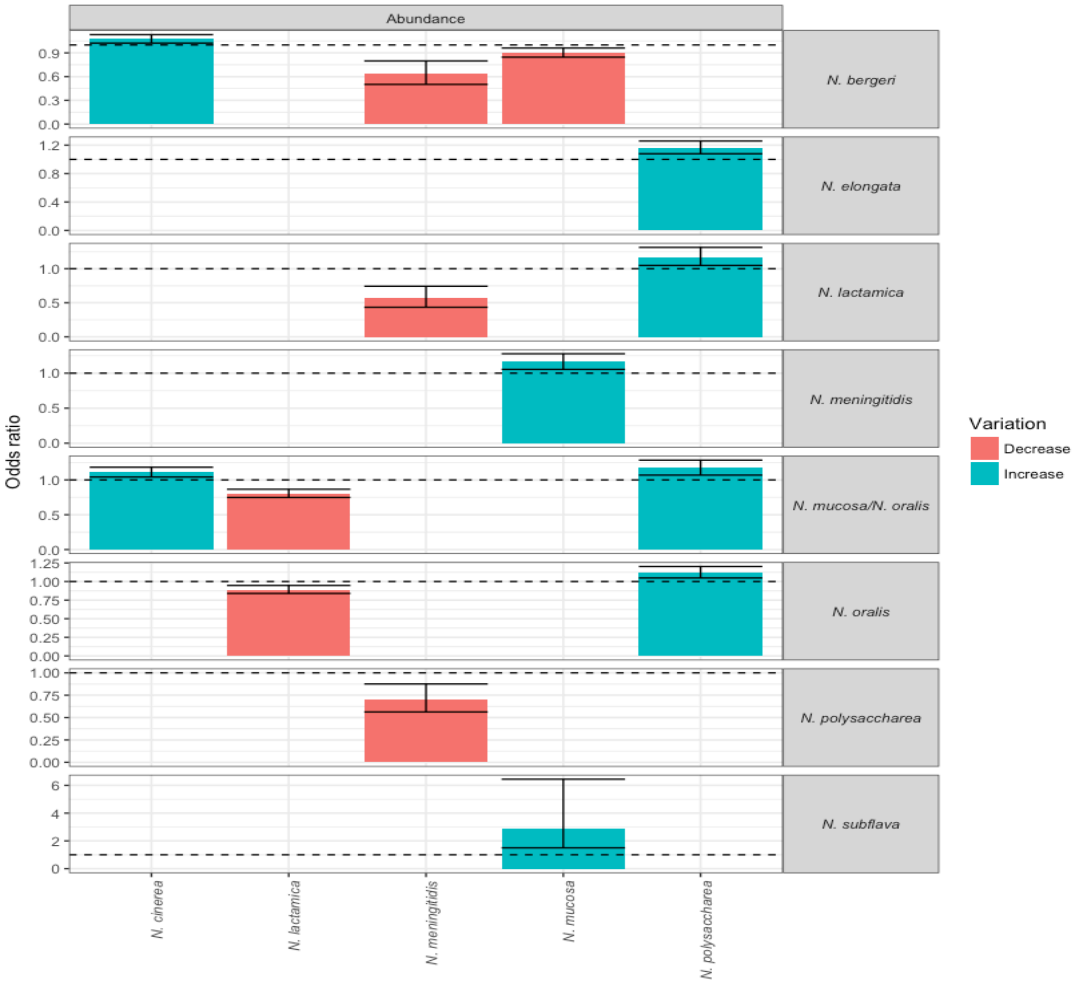


Figure 87 Odd ratios calculated for each species, marked in the X axis, when considering the abundance of each of the species marked in the Y axis labels. The reference group considered when calculating the odds was those colonized, hence, odd ratios above 1 indicate an increase in odds of having a higher abundance or presence when the species in the Y label is present.

8.4.2. Effect of early life events in neisserial populations

In order to determine the effect of various life events in carriage of certain *Neisseria* species, I used samples from the Dormice cohort. Extensive metadata was collected form infants at each sample collection time point, which was used in conjunction with *rplF* sequencing data to determine potential significant effects. As each infant enrolled in the Dormice cohort had several samples taken at

different time points, time series analysis methods can be used to identify potential time frames in which a significant difference may appear, instead of measuring changes at each time point.

Several methods have been developed with the intent of facilitating time series analysis for sequencing data. One of these, NMIT (175), relies on performing a PERMANOVA analysis at each time point, and then identifying correlations over time within groups of samples. Although this method seemed pertinent to use at first, the difficulty in its interpretability made it difficult to obtain any valuable insights. Another method, ANANKE (176), can provide more intuitive results than those of NMIT, and should work well in conjunction with the MED method used for sequence variant identification, but requires more time points for its algorithm to function properly than those collected for the Dormice dataset, and could not be used. Other methods, such as linear mixed effects models, have been used previously in microbiome studies (177). These methods work similarly to linear models but allow the calculation of varying intercepts and slopes for the different groups in the data. These methods could not be used for this analysis as the data did not meet the normality criteria for residuals and mixed effect values.

The metagenomeSeq package, used in Chapters 6 and 7 for the differential abundance analysis, also provides a method to model time series, using the fitTimeSeries function, and determine differences in abundance between two groups. This method relies on the gss (178) package in R. This algorithm considers the development of individual species over time for each set of samples taken from the same infant. It then tries to fit the average normalised relative abundance of one species from all infants to a smooth curve and assess the difference in area in between the groups being compared. The algorithm uses a non-parametric approach that relies on bootstrapping the dataset to find the underlying distribution of the data and identify significant differences in abundance. For this

analysis bootstrapping was performed 1000 times. This method allows us to identify intervals of time that correlate to a life event. In order to use this approach, the log10 of normalized reads were used to estimate differences, as suggested by the developer (163).

For the following analysis, I will present the results obtained after analysing several variables. Variables were set to have to levels, (*i.e.* Yes/No, High/Low), to accommodate the data to the limitations of the method. The levels and number of samples per level are described for each variable. I will present graphs similar to those shown in Figure 88. These show the results obtained after modelling the data using the fitTimeSeries function (Figure 88 A), with each level of the variable coloured differently. The solid line represents the predicted average, and the coloured area in between the dashed lines represents the 95% confidence interval. Finally, the area enclosed in between the two black dotted lines highlights the window of time in which significant differences appeared. The p-value showed at the top left corner of the figure refers to the p-value of this time window. Figure 88B shows boxplots for each of the two levels in the variable of interested per collection time point.

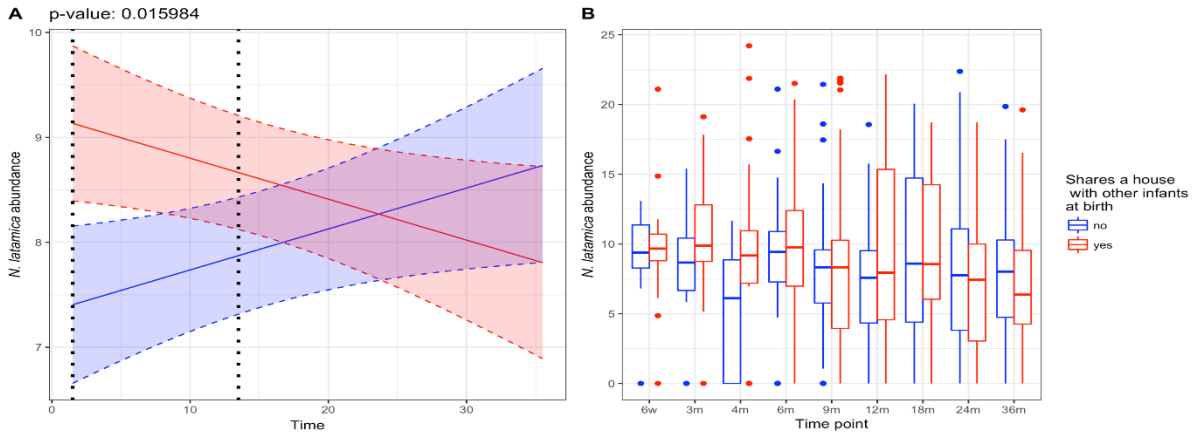


Figure 88 Example of output values obtained from the fitTimeSeries data (A), and box plots generated per time point and level in the variable of interest. Both figures show normalized read counts in the Y axis, and the time variable showing the months of life for figure A, and each sample collection time point for figure B on the X axis.

A Fisher's exact test was also used to calculate the odds ratio of each species that had a significant difference using the fitTimeSeries models. Two odd ratio values were calculated, one considering abundance of each species, and another considering presence or absence.

8.4.2.1. Effects of crowding in Neisseria species carriage

To assess the effect of crowding, which has previously been positively correlated with *N. meningitidis* carriage (87,88), I firstly considered the values from the conditions at birth, and then the values throughout life. Very similar results were obtained for both. To avoid repetition, I will only show those obtained when considering the data at birth. The crowding index was considered as the number of people living in the household per room (not considering the kitchen and restrooms). I divided infants in two categories, 'High' if the value was above one, or 'Low' if it was equal or lower than one. The distribution of crowding index values, and the number of samples per level are shown in Figure 89. The number of infants per level and time point are shown in Table 30.

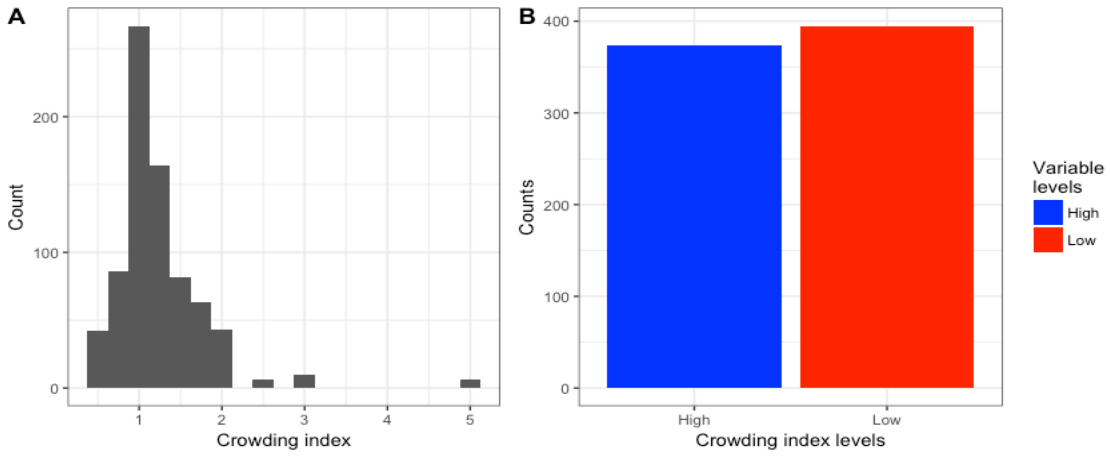


Figure 89 Distribution of crowding index values (A), and the number of samples per level (B). Figure A shows the number of samples aggregated by crowding index. Figure B shows the total number of samples when divided by into a 'low' or 'High' crowding index. A low crowding index is defined as a ratio of people living in a household devided by the number of rooms (excluding kitchen and bathrooms) equal or lower than one. A High crowding index was defined as a ratio higher than one.

Table 30 Number of infants per time point in each level considered.

LEVEL	6W	3M	4M	6M	9M	12M	18M	24M	36M
HIGH	13	23	24	43	55	63	59	54	44
LOW	16	19	20	53	55	62	63	62	53

This method identified three different species that differed significantly when considering high or low crowding index. *N. lactamica* showed an increase in the average number of counts in infants living in households with a crowding index above one, from 1.5 months until 16.5 months of life (Figure 90). This was paired with a significant difference in the abundance odds ratios, but not in the odds of this species being present (Table 31).

.

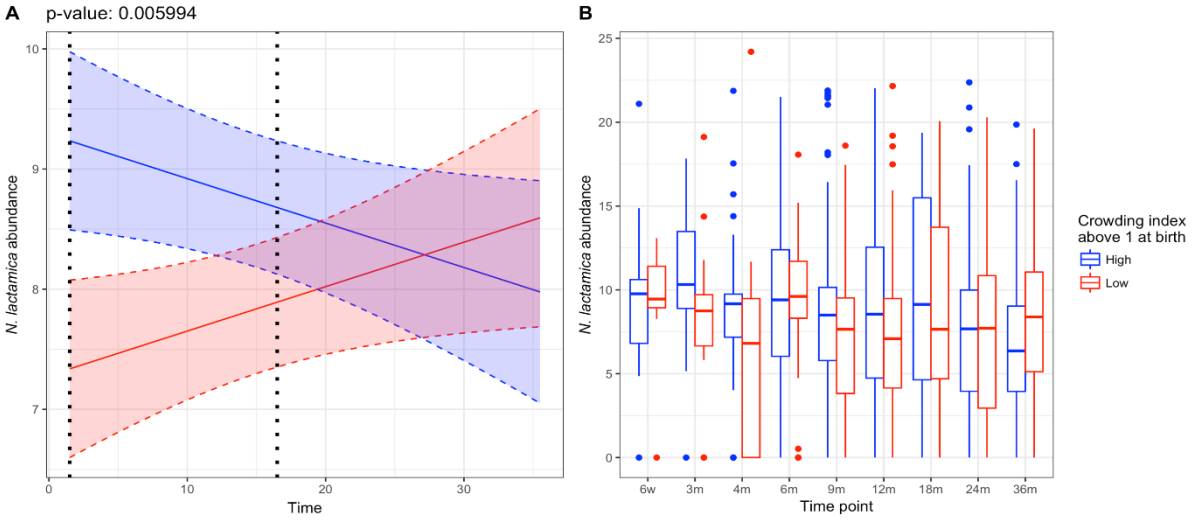


Figure 90 *A)* The modelled difference in *N. lactamica* abundance based on normalised read counts and B) the distribution of the normalized abundance data over time with regard to crowding at birth Both figures show normalized read counts in the Y axis, and the time variable showing the months of life for figure A, and each sample collection time point for figure B on the X axis.

N. subflava and *N. oralis* presented very similar patterns in their progression depending on crowding factors, both showing increases in abundance after 10.5 months in infants with a low crowding index. These two species also showed no differences when measuring the odds of them

being present in samples. When testing the odds of their abundance, *N. oralis* showed a significant difference, although *N. subflava* was not significantly different, but it was very close (p=0.06)

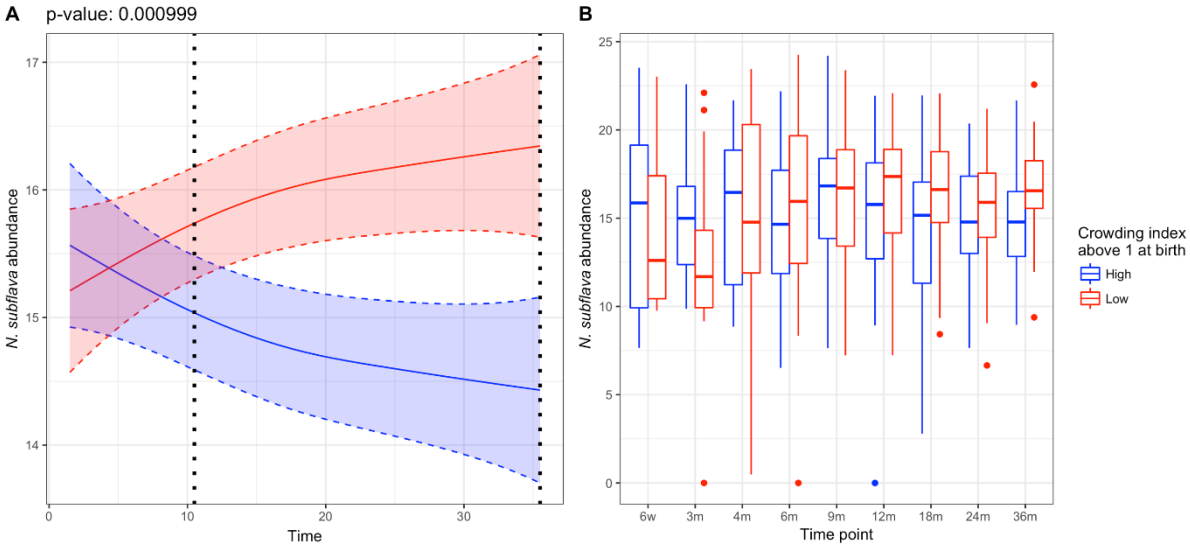


Figure 91 A) The modelled difference in *N. subflava* abundance based on normalised read counts and B) the distribution of the normalized abundance data over time with regard to crowding at birth. Both figures show normalized read counts in the Y axis, and the time variable showing the months of life for figure A, and each sample collection time point for figure B on the X axis.

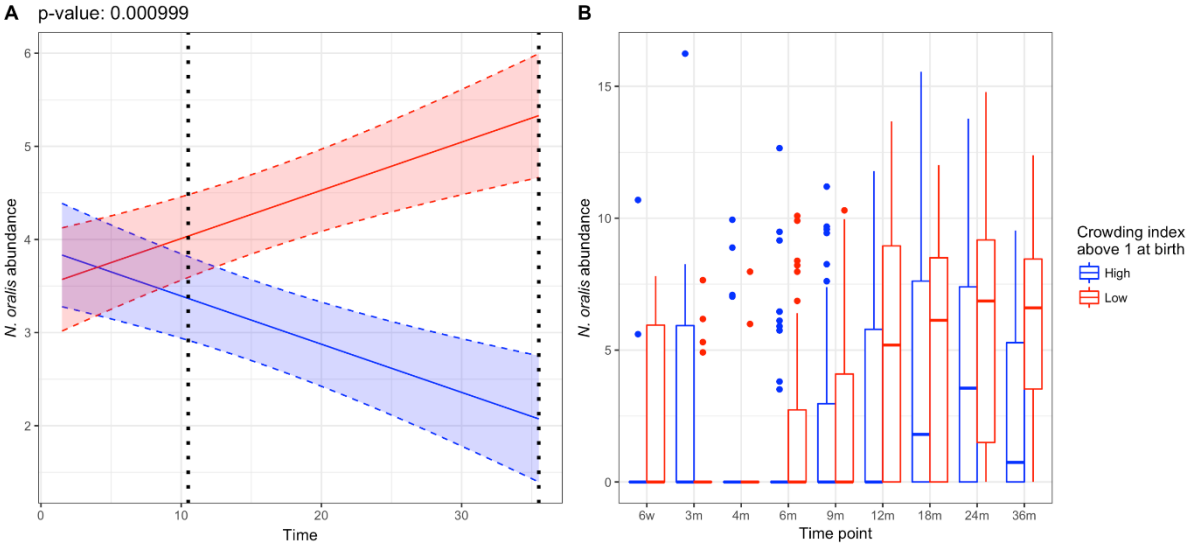


Figure 92 A) The modelled difference in *N. oralis* abundance based on normalised read counts and B) the distribution of the normalized abundance data over time with regard to crowding at birth. Both figures show normalized read counts in the Y axis, and the time variable showing the months of life for figure A, and each sample collection time point for figure B on the X axis.

Another variable measured was the presence of other infants in the household. This variable provided very similar results to those measured when considering the crowding index, as it could be expected that a crowded household would include other infants in it. To avoid repetition these results will not be shown.

Table 31 Odds ratios calculated based on abundance and presence of *N. lactamica*, *N. subflava* and *N. oralis* when considering infants in a high or low crowding index household. The odds ratio (Odds estimate) is presented with confidence intervals (Conf. int low and Conf. int High), the correspondent p-value, and the adjusted p-value using the Bonferroni correction.

ODDS		ODDS ESTIMATE	CONF.INT LOW	CONF.INT HIGH	P-VALUE	ADJ.P-VALUE
ABUNDANCE	*N. lactamica*	1.1430	1.0848	1.2043	0.0000	0.0000
	N. subflava	0.9445	0.9067	0.9839	0.0061	0.0606
	N. oralis	0.6953	0.6403	0.7547	0.0000	0.0000
PRESENCE	*N. lactamica*	0.4591	1.0632	0.9018	1.2535	0.9689
	N. subflava	0.9067	1.0119	0.8655	1.1829	0.9689
	N. oralis	0.0131	0.7615	0.6108	0.9477	0.1307

8.4.2.2. Effect of the use of a pacifier

Another factor previously described as a potential source of transmission of *N. meningitidis* is the use of a pacifier (88). The use or not of a pacifier was coded as 'Yes' or 'No', respectively. The number of samples per time point is shown in Table 32, and the total number per level is shown in Figure 93.

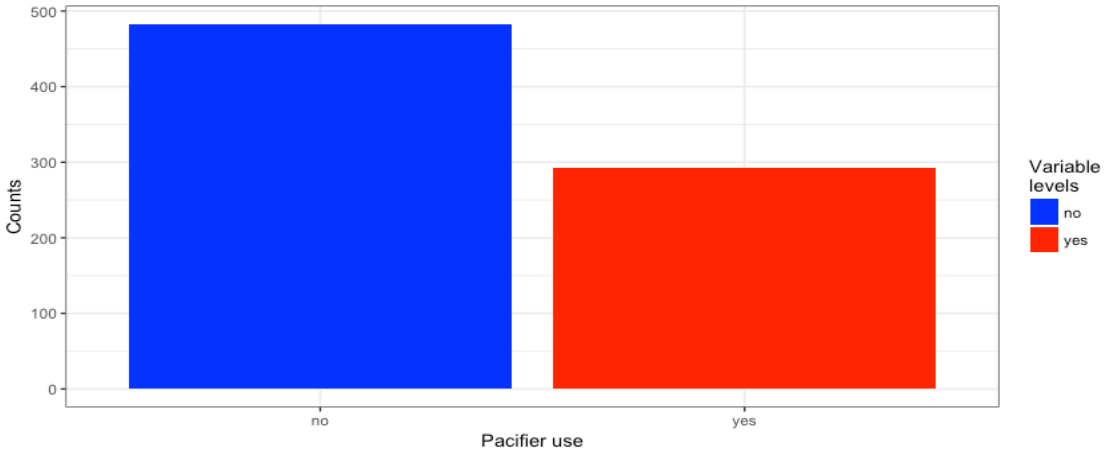


Figure 93 Number of samples (counts) per level when considering the use of pacifier.

Table 32 Number of infants per time point using (Yes) or not using (No) a pacifier.

LEVEL	6W	3M	4M	6M	9M	12M	18M	24M	36M
NO	18	20	230	59	72	75	78	76	62
YES	11	18	20	37	46	41	44	40	35

For this variable, three different species showed significant differences. *N. lactamica* (Figure 94) was related to an increase in abundance in those infants not using a pacifier in between 1.5 and 18 months, and 9.5 and 5.5 months, respectively. Sequences from the *N. mucosa/N. oralis* group (Figure 96) and *N. meningitidis* (Figure 95) showed an increase correlated with pacifier use in between 14.5 and 26 months, and 13 and 36 months, respectively. When considering the odds of the overall presence of these four species, only *N. lactamica* and *N. bergeri* showed a significant difference, while the other two did not. None of the species showed a significant difference in terms of presence in a particular group.

Table 33 Odds ratios calculated based on abundance and presence of *N. lactamica*, *N. bergeri, N. meningitidis* and *N. mucosa/N.* oralis when considering the use of a pacifier. The odds ratio (Odds estimate) is presented with confidence intervals (Conf. int low and Conf. int High), the correspondent p-value, and the adjusted p-value using the Bonferroni correction.

ODDS	SPECIES	ESTIMATE	CONF.INT LOW	CONF.INT HIGH	P-VALUE	ADJ.P-VALUE
ABUNDANCE	*N. lactamica*	0.9076	0.8613	0.9564	0.0003	0.0026
	N. bergeri	0.8103	0.7361	0.8915	0.0000	0.0001
	N. meningitidis	1.2282	1.0227	1.4758	0.0253	0.2529
	N. mucosa/ N. oralis	1.0167	0.9632	1.0731	0.5472	1.0000
PRESENCE	*N. lactamica*	0.5527	1.0528	0.8881	1.2499	0.9304
	N. bergeri	0.8517	1.0307	0.8019	1.3302	0.9304
	N. meningitidis	0.0195	0.5945	0.3820	0.9257	0.1950
	N. mucosa/ N. oralis	0.9304	0.9912	0.8328	1.1816	0.9304

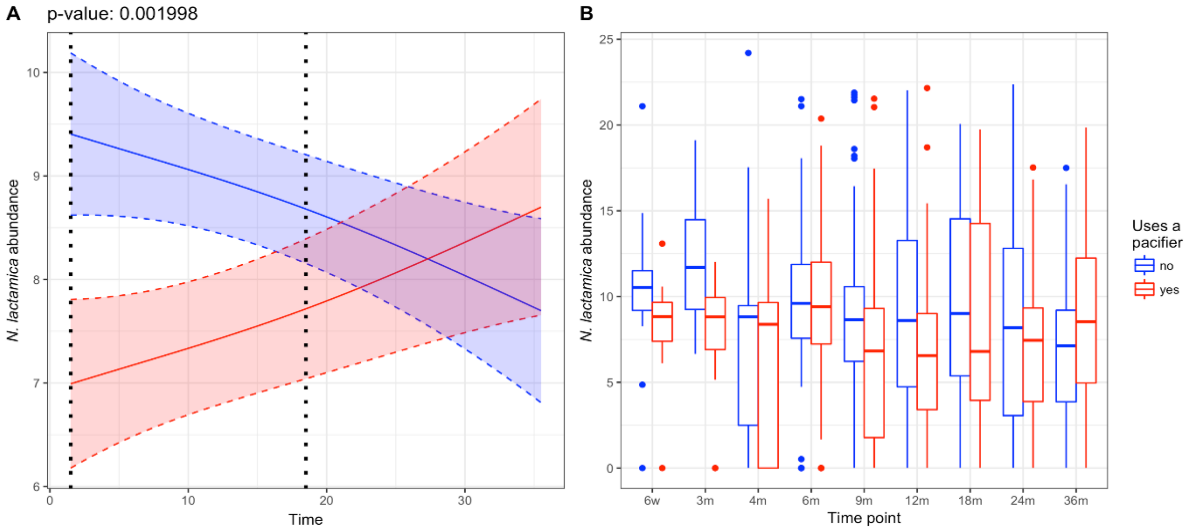


Figure 94 A) The modelled difference in *N. lactamica* abundance based on normalised read counts and B) the distribution of the normalized abundance data over time with regard s to the use of a pacifier. Both figures show normalized read counts in the Y axis, and the time variable showing the months of life for figure A, and each sample collection time point for figure B on the X axis.

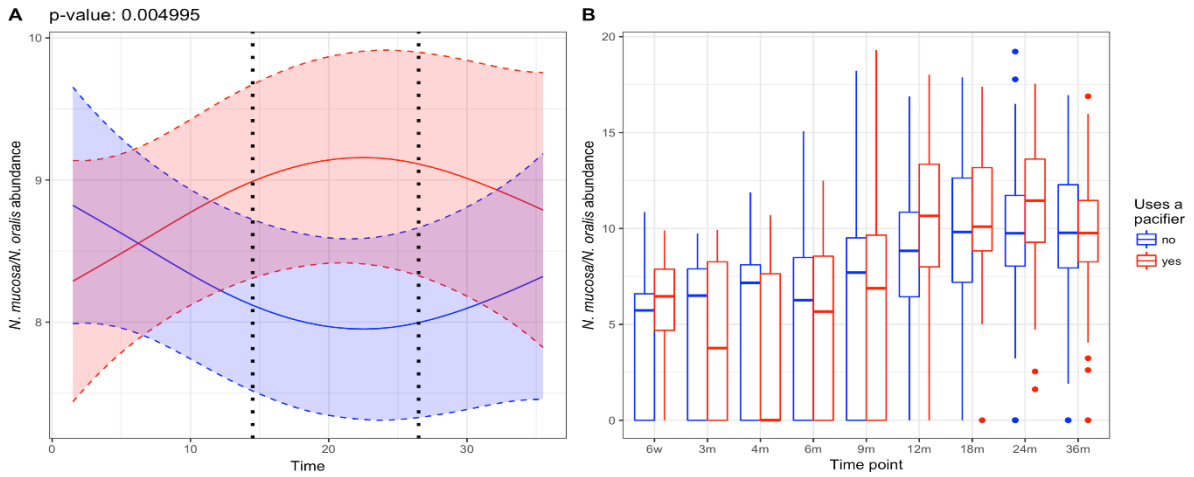


Figure 96 A) modelled difference in *N. mucosa/N. oralis* abundance based on normalised read counts and B) the distribution of the normalized abundance data over time with regard s to the use of a pacifier. Both figures show normalized read counts in the Y axis, and the time variable showing the months of life for figure A, and each sample collection time point for figure B on the X axis.

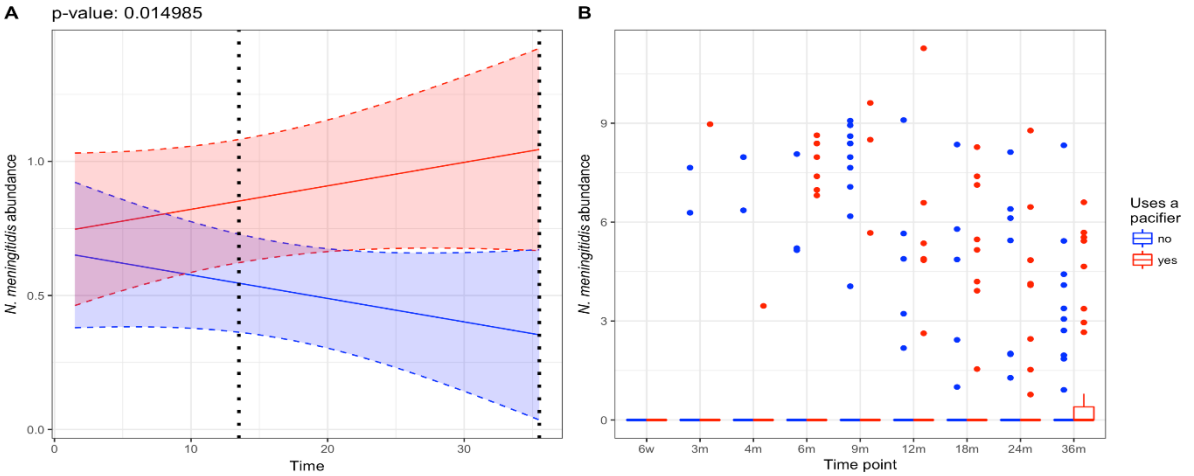


Figure 95 A) modelled difference in *N. meningitidis* abundance based on normalised read counts and B) the distribution of the normalized abundance data over time with regard s to the use of a pacifier. Both figures show normalized read counts in the Y axis, and the time variable showing the months of life for figure A, and each sample collection time point for figure B on the X axis.

8.4.2.3. Effect of living with a smoker in the household

Smoking, and the presence of a smoker in the household is another factor that has been linked to the increase in carriage of *N. meningitidis* (11,89). Similar to the data presented for the crowding index, data for the presence of a smoker was collected at birth and at each sample collection time

point. Values were coded as 'Yes' or 'No' for presence or absence of a smoker. The distribution of samples per time point and pr variable level are shown in Table 34 and Figure 97.

Table 34 Number of infants per time point living with (Yes) or without (No) a smoker in their household.

	6W	3M	4M	6M	9M	12M	18M	24M	36M
NO	24	27	33	7 5	88	87	90	89	75
YES	5	12	9	20	29	28	31	26	21

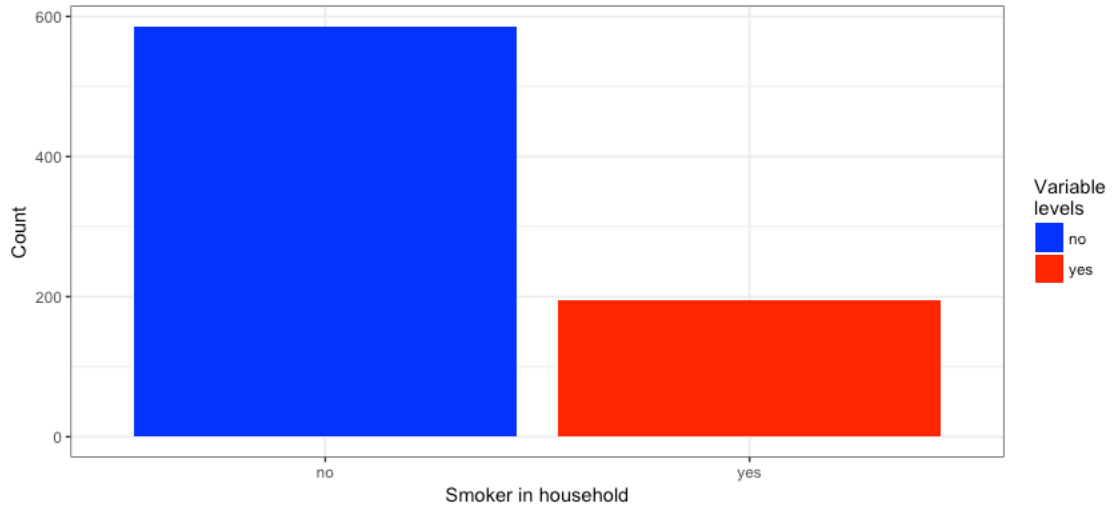


Figure 97 Number of samples (counts) per level when considering the presence of a smoker in the house.

The presence of smokers in the house had an impact on several *Neisseria* species. *N. bergeri* had a significant increase in those infants co-habiting with a smoker from birth until 18 months (Figure 98). *N. cinerea* had an increase, but up to 17.5 months (*Figure 99*). *N. mucosa* also had an increase in abundance from birth up to 7 months (Figure 100). Conversely, *N. oralis* and *N. subflava* were increased in households without smokers from 11.5 (Figure 101) and 21.5 (Figure 102) months onwards, respectively.

As tested before, the overall odds of the species abundance and presence were also calculated for each. Only *N. bergeri* and *N. oralis* had an increase in the odds of having a higher abundance in those infants living with smokers. Furthermore, *N. oralis* also had increased odds in prevalence in this level.

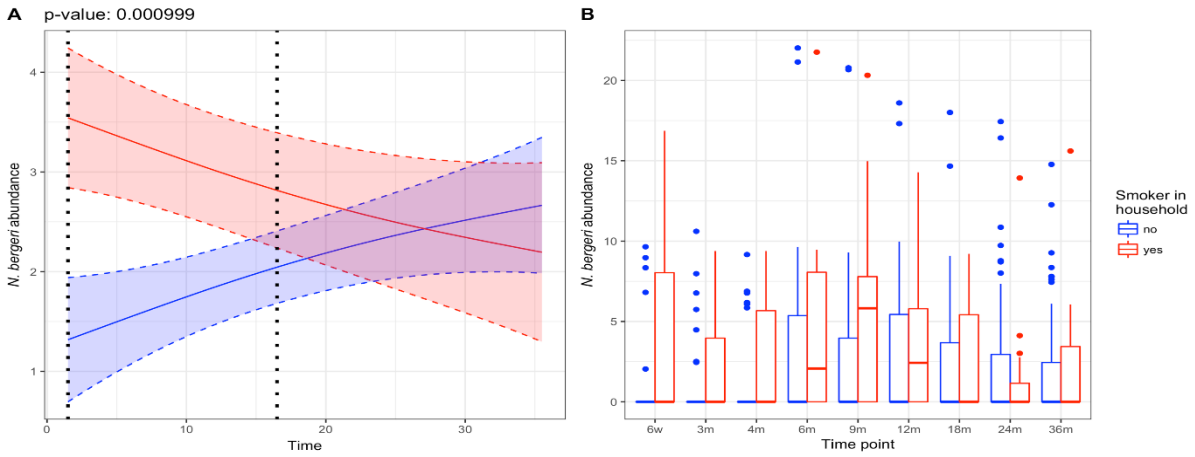


Figure 98 A) modelled difference in *N. bergeri* abundance based on normalised read counts and B) the distribution of the normalized abundance data over time when considering infants living with or without a smoker in the house. Both figures show normalized read counts in the Y axis, and the time variable showing the months of life for figure A, and each sample collection time point for figure B on the X axis.

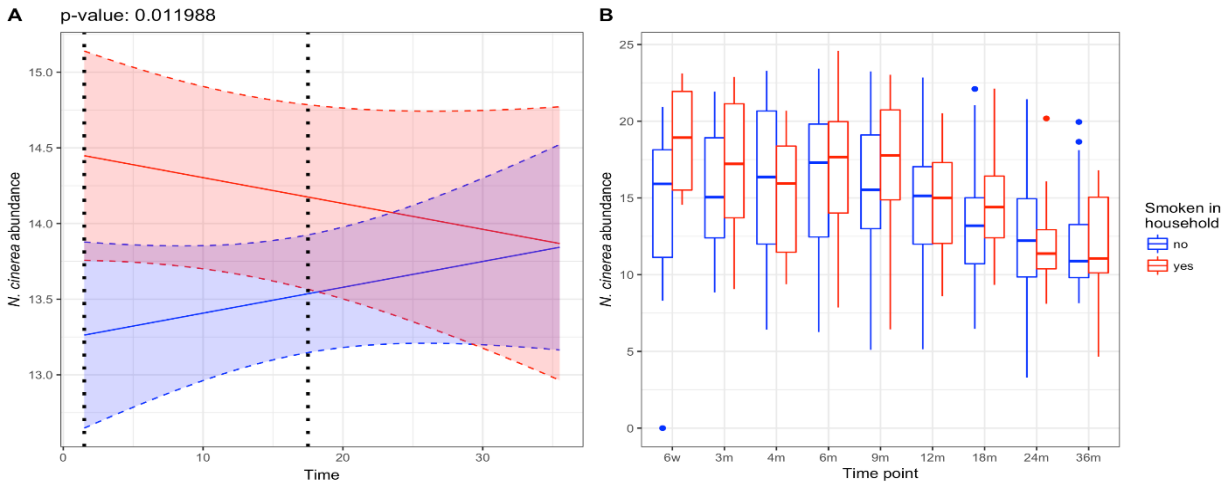


Figure 99 A) modelled difference in *N. cinerea* abundance based on normalised read counts and B) the distribution of the normalized abundance data over time when considering infants living with or without a smoker in the house. Both figures show normalized read counts in the Y axis, and the time variable showing the months of life for figure A, and each sample collection time point for figure B on the X axis.

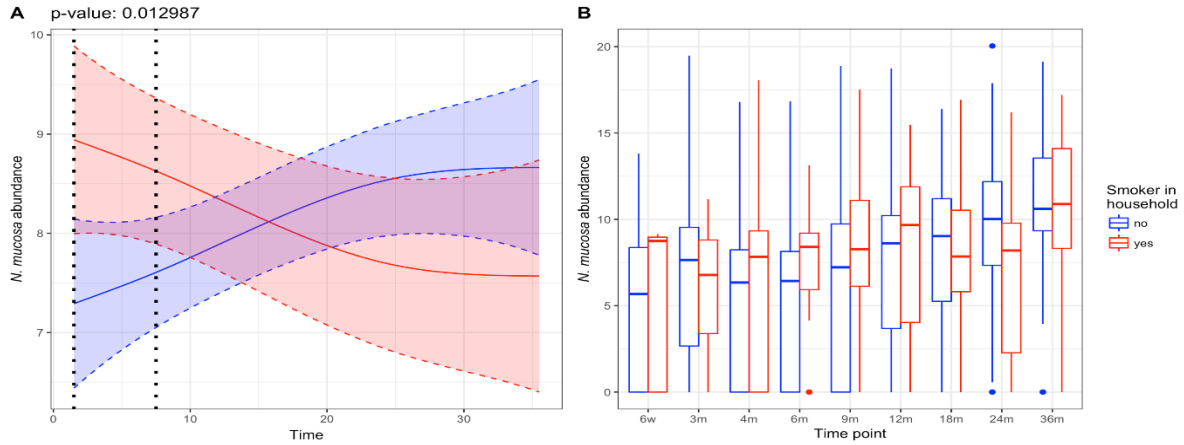


Figure 100 A) modelled difference in *N. mucosa* abundance based on normalised read counts and B) the distribution of the normalized abundance data over time when considering infants living with or without a smoker in the house. Both figures show normalized read counts in the Y axis, and the time variable showing the months of life for figure A, and each sample collection time point for figure B on the X axis.

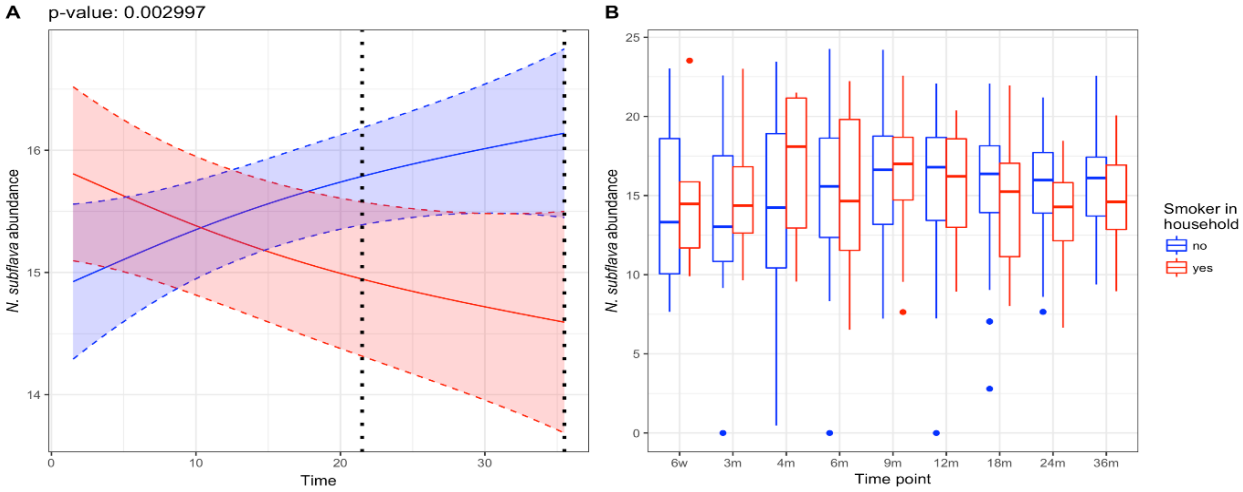


Figure 101 A) modelled difference in *N. subflava* abundance based on normalised read counts and B) the distribution of the normalized abundance data over time when considering infants living with or without a smoker in the house. Both figures show normalized read counts in the Y axis, and the time variable showing the months of life for figure A, and each sample collection time point for figure B on the X axis.

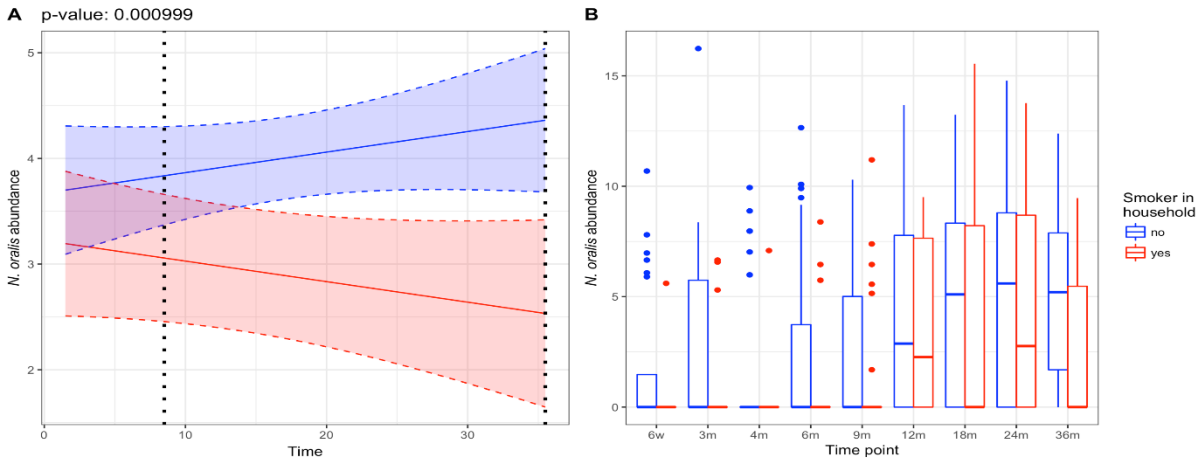


Figure 102 A) modelled difference in *N. oralis* abundance based on normalised read counts and B) the distribution of the normalized abundance data over time when considering infants living with or without a smoker in the house. Both figures show normalized read counts in the Y axis, and the time variable showing the months of life for figure A, and each sample collection time point for figure B on the X axis.

Table 35 Odds ratios calculated based on abundance and presence of *N. bergeri, N. cinerea, N. mucosa, N. oralis* and *N. subflava* when considering infants living with a smoker in the household. The odds ratio (Odds estimate) is presented with confidence intervals (Conf. int low and Conf. int High), the correspondent p-value, and the adjusted p-value using the Bonferroni correction.

ODDS	SPECIES	ODDS ESTIMATE	CONF.INT LOW	CONF.INT HIGH	P-VALUE	ADJ.P-VALUE
ABUNDANCE	*N. bergeri*	0.6724	0.6080	0.7442	0.0000	0.0000
	N. cinerea	0.9561	0.9105	1.0041	0.0704	0.1761
	N. mucosa	0.2049	0.9607	0.9030	1.0225	0.3999
	N. oralis	1.4198	1.2840	1.5723	0.0000	0.0000
	N. subflava	1.0398	0.9910	1.0913	0.1114	0.1857
PRESENCE	*N. bergeri*	0.7242	0.5573	0.9469	0.0147	0.0734
	N. cinerea	1.0010	0.8338	1.2055	1.0000	1.0000
	N. mucosa	0.7310	0.9681	0.7971	1.1802	1.0000
	N. oralis	1.5004	1.1425	1.9930	0.0029	0.0288
	N. subflava	0.9859	0.8250	1.1817	0.8928	1.0000

8.4.2.4. Effect of pets in household

Pets have been associated with alterations in the upper respiratory tract microbiome (179), and have been associated with a decrease in carriage of *N. meningitidis* (180). It was decided to include the presence of pets, and their effect on *Neisseria* species carriage. For this analysis, two different

variables were considered, first all pets, and then just furry pets, which in this analysis included dogs, cats and rabbits. A breakdown of the number of samples with regards to the presence of a pet is shown in Figure 103 and Table 36.

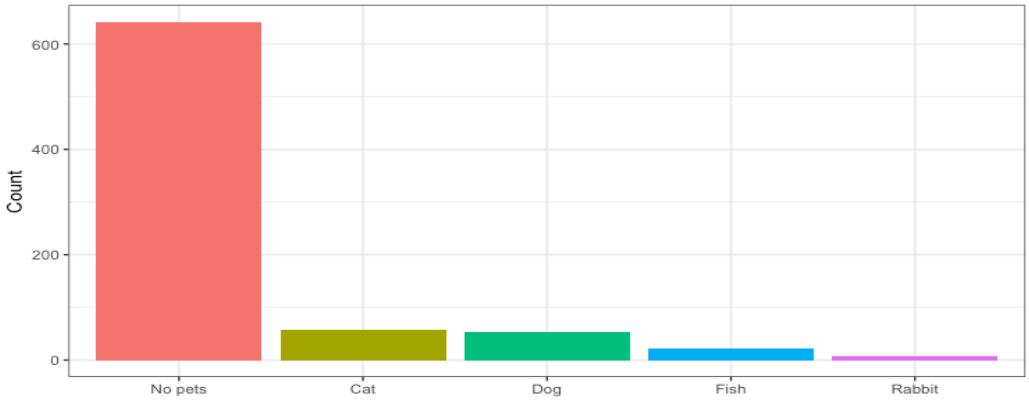


Figure 103 Number of samples collected (counts) in households without pets (red), and those were at least one pet was present (Cat = yellow, Dog = green, Fish = blue, Rabbit = purple).

Table 36 Number of infants per time point not living with pets (No), those living with pets (Yes-All), and those living with furry pets in particular (Yes – Furry).

	6W	3M	4M	6M	9M	12M	18M	24M	36M
NO	23	33	36	77	96	94	99	94	80
YES - ALL	6	6	6	18	21	21	22	21	16
YES - FURRY	4	6	6	15	17	18	18	18	14

When considering all pets, *N. elongata* saw an increase from birth until 13.5 months in infants living without pets (Figure 104), and *N. bergeri* was increased from 18 months onwards in infants living with pets (Figure 105). However, when considering furry pets only, *N. elongata* did not show any significant differences, and *N. bergeri* saw a similar increase, but from 2 months onwards. *N. lactamica* appeared to have a significant increase in abundance in infants living with furry pets (Figure 106).

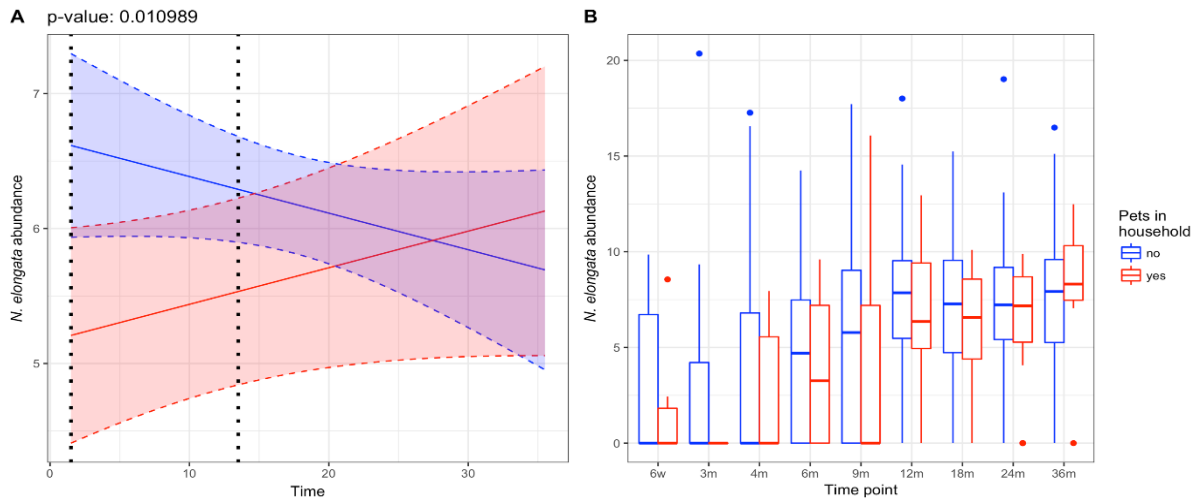


Figure 104 A) modelled difference in *N. elongata* abundance based on normalised read counts and B) the distribution of the normalized abundance data over time when considering infants living in a household with a pet. Both figures show normalized read counts in the Y axis, and the time variable showing the months of life for figure A, and each sample collection time point for figure B on the X axis.

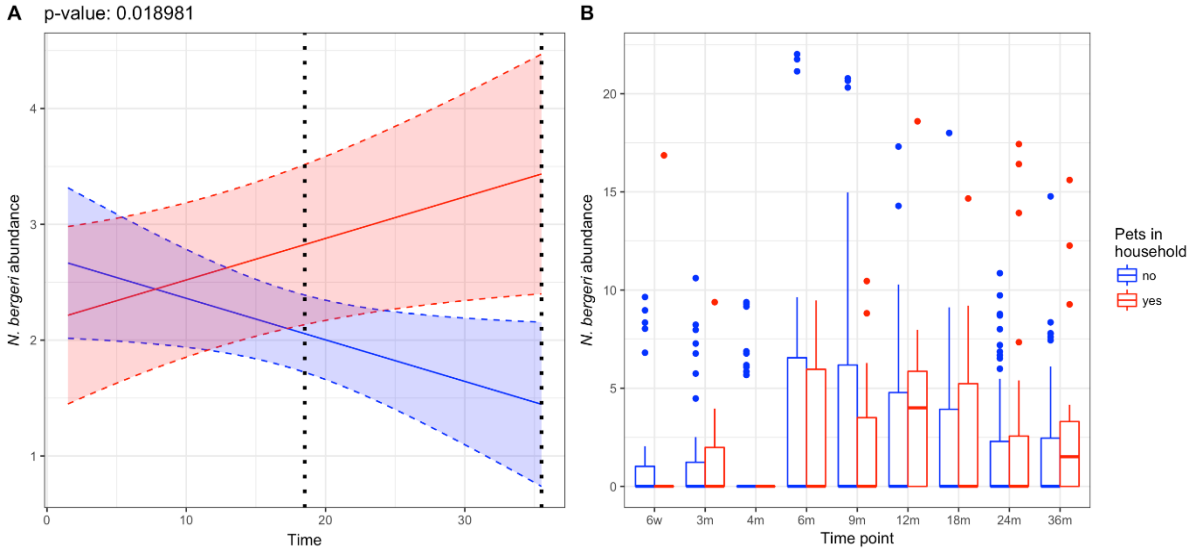


Figure 105 A) modelled difference in *N. bergeri* abundance based on normalised read counts and B) the distribution of the normalized abundance data over time when considering infants living in a household with a pet. Both figures show normalized read counts in the Y axis, and the time variable showing the months of life for figure A, and each sample collection time point for figure B on the X axis.

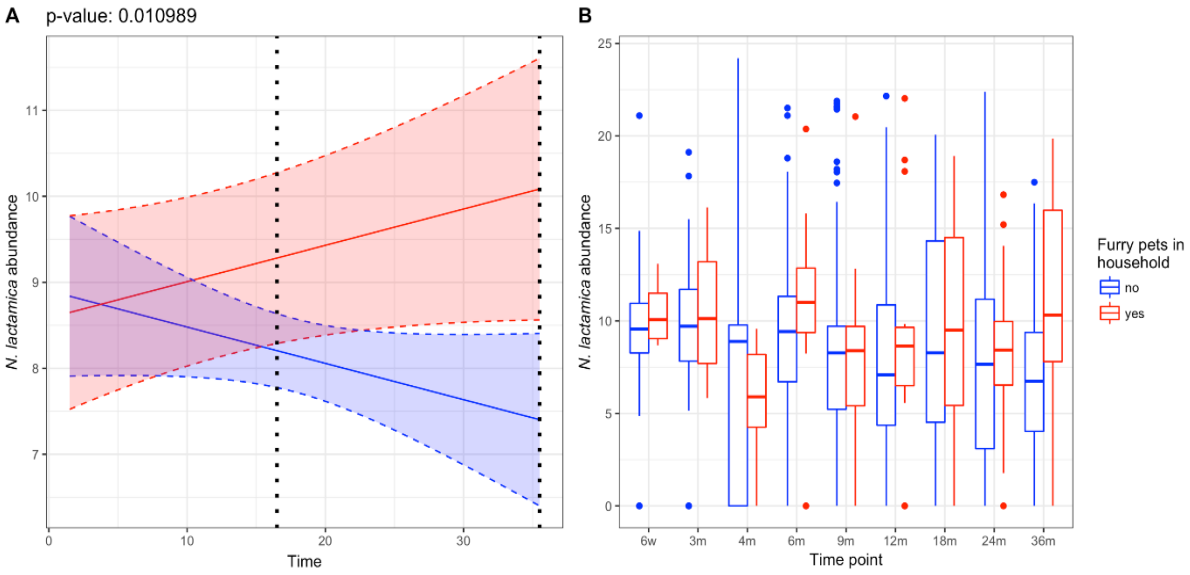


Figure 106 A) modelled difference in *N. lactamica* abundance based on normalised read counts and B) the distribution of the normalized abundance data over time when considering infants living in a household with a furry pet. Both figures show normalized read counts in the Y axis, and the time variable showing the months of life for figure A, and each sample collection time point for figure B on the X axis.

The analysis of the odds based on the abundance of each species revealed a larger number of differences than the time series analysis(Table 37). However, the analysis based on presence showed no difference for any species. When considering all pets, *N. bergeri* and *N. elongata* had increased odds mirroring the time series analysis. *N. lactamica* also showed an increase in odds in infants with pets, while *N. subflava, N. meningitidis* and *N. polysaccharea* saw increased odds of their abundance being increased in infants without pets. When considering just furry pets, *N. lactamica* had similar results to the time series analysis, while *N. bergeri* did not show a significant increase in odds. *N. subflava, N. elongata* and *N. meningitidis* showed again an increase in odds in infants with without pets.

Table 37 Odds ratios calculated based on abundance for all *Neisseria* species when considering infants living with a pet (pets-All), or just furry pets (Pets-Furry). The odds ratio (Odds estimate) is presented with confidence intervals (Conf. int low and Conf. int High), the correspondent p-value, and the adjusted p-value using the Bonferroni correction

	SPECIES	ESTIMATE	CONF.INT LOW	CONF.INT HIGH	P-VALUE	ADJ.P-VALUE
PETS - ALL	*N. lactamica*	0.9128	0.8540	0.9761	0.0074	0.0148
	N. bergeri	0.8345	0.7423	0.9398	0.0027	0.0091
	N. subflava	1.0795	1.0228	1.1397	0.0052	0.0129
	N. elongata	1.1404	1.0488	1.2413	0.0018	0.0091
	N. cinerea	0.9544	0.9040	1.0077	0.0905	0.1131
	N. meningitidis	1.3835	1.0640	1.8236	0.0145	0.0241
	N. polysaccharea	1.0951	1.0109	1.1874	0.0251	0.0359
	N. oralis	1.0636	0.9560	1.1853	0.2648	0.2942
	N. mucosa/ N. oralis	0.9981	0.9304	1.0714	0.9573	0.9573
	N. mucosa	0.8953	0.8354	0.9600	0.0018	0.0091
PETS - FURRY	*N. lactamica*	0.8936	0.8329	0.9594	0.0019	0.0184
	N. bergeri	0.9243	0.8133	1.0532	0.2311	0.3302
	N. subflava	1.0808	1.0203	1.1453	0.0077	0.0193
	N. elongata	1.1403	1.0427	1.2486	0.0037	0.0184
	N. cinerea	0.9813	0.9261	1.0403	0.5250	0.5250
	N. meningitidis	1.4736	1.1049	2.0009	0.0064	0.0193
	N. polysaccharea	1.0545	0.9691	1.1487	0.2206	0.3302
	N. oralis	0.9605	0.8604	1.0744	0.4621	0.5250
	N. mucosa/ N. oralis	0.9757	0.9057	1.0518	0.5179	0.5250
	N. mucosa	0.9263	0.8601	0.9984	0.0434	0.0868

8.4.2.5. Effects of feeding in neisserial populations

Breastfeeding has been proposed as a protective factor towards the development of diseases, such as pneumococcal disease (181,182). It was thus considered a variable to be included in this analysis to identify its potential impact in *Neisseria*. As opposed to the variables previously mentioned, differences were identified when considering the feeding routine at birth, and the feeding at each sample collection time point. Results for both sets will be presented here. First, considering feeding at birth (Figure 107), the majority of infants were breastfed, with the next most numerous group being the samples from infants that received mixed feeding. Due to the very low number of samples from

infants that were formula-fed at birth (two infants, 11 samples in total), this group was excluded from the analysis.

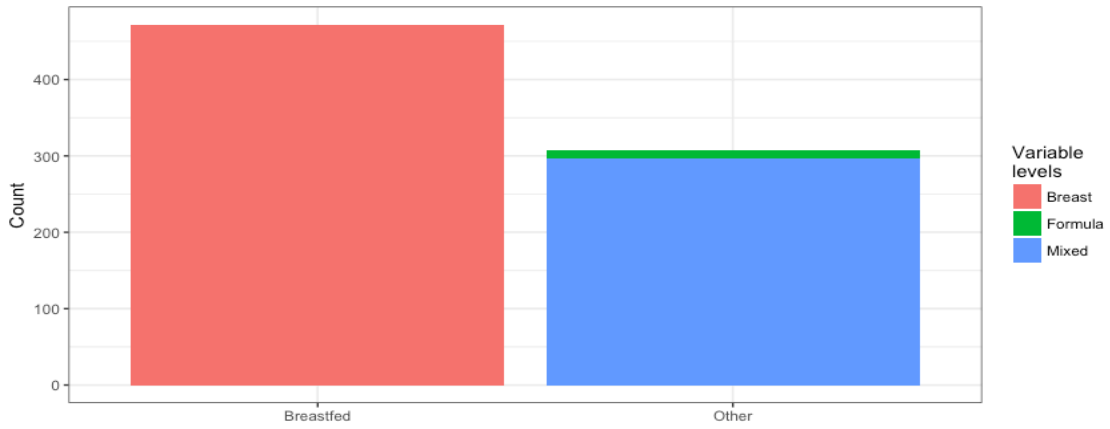


Figure 107 Number of samples (Count) from infants that were breastfed, formula fed, or had mix feeding at birth. Bars representing the number of samples per level are coloured according to the legend of the right.

When classifying samples according to the feeds received prior to sample collection, the majority of samples were collected from infants being breast fed, followed by those having a mixed feeding, and then a formula-fed diet. Finally, there were 12 samples, from two infants, that were weaned. This group was excluded from the analysis due to low number of samples (Figure 108). The final number of samples per time point is show in Table 38.

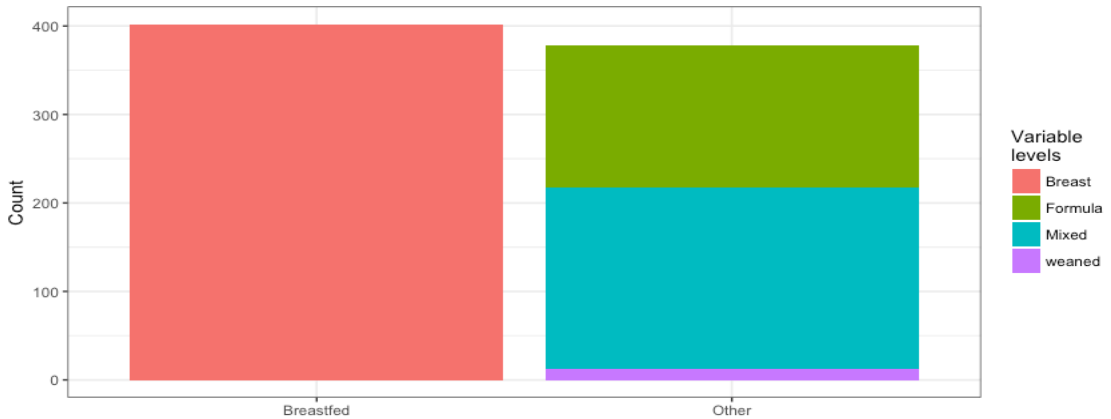


Figure 108 Number of samples (Count) from infants that were breastfed, formula fed, or had mix feeding prior to sample collection. . Bars representing the number of samples per level are coloured according to the legend of the right.

Table 38 Number of infants per time point and feeding routine considered for the final analysis. The data is show when considering the feeding at birth, and at each sample collection visit.

		6W	3M	4M	6M	9M	12M	18M	24M	36M
FEEDING AT BIRTH	Breast	18	23	23	59	74	72	74	72	57
	Mixed	11	19	21	35	42	42	46	42	39
FEEDING AT VISIT	Breast	19	22	20	53	58	59	64	60	47
	Formula	5	8	11	22	24	23	25	22	21
	Mixed	5	12	13	19	34	32	31	32	27

When considering feeding at birth, *N. cinerea* and *N. subflava* showed an increase in abundance from birth until 14 and 8 months, respectively, in mix-fed infants. *N. mucosa* was increased in breastfed infants from 2 to 18 months. These three species showed an opposite trend when testing the mix fed infants.

The analysis of the feeding routines at each sample collection visit, *N. oralis* was increased in breastfed infants form 16 months onwards. An opposite trend was identified in mixed-fed infants but not in formula-fed infants. Formula fed infants had increased abundances of *N. subflava* and *N. bergeri* from birth until 7 and 8 months, respectively, while mix-fed infants had an opposite trend for these two species. A summary of all variations is shown in Figure 109.

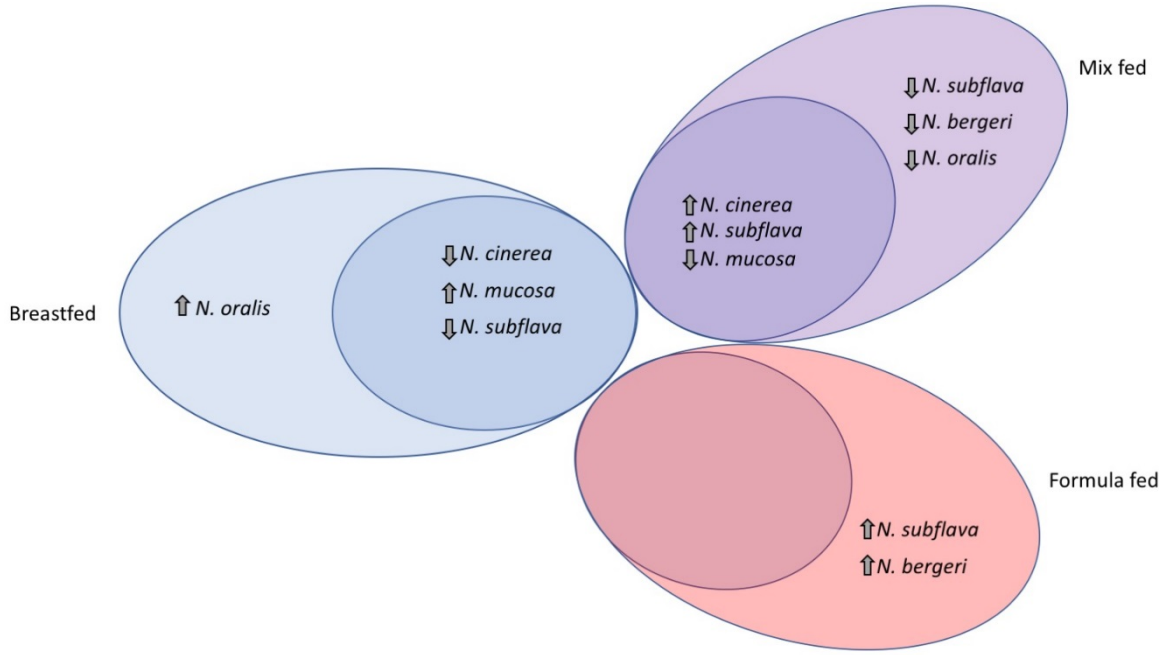


Figure 109 Increase or decrease in abundance measured for each feeding routine. In each ellipsis, the inner darker circles show the results obtained from birth feeding data, while the others show the data for the feeding at the time of sample collection. The arrows show an increase (up) or a decrease (down) measured for each species.

Finally, when calculating the odds of increased abundance per feeding routine, different species were significantly different (Table 39). Breast fed infants showed increased odds for *N. lactamica, N. oralis* and *N. mucosa,* while seeing a decrease in *N. meningitidis* and *N. polysaccharea*. These differences were measured for both feeding at birth and per visit. There was not increased or decreased odds of presence of any *Neisseria* species for any species or feeding routine.

Table 39 Odds ratios calculated based on abundance for *N. lactamica, N. meningitidis, N. oralis* and *N. mucosa* when considering breastfed infants at birth and at each sample collection visit. The odds ratio (Odds estimate) is presented with confidence intervals (Conf. int low and Conf. int High), the correspondent p-value, and the adjusted p-value using the Bonferroni correction.

	SPECIES	ESTIMATE	CONF.INT LOW	CONF.INT HIGH	P-VALUE	ADJ.P-VALUE
FEEDING AT BIRTH BREASTFED	*N. lactamica*	1.0845	1.0282	1.1439	0.0027	0.0266
	N. meningitidis	0.7599	0.6336	0.9120	0.0026	0.0261
	N. polysaccharea	0.8891	0.8372	0.9443	0.0001	0.0012
	N. oralis	1.1366	1.0466	1.2350	0.0021	0.0214
	N. mucosa	1.1395	1.0776	1.2053	0.0000	0.0000
FEEDING AT VISIT BREASTFED	*N. lactamica*	1.0845	1.0282	1.1439	0.0027	0.0266
	N. meningitidis	0.7599	0.6336	0.9120	0.0026	0.0261
	N. polysaccharea	0.8891	0.8372	0.9443	0.0001	0.0012
	N. oralis	1.1366	1.0466	1.2350	0.0021	0.0214
	N. mucosa	1.1395	1.0776	1.2053	0.0000	0.0000

8.5. Discussion

The carriage of *Neisseria* species seems to be quite variable. While, as seen in previous results (Chapter 7, Section 7.4.5) most infants are colonized by *N. cinerea* and *N. subflava,* the proportions of these two species vary in every time point for a given infant. Furthermore, other *Neisseria* species also colonize the upper respiratory tract, but the number of time points in which these appear, and the proportion of reads they represent is also varied. However, there seems to be a trend for *N. lactamica* and *N. polysaccharea*, and perhaps to a lesser extent *N. bergeri,* to become the most prevalent species in the *Neisseria* microbiome when present. These trends are consistent with the ordination results obtained in Chapter 7 (Section 0) in which *N. lactamica* and *N. polysaccharea* seemed to appear in separate clusters from the rest of *Neisseria* species. It could be possible that these two species generate an environment in which other *Neisseria* species may not thrive. This could be one of the reasons why *N. lactamica* has been historically associated with protection towards *N. meningitis* colonization (9). A similar hypothesis could be stated for *N. polysaccharea* and *N. bergeri*, *i.e.* an exclusion effect on other *Neisseria* species, in addition to any potential antigenic similarity to *N.*

meningitidis (45). *N. bergeri* was first described by Bennet *et al.* (52) as a separate species, very similar to *N. polysaccharea*. This tendency towards monopolizing the neisserial niche could be due to their relatedness.

When looking at the odds of individual *Neisseria* species having a higher abundance, or being present, based on the presence of another, some trends were apparent. *N. cinerea* and *N. subflava* had their odds reduced by almost all other species, which could be expected as these two species are the two most prominent species in the *Neisseria* population, and other species appearing will likely overtake part of the total population.

A suggested synergistic relationship between *N. mucosa, N. oralis* and *N. elongata* was found, *i.e.* an increase in the odds of all three species having a higher abundance when any of the others were present. The odds of having a higher carriage of *N. polysaccharea* were also increased in infants that had been previously colonized by any of the three species. However, *N. polysaccharea* seemed to decrease the odds of *N. mucosa* and *N. elongata* being present. Meanwhile, *N. meningitidis* was more likely to appear in those infants that were also colonized by *N. lactamica* or *N. polysaccharea* at the same time point, although the opposite effect was seen when considering the colonization of these two species at the previous time point. A similar trend was observed in which *N. polysaccharea* had lower odds of appearing if the infants were also colonized with *N. mucosa* or *N. elongata*, but the odds increased if *N. oralis, N. mucosa/N. oralis, N. lactamica* or *N. elongata* were present at the previous time point. *N. subflava* also contributed towards *N. mucosa* (and *N. mucosa*/*N. oralis*), while decreasing the odds of *N. bergeri* and *N. cinerea*.

The results seen with *N. lactamica* and *N. meningitidis* are especially intriguing as they seem to partially contradict the previous literature suggesting that *N. lactamica* colonization can protect against meningococcus colonization and invasive disease (it should be noted that the available data is suggestive but not conclusive). While *N. meningitidis* had lower odds of being present when *N. lactamica* was found in the previous time point, which would agree with previous literature, it is difficult to identify a theory which explains why *N. meningitidis* would have higher odds of appearing in infants already colonized by *N. lactamica*. One possibility is that *N. meningitidis* takes advantage of the disrupted environments generated by, *N. lactamica, N. polysaccharea,* and *N. bergeri,* which seemed to predominate the neisserial population when present, as hypothesized before. However, if *N. meningitidis* overcome this effect, their survival may be enhanced, leading to a higher abundance.

Looking at the broad picture created by the data, *N. cinerea, N. bergeri* and *N. lactamica* may act as independent agents, colonizing, and utilizing their gene repertoire to predominate the neisserial population. Meanwhile, *N. polysaccharea, N. subflava, N. elongata, N. mucosa, N. oralis* seem to generate a cycle of inhibition of growth among themselves.

While the interaction of *Neisseria* species among themselves seems to play a role in the development of the neisserial population, certain life events may also favour some species over others. Several publications (88,180,183,184) measured the effect of certain variables as correlates of *N. meningitidis* and *N. lactamica* colonization. Here I present data on how certain life events may impact *Neisseria* species. However, due to the lack of research done in most *Neisseria* species, it is difficult to draw robust conclusions, and most of the arguments should be considered hypothesis-generating.

When considering the effect of the crowding index, be it at birth or throughout the first three years of life, or the presence of other children in the same household, *N. lactamica* abundance was increased earlier in life in infants in crowded households. Although purely speculative, this could be due to *N. lactamica* being more likely to be passed on from their older sibling. *N. oralis* and *N. subflava* showed an opposite trend, being more abundant later on in life in households that were less crowded, consistent with the trend seen earlier with the odds of colonization. Increases in *N. meningitidis* have been associated with high crowding, but usually in studies where the population tends to have higher levels of crowding (87,112), while others similar to those in our data found no correlation (11).

The use of a pacifier also affected *N. lactamica*, it being more abundant in those infants that did not use one up to 18 months, while *N. meningitidis* abundance was increased in those that used one from 13 months onwards. This difference could be an effect that derives from infants that do not use a pacifier being fed more frequently than those that use one, and the feeding may enhance *N. lactamica* growth. This may in turn provide a passive immunity towards *N. meningitidis* that infants using a pacifier may be less likely to acquire. Pacifiers have also been hypothesized as potential transmitter of the *N. meningitidis* among children (88).

Smoking and passive smoking have also been associated with an increase in carriage of *N. meningitidis* (116,185). In this analysis, *N. cinerea* and *N. bergeri* and *N. mucosa* abundance was increased in infants with a smoker in the household earlier in life, while *N. oralis* and *N. subflava* saw an increase in abundance. However, no difference was identified for *N. meningitidis*. A meta-analysis of the study of second-hand smoking in pediatric settings (186) found positive correlations with *N. meningitidis, S. pneumonaie* and *H. influenzae.* However, the majority of these studies dealt with

numbers of samples much higher than the ones presented here, which could explain the disparity in between results.

Neisseria has been shown to be transmitted from dogs to owners (187), and has been associated with changes in the microbiome (155,188). However, none of these studies have dealt with the effects of pets on *Neisseria* populations. Infants living with pets had an increased abundance of *N. lactamica* and *N. bergeri* later in life, while they also had a decrease in *N. elongata* earlier on. Although no significant differences were identified for *N. meningitidis, N. polysaccharea* or *N. subflava* in the longitudinal analysis, their odds ratio was increased in those infants without pets. Several *Neisseria* species have been associated with pets, such as cats and dogs (4), however, none of these species have been identified in them. It could be possible that pets are not colonized by them permanently but could act as transient carriers of human bacteria.

Finally, the feeding routine of the infants seemed to impact the development of neisserial populations. Each feeding pattern seemed to lead to changes in different species, both taking into consideration feeding at birth and throughout the first three years of life. The differences seen considering the feeding at birth were balanced in between infants having a mixed diet and those that were breastfed, with the exception of the decrease in *N. polysaccharea* in breastfed infants. The differences were more spread later on in life, although mixed fed infants had a decrease in *N. subflava, N. bergeri* and *N. oralis*, while breastfed infants saw an increase in *N. oralis,* and formula fed infants had increased abundance on the other two species. When considering just the overall odds of each species, *N. lactamica* odds were increased in breastfed infants, and *N. meningitidis* was decreased. This suggests that breastfeeding could provide either nutrients that benefit the growth of *N.*

lactamica, or maternal antibodies being passed on through the breastmilk could prevent *N. meningitidis* from colonizing or growing.

When looking at all the variables together, there seems to be a certain trend in the odds, in which *N. lactamica*, *N. bergeri* and *N. cinerea* always seem to oppose the trend of *N. meningitidis*, which is usually accompanied by *N. mucosa* and *N. polysaccharea*. This could fit into the data proposed earlier when comparing the odds of abundance and presence of each bacteria. This could suggest that *N. cinerea, N. bergeri* and *N. lactamica* may be the true agents protecting from colonization of *N. meningitidis* and the other species may generate a more advantageous environment for its establishment in the upper respiratory tract.

9. Chapter 9: Discussion

9.1. *rplF as a Neisseria speciation tool*

Since the development of the *rplF* gene as a speciation tool for *Neisseria* species (40), it has been applied on several occasions to identify *Neisseria* species in human specimens (51,189), as a supplementary tool to identify new *Neisseria* species (43,44), and has inspired a similar approach using the *rpsB* gene to identify *Streptococcus* species (190). While the implementation of neisserial *rplF* as described in its original publication is applicable to Sanger sequencing, several steps had to be optimized to enable its use by the Illumina NGS platform.

After validation of the method to correctly identify *Neisseria* species in both DNA extracted from human samples and pure bacterial cultures, two issues were identified as potentially problematic for downstream processes: The seeming lack of variability of *rplF* sequences, and the mismatches in the primer binding sequence of the Forward primer used.

9.1.1. Problems arising from lack of sequence diversity

Neisseria species usually associated with human hosts showed a high degree of similarity in *rplF* amplicon sequences, unlike those associated with animal hosts which were more divergent. Sequence similarity may not be an issue as long as the quality of the sequencing is satisfactory, but it is nonetheless an inherent obstacle of the *rplF* sequencing approach.

While it is possible to identify cultured *Neisseria* isolates after extracting DNA, it proved more complicated to do so from DNA extracted from a throat swab when using Sanger sequencing. Several *rplF* variants, even from the same species, could be present in the swab, hindering sequencing reaction quality, as shown in Chapter 3 (Section 3.3.3). While there may not be a possibility of bypassing this problem when using Sanger sequencing, it is an issue that can be addressed if *rplF* gene is used in a NGS setting as presented here. However, the use of *rplF* in an NGS platform also carries another layer of biological complexity in the form of human and bacterial DNA co-occurring in the sample.

After comparing *rplF* primer binding region and amplicon sequences to those of other *Neisseriaceae* and upper respiratory tract commensals, there was evidence to suggest that only *Bergeriella, Simonsiella, Kingella* and *Conchiformibius* sequences could bind the rplF primers, although they would probably do so at a lower efficiency due to the dissimilarities. Despite this possibility, the difference in *rplF* amplicon sequences was sufficient to be confident that if sequences from these genera were amplified in our samples, they would not be wrongly classified as *Neisseria*. Evidence of this was seen after processing the sequencing reads from the rplF libraries, where some of the sequences identified belonged to *Simonsiella muellieri rplF* sequences, which were below 85% similarity to *Neisseria rplF* sequences.

Another issue that arose from the similarity of *Neisseria rplF* sequences was the almost complete identity of *M. cerebrosus* sequences to those of *N. mucosa*. While the analysis could suggest that *M. cerebrosus* may be *N. sicca/N. mucosa*, the analysis was insufficient to say so. While other publications have relied on the use of ANI scores to differentiate *Neisseria* species (44,191,192), it is also important to note that species identification based on nucleotide sequence methods may be hindered in highly

recombinogenic bacteria, such as *Neisseria*, as the high degree of recombination may lead to blurry lines delimiting species (193).While the ANI score method has been applied to other recombinogenic bacteria (194) with great success, no analysis has been performed to further validate its use with *Neisseria* sequences, as most rely on MLST/rMLST schemes to speciate *Neisseria*. *M. cerebrosus* was not identified during the microbiome analysis of either Dormice or Oxford samples, making it possible to dismiss the bacterium as a confounding factor on *Neisseria* population analyses.

The method used to classify *rplF* sequences also proved to be a key factor in the species assignment process. BLAST-like algorithms, whichrely on overall dissimilarity of sequences, may not be an optimal way of classifying *rplF* amplicons, as the differences seem to arise from changes in specific bases, rather than changes across the amplicon. The minimum entropy algorithm (99) (which is an extension of the oligotyping algorithm) focuses on the more entropic bases of an amplicon instead of considering it as a whole, providing a much higher accuracy when clustering *rplF* amplicons. Coupling a relatively non-standard clustering method, such as MED, in QIIME and later on in phyloseq, did not prove to be particularly challenging, and provided very reliable results. Using MED could also be beneficial for future studies, even if they rely solely on Sanger sequencing outputs, to improve its bioinformatics pipeline performance.

Another important factor pertinent to this topic is the reference sequences used to classify the *rplF* sequencing reads. For all the steps in the analysis, I used the sequences provided at the *Neisseria* PubMLST database. These sequences follow the guidelines proposed by Bennett *et al.* (52) for *Neisseria* species classification. In these, *N. sicca, N. flava* and *N. flavescens* were not considered species and their sequences were split between *N. subflava* and *N. mucosa*, and a proposed division of *N. polysaccharea* was considered a new species, named *N. bergeri.* The lack of consensus when

denoting *Neisseria* species, and the groups they belong to, could lead to mistakes in species assignment and spurious results. While Bennett *et al*. have made remarkable advances in terms of delimiting species (and genospecies) in the *Neisseria* genus (52,114), more research and validation is necessary in this particular field, as this is a limiting factor, and regardless of the method used to assign species to a particular sequence, it will always be, at most, as good as the quality of the reference sequences used.

9.1.2. Implication of the rplF primer mismatches

When comparing the published *rplF* primers to corresponding gene sequences, only *N. polysaccharea* and *N. meningitidis* matched the primers perfectly. Other human-colonizing *Neisseria* had one mismatch, *N. gonorrhoeae* had two, and animal-colonizing *Neisseria* had three or more. It was decided to compare the efficiency of the published primers to a pair including an ambiguous base in the 12th nucleotide (Y, which accounts for C and T). The analysis showed no significant differences among the results obtained when sequencing artificial bacterial mixtures. However, a slight trend was identified whereby *N. polysaccharea* and *N. meningitidis* are favoured when using the published *rplF* compared to the modified primers. The use of the original primers will not affect Sanger sequencing results on amplicons obtained from pure bacterial cultures, but it will lead to an overrepresentation of the species it binds to perfectly, if present, in studies such as the one presented here.

As the samples in this study were of paediatric origin, it was decided not to focus on the extra ambiguous base found in *N. gonorrhoeae,* located in the 8th base of the Forward primer. For future studies, especially if they are focused on adult populations, it would be advisable to use a Forward primer including both ambiguities, to avoid possible underrepresentation of this species. However, as

this mismatch is located further upstream than the one identified at the 12th base, it may not be as impactful. Another consideration for adult population studies is that, as opposed to other *Neisseria*, *N. gonorrhoeae* is an intracellular pathogen (195), and the method used to extract DNA from the sample may bias the species identification and should be optimized.

One final consideration regarding the primer sequences is that the optimized method may only be applicable to only those *Neisseria* species colonizing humans. As shown Chapter 3 (Section 3.3.4) *Neisseria* species usually only associated with animals, such as *N. dentiae, N. arctica, N. canis*, have at least three mismatches when compared to the original published primers. While this discordance may not be as impactful if the Sanger sequencing method is used on pure bacterial cultures, it will likely have a higher impact on NGS protocols such as the one presented in this thesis. This may mean that DNA from these species may not be amplified, or amplified with a very low efficiency, and will be missed or underrepresented in the final analysis.

9.2. Development of the rplF library preparation and data processing pipeline

The development of a reliable library preparation method was the most challenging part of this project. While the use of custom amplicon sequencing using the MiSeq is supported by Illumina, there is a lack of support towards the development of new applications in the platform. The use of the 16S rRNA gene as a marker has been supported by Illumina, as shown by the publication of their own protocol used to amplify variable regions and create MiSeq compatible libraries, although these rely on TruSeq kit usage (https://support.illumina.com/documents/documentation/chemistry_documentation/16s/16s-metagenomic-library-prep-guide-15044223-b.pdf). While it is

understandable that Illumina will encourage the use of their own kits, it may not be feasible to use TruSeq kits in large scale studies due to economic constraints.

For the development of *rplF* libraries, I based the first attempts on a single-step PCR approach, which had been used successfully in 16S rRNA gene sequencing experiments before (196). Despite this, the results obtained for all the libraries prepared using this method were unsatisfactory. When considering the overall results and trends observed when using the single-step method, it was first hypothesised that the issues may again be underlying lack of sequencing diversity in the *rplF* gene, and it was compensated initially by lowering the concentration of the library being added into the MiSeq flow-cell, and by increasing the concentration of PhiX control. This, however, only marginally helped the sequencing quality.

The use of the two-step PCR approach, and the later addition of the extra heat denaturing step of the final library-PhiX mixture, greatly improved the sequence quality, and none of the of the issues faced when using the single-step protocol resurfaced. While these results were adequate, they also do not fully align with the idea that the single-step PCR did not work as intended because of the lack of sequence diversity.

The proportion of PhiX used in the two-step approach (20% spike, 12.5 pM) was higher than the proportion used in most 16S rRNA gene library sequencing experiments (20% spike, 8 pM), and this was probably a contributing factor towards the good quality obtained. However, when using the two-step approach, the first bases sequenced will be identical for all DNA molecules, as the first

nucleotides sequenced belong to the PCR primers. Thus, it is possible to believe that the single-step protocol might have not failed due to the lack of sequence diversity.

There might be two possible explanations as to why the single-step PCR approach did not provide satisfactory results. One could be that the *rplF* amplicons bind to the flow cell with a higher affinity than PhiX, which could be at least partially true, as even with the two-step approach, a larger amount of PhiX was required. Another explanation could be that the custom sequencing primers, included only in the single-step approach, may have had unforeseen interactions with the PhiX DNA, leading to spurious sequence results. Proving either of these two hypotheses may be overly challenging, as there is no description of the proprietary Illumina buffer solutions used by the MiSeq instrument.

Regardless of possible explanations as to why the single-step approach failed, the two-step PCR approach was successful, providing high quality reads for all libraries prepared. Although the use of this protocol involves extra costs due to the added PCR and clean-up steps, it proved to be more reliable than the single-step PCR for *rplF* library generation purposes, and the primers used in the second step of the PCR can be reused in other MiSeq amplicon sequencing applications, as they only rely on the overlapping sequences added to the first amplicon.

9.3. Effect of Bexsero® on the infant upper respiratory tract microbiome

The comparison of the microbiome of infants from the Dormice and Oxford cohort showed a great degree of similarity among the proportions of species identified at comparable time points. When integrating both sample sets, two trends regarding *Neisseria* colonization were apparent:

- *Neisseria* carriage starts early on in life, colonizing almost 25% of infants by the 6-week time point, and steadily increases up to one year of life, when almost all infants seem to be colonized.
- The abundance of *Neisseria* reads drastically increased at the 9-month compared to previous time points and comprised 12-15% of the microbiome.

The analysis presented suggests that there is a small, but significant, change in the microbiome of Bexsero®-vaccinated infants when compared to non-vacinees. Due to the nature of the samples collected, it was not possible to determine if the effect was due to a difference in geographical location of the two sets of samples used, or was a real effect caused by vaccination. Due to the lack of a 12-month control group from Oxford infants it was not possible to assess this difference. Although both of these cities are relatively close, the infants could be exposed to different factors, such as different levels air pollution or crowding, which could lead to the subtle differences identified between these two sample sets. The lack of medical metadata collected for the Oxford cohort is one of the weaknesses of the analysis presented. Had it been available, it would have been possible to model the effect of Bexsero® to a higher level of detail.

Nevertheless, if we consider the differences seen when comparing vaccinated and unvaccinated infants as exclusively due to Bexsero® vaccination, it still a very small proportion of the microbiome that was affected, around 3% by the one-year time point, and may not pose a real effect in children's normal microbiome development.

9.4. Effect of Bexsero® on the infant upper respiratory tract Neisseria population

When comparing species colonizing the *Neisseria* niche of the upper respiratory tract microbiome, a similar degree of similarity was found for both Oxford and Dormice samples to that observed when comparing the microbiome data. Both sets showed that the majority of the *Neisseria* niche is formed by *N. subflava* and *N. cinerea* at all time points tested. It also showed that infants are colonized by a vast number of different *Neisseria* species in different proportions throughout the first three years of life.

Very little is known about the carriage of other *Neisseria* species in the paediatric population apart from *N. lactamica* and *N. meningitidis*. Diallo *et al.* (51) published data of a large study, over 7000 participants across different African countries, in which *Neisseria* species were identified using *rplF* sequences by plating oropharyngeal swabs in Thayer Martin media, followed by Sanger sequencing to identify isolated colonies. Their results showed that for infants in between 0 and 4 years, *N. lactamica* was the most carried isolate, followed by *N. meningitidis, N. polysaccharea, N. bergeri* and *N. subflava.* The data contrast with that reported in this thesis. For example, in contrast to this study Diallo *et al.* (51) found *N. subflava* had a very low carriage rate (below 0.5%), and *N. cinerea* was not present. This

could be due to the differences in populations studied (*e.g.* ethnicity or environment) but it could also point to one of the underlying problems with culture-based analysis, as not all species may grow as efficiently under laboratory conditions and may lead to skewed results. This is in turn one of the main advantages of using sequencing-based methods to assess carriage, as it can bypass those inherent problems.

When focusing on the differences in *Neisseria* species among vaccinated and unvaccinated infants, very few significant differences were identified, suggesting that vaccination with Bexsero® may not have a significant impact in *Neisseria* carriage. However, a potential side effect of the vaccine could be an increase in *N. lactamica*. This effect was identified when using differential abundance methods, and when comparing the sequencing data alongside the qPCR data, especially in those infants that received two doses of the vaccine. This could mean that Bexsero® may lead to an unforeseen secondary effect *i.e.* infants to become less prone to *N. meningitidis*, not necessarily by the immunological response caused by the vaccine, but by increasing *N. lactamica* colonization. Such an affect may be transient, as the difference seen is not as pronounced in infants that had received more than 2 doses.

9.5. Early life events and Neisseria populations

Infants showed different patterns of development in their *Neisseria* populations. As mentioned earlier, *N. subflava* and *N. cinerea* were present in all infants, although in vastly different proportions. However, there seemed to be a trend where three species, when present, tended to dominate the *Neisseria* population, *i.e. N. polysaccharea, N. lactamica* and *N. bergeri.* There seemed to be an

interplay among different *Neisseria* species in which they favoured or hindered the abundance and presence of others, both when considering co-colonization at the same time, and when considering the presence of other species at the previous time point. These data suggest that *Neisseria* species not only can co-habit in the upper respiratory tract, but can also, at least partially, modulate the dynamics of other species present.

N. meningitidis seemed to display a dual behaviour when considering its relation to *N. lactamica* and *N. polysaccharea*, both decreasing the likelihood of *N. meningitidis* being present at the next time point, but increasing meningococcal abundance if both co-occurred at the same time. This suggests that *N. meningitidis* is compromised in its ability to colonize in the presence of *N. lactamica* or *N. polysaccharea*, but if it succeeds in establishing itself in the upper respiratory tract, it may take advantage of the disrupted neisserial niche to increase its abundance.

Several life events were associated with changes in *Neisseria* populations, either during a period of time, or when considering the overall odds. Focusing on the interplay between *N. lactamica* and *N. meningitidis*, infants that were breastfed and lived with a pet in their household were less likely to carry *N. meningitidis*. However, due to the low numbers of infants carrying this bacterium it is difficult to determine the significance and impact of these results.

9.6. Future work

Samples used in the analysis presented are from two different cohorts (Dormice and Oxford). Further work should evaluate an independent cohort(s), not only to validate the *rplF* library preparation method and data sequencing pipeline, but also to ascertain the general applicability of the results.

To continue with the work here presented regarding *rplF* sequencing, it would be suggested to sequence the samples from the Dormice infants that could not be included in the analysis here presented. This would involve c. 10 extra MiSeq sequencing experiments. The new data would increase the number of samples included in the analysis, and its robustness, which will especially provide more robustness to claims regarding *N. meningitidis* carriage.

While the focus has been set on those infants that were colonized by *Neisseria* it could be on interest to determine what factors could be involved in the absence of *Neisseria* in the samples especially in between birth and the first year of life. However, this could prove difficult to do, as it would be necessary to determine the limit of detection of the sequencing protocol to determine a viable threshold to discriminate positive and negative samples.

Another point of focus would be to perform multivariate analyses including the metadata collected for the Dormice samples. More variables that have not been considered, such as the presence of viruses, antibiotics usage, upper respiratory tract infections or fever episodes could also be included in the analyses. Also, due to different collection times of samples in each cohort, slight

differences among the two groups, in particular for *N. meningitis,* could arise due to the changing incidence rates in the different years in which samples were collected (13). Furthermore, the season in which each sample was collected could also be considered, as incidence of *N. meningitis* has been shown to differ thoughout a given year (197).

N. bergeri was described by Bennet *et al.* as a divergent *N. polysaccharea* isolate (52). However, there has been no validation of *N. bergeri* as a separate species. The analysis here proposed could be re-processed considering *N. bergeri* reads as *N. polysaccharea*.

The data showed so far has been focused at the species level, however, a similar analysis could be performed using sequence variants instead of species as the taxonomic level considered for the analyses. It would be of great interest to determine the changes within single species using sequence variants as proxies for different isolates. This could help understand the dynamics of specific species and the metadata could be coupled to determine what effects lead to a higher or lower turnover of sequence variants.

References

1. Knapp JS. Historical perspectives and identification of *Neisseria* and related species. *Clinical microbiology reviews*. 1988;1(4): 415–431.

2. Read RC. *Neisseria meningitidis*; clones, carriage, and disease. *Clinical Microbiology and Infection*. [Online] 2014;20(5): 391–395. Available from: doi:10.1111/1469-0691.12647

3. Nieto JAS, Marcos C, Vindel A. Multicolonization of human nasopharynx due to *Neisseria spp*. *International Microbiology*. 1998;1(1): 59–63.

4. Liu G, Tang CM, Exley RM. Non-pathogenic *Neisseria*: members of an abundant, multi-habitat, diverse genus. *Microbiology (Reading, England)*. [Online] Microbiology Society; 2015;161(7): 1297–1312. Available from: doi:10.1099/mic.0.000086

5. Moxon ER, Jansen VAA. Phage variation: understanding the behaviour of an accidental pathogen. *Trends in microbiology*. [Online] 2005;13(12): 563–565. Available from: doi:10.1016/j.tim.2005.10.004

6. Olsen SF, Djurhuus B, Rasmussen K, Joensen HD, Larsen SO, Zoffman H, et al. Pharyngeal carriage of *Neisseria meningitidis* and *Neisseria lactamica* in households with infants within areas with high and low incidences of meningococcal disease. *Epidemiology and Infection*. Cambridge University Press; 1991;106(3): 445–1015.

7. Yazdankhah SP, Caugant DA. *Neisseria meningitidis*: an overview of the carriage state. *Journal of medical microbiology*. [Online] 2004;53(Pt 9): 821–832. Available from: doi:10.1099/jmm.0.45529-0

8. Oliver KJ, Reddin KM, Bracegirdle P, Hudson MJ, Borrow R, Feavers IM, et al. *Neisseria lactamica* protects against experimental meningococcal infection. *Infection and Immunity*. [Online] 2002;70(7): 3621–3626. Available from: doi:10.1128/IAI.70.7.3621-3626.2002

9. Li Y, Zhang Q, Winterbotham M, Mowe E, Gorringe A, Tang CM. Immunization with live *Neisseria lactamica* protects mice against meningococcal challenge and can elicit serum bactericidal antibodies. *Infection and Immunity*. [Online] American Society for Microbiology; 2006;74(11): 6348–6355. Available from: doi:10.1128/IAI.01062-06

10. Deasy AM, Guccione E, Dale AP, Andrews N, Evans CM, Bennett JS, et al. Nasal Inoculation of the Commensal *Neisseria lactamica* Inhibits Carriage of *Neisseria meningitidis* by Young Adults: A Controlled Human Infection Study. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America*. [Online] 2015;60(10): 1512–1520. Available from: doi:10.1093/cid/civ098

11. MacLennan J, Kafatos G, Neal K, Andrews N, Cameron JC, Roberts R, et al. Social Behavior and Meningococcal Carriage in British Teenagers. *Emerging infectious diseases*. [Online] Centers for Disease Control and Prevention; 2006;12(6): 950–957. Available from: doi:10.3201/eid1206.051297

12. Cleary PR, Calvert N, Gee S, Graham C, Gray S, Kaczmarski E, et al. Variations in *Neisseria meningitidis* carriage by socioeconomic status: a cross-sectional study. *Journal of public health (Oxford, England)*. [Online] 2015;38(1): 61–70. Available from: doi:10.1093/pubmed/fdv015

13. Halperin SA, Bettinger JA, Greenwood B, Harrison LH, Jelfs J, Ladhani SN, et al. The changing and dynamic epidemiology of meningococcal disease. *Vaccine*. [Online] 2012;30: B26–B36. Available from: doi:10.1016/j.vaccine.2011.12.032

14. Zemanick ET, Wagner BD, Robertson CE, Stevens MJ, Szefler SJ, Accurso FJ, et al. Assessment of airway microbiota and inflammation in cystic fibrosis using multiple sampling methods. *Annals of the American Thoracic Society*. [Online] 2015;12(2): 221–229. Available from: doi:10.1513/AnnalsATS.201407-310OC

15. Hu J, Iragavarapu S, Nadkarni GN, Huang R, Erazo M, Bao X, et al. Location-Specific Oral Microbiome Possesses Features Associated With CKD. *Kidney international reports*. [Online] 2018;3(1): 193–204. Available from: doi:10.1016/j.ekir.2017.08.018

16. Hoggard M, Biswas K, Zoing M, Wagner Mackenzie B, Taylor MW, Douglas RG. Evidence of microbiota dysbiosis in chronic rhinosinusitis. *International Forum of Allergy & Rhinology*. [Online] 2016;7(3): 230–239. Available from: doi:10.1002/alr.21871

17. Janda WM, Bradna JJ, Ruther P. Identification of *Neisseria spp., Haemophilus spp*., and other fastidious gram-negative bacteria with the MicroScan *Haemophilus-Neisseria* identification panel. *Journal of Clinical Microbiology*. 1989;27(5): 869–873.

18. Woo PCY, Lau SKP, Teng JLL, Tse H, Yuen KY. Then and now: use of 16S rDNA gene sequencing for bacterial identification and discovery of novel bacteria in clinical microbiology laboratories. *Clinical microbiology and infection : the official publication of the European Society of Clinical Microbiology and Infectious Diseases*. [Online] 2008;14(10): 908–934. Available from: doi:10.1111/j.1469-0691.2008.02070.x

19. Tibayrenc M, Ayala FJ. How clonal are *Neisseria* species? The epidemic clonality model revisited. *Proceedings of the National Academy of Sciences of the United States of America*. [Online] 2015;112(29): 8909–8913. Available from: doi:10.1073/pnas.1502900112

20. Brehony C, Jolley KA, Maiden MCJ. Multilocus sequence typing for global surveillance of meningococcal disease. *FEMS microbiology reviews*. [Online] Blackwell Publishing Ltd; 2007;31(1): 15–26. Available from: doi:10.1111/j.1574-6976.2006.00056.x

21. Fazio C, Neri A, Renna G, Vacca P, Antonetti R, Barbui AM, et al. Persistent occurrence of serogroup Y/sequence type (ST)-23 complex invasive meningococcal disease among patients aged five to 14 years, Italy, 2007 to 2013. *Euro surveillance : bulletin Européen sur les maladies transmissibles = European communicable disease bulletin*. [Online] 2015;20(45): e35699. Available from: doi:10.2807/1560-7917.ES.2015.20.45.30061

22. Maiden MCJ, Harrison OB. Population and Functional Genomics of *Neisseria* Revealed with Gene-by-Gene Approaches. Kraft CS (ed.) *Journal of Clinical Microbiology*. [Online] American Society for Microbiology; 2016;54(8): 1949–1955. Available from: doi:10.1128/JCM.00301-16

23. Jolley KA, Bliss CM, Bennett JS, Bratcher HB, Brehony C, Colles FM, et al. Ribosomal multilocus sequence typing: universal characterization of bacteria from domain to strain. *Microbiology*. [Online] Society for General Microbiology; 2012;158(Pt 4): 1005–1015. Available from: doi:10.1099/mic.0.055459-0

24. Pérez-Losada M, Cabezas P, Castro-Nallar E, Crandall KA. Pathogen typing in the genomics era: MLST and the future of molecular epidemiology. *Infection, Genetics and Evolution*. [Online] 2013;16: 38–53. Available from: doi:10.1016/j.meegid.2013.01.009

25. Chen Y, Frazzitta AE, Litvintseva AP, Fang C, Mitchell TG, Springer DJ, et al. Next generation multilocus sequence typing (NGMLST) and the analytical software program MLSTEZ enable efficient, cost-effective, high-throughput, multilocus sequencing typing. *Fungal genetics and biology : FG & B*. [Online] 2015;75: 64–71. Available from: doi:10.1016/j.fgb.2015.01.005

26. Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Lozupone CA, Turnbaugh PJ, et al. Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proceedings of the National Academy of Sciences of the United States of America*. [Online] National Acad Sciences; 2011;108 Suppl 1(Supplement_1): 4516–4522. Available from: doi:10.1073/pnas.1000080107

27. Lozupone CA, Stombaugh J, Gonzalez A, Ackermann G, Wendel D, Vazquez-Baeza Y, et al. Meta-analyses of studies of the human microbiota. *Genome Research*. [Online] 2013;23(10): 1704–1714. Available from: doi:10.1101/gr.151803.112

28. Earl CS, An S-Q, Ryan RP. The changing face of asthma and its relation with microbes. *Trends in microbiology*. [Online] 2015;23(7): 408–418. Available from: doi:10.1016/j.tim.2015.03.005

29. Marchesi JR, Adams DH, Fava F, Hermes GDA, Hirschfield GM, Hold G, et al. The gut microbiota and host health: a new clinical frontier. *Gut*. [Online] 2016;65(2): 330–339. Available from: doi:10.1136/gutjnl-2015-309990

30. Dickson RP, Martinez FJ, Huffnagle GB. The role of the microbiome in exacerbations of chronic lung diseases. *Lancet*. [Online] 2014;384(9944): 691–702. Available from: doi:10.1016/S0140-6736(14)61136-3

31. Si J, Lee C, Ko G. Oral Microbiota: Microbial Biomarkers of Metabolic Syndrome Independent of Host Genetic Factors. *Frontiers in Cellular and Infection Microbiology*. [Online] 2017;7: 516. Available from: doi:10.3389/fcimb.2017.00516

32. Sakwinska O, Bastic Schmid V, Berger B, Bruttin A, Keitel K, Lepage M, et al. Nasopharyngeal Microbiota in Healthy Children and Pneumonia Patients. *Journal of Clinical Microbiology*. [Online] 2014;52(5): 1590–1594. Available from: doi:10.1128/JCM.03280-13

33. Jiang W, Ling Z, Lin X, Chen Y, Zhang J, Yu J, et al. Pyrosequencing Analysis of Oral Microbiota Shifting in Various Caries States in Childhood. *Microbial Ecology*. [Online] Springer US; 2014;67(4): 962–969. Available from: doi:10.1007/s00248-014-0372-y

34. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome biology*. [Online] BioMed Central; 2014;15(3): R46. Available from: doi:10.1186/gb-2014-15-3-r46

35. Lassalle F, Spagnoletti M, Fumagalli M, Shaw L, Dyble M, Walker C, et al. Oral microbiomes from hunter-gatherers and traditional farmers reveal shifts in commensal balance and pathogen load linked to diet. *Molecular Ecology*. [Online] 2017. Available from: doi:10.1111/mec.14435

36. Donati C, Zolfo M, Albanese D, Tin Truong D, Asnicar F, Iebba V, et al. Uncovering oral *Neisseria* tropism and persistence using metagenomic sequencing. *Nature microbiology*. [Online] 2016;1(7): 16070. Available from: doi:10.1038/nmicrobiol.2016.70

37. Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, Gordon JI. The human microbiome project. *Nature*. [Online] 2007;449(7164): 804–810. Available from: doi:10.1038/nature06244

38. La Scola B, Gundi VAKB, Khamis A, Raoult D. Sequencing of the *rpoB* gene and flanking spacers for molecular identification of Acinetobacter species. *Journal of Clinical Microbiology*. [Online] 2006;44(3): 827–832. Available from: doi:10.1128/JCM.44.3.827-832.2006

39. Alexandre A, Laranjo M, Young JPW, Oliveira S. *dnaJ* is a useful phylogenetic marker for alphaproteobacteria. *International Journal of Systematic and Evolutionary Microbiology*. [Online] Society for General Microbiology; 2008;58(Pt 12): 2839–2849. Available from: doi:10.1099/ijs.0.2008/001636-0

40. Bennett JS, Watkins ER, Jolley KA, Harrison OB, Maiden MCJ. Identifying *Neisseria* species by use of the 50S ribosomal protein L6 (*rplF*) gene. *Journal of Clinical Microbiology*. [Online] American Society for Microbiology; 2014;52(5): 1375–1381. Available from: doi:10.1128/JCM.03529-13

41. Harrison OB, Bray JE, Maiden MCJ, Caugant DA. Genomic Analysis of the Evolution and Global Spread of Hyper-invasive Meningococcal Lineage 5. *EBioMedicine*. [Online] 2015;2(3): 234–243. Available from: doi:10.1016/j.ebiom.2015.01.004

42. The MenAfriCar consortium. Meningococcal carriage in the African meningitis belt. *Tropical Medicine & International Health*. [Online] 2nd ed. 2013;18(8): 968–978. Available from: doi:10.1111/tmi.12125

43. Volokhov DV, Amselle M, Bodeis-Jones S, Delmonte P, Zhang S, Davidson MK, et al. *Neisseria zalophi sp. nov.*, isolated from oral cavity of California sea lions (Zalophus californianus). *Archives of microbiology*. [Online] 2018. Available from: doi:10.1007/s00203-018-1499-x

44. Hansen CM, Himschoot EA, Hare RF, Meixell BW, Hemert CV, Hueffer K. *Neisseria arctica sp. nov.*, isolated from nonviable eggs of greater white-fronted geese (Anser albifrons) in Arctic Alaska. *International Journal of Systematic and Evolutionary Microbiology*. [Online] 2017;67(5): 1115–1119. Available from: doi:10.1099/ijsem.0.001773

45. Muzzi A, Mora M, Pizza M, Rappuoli R, Donati C. Conservation of meningococcal antigens in the genus *Neisseria*. *mBio*. [Online] American Society for Microbiology; 2013;4(3): e00163–13. Available from: doi:10.1128/mBio.00163-13

46. Lucidarme J, Gilchrist S, Newbold LS, Gray SJ, Kaczmarski EB, Richardson L, et al. Genetic Distribution of Noncapsular Meningococcal Group B Vaccine Antigens in *Neisseria lactamica*. *Clinical and Vaccine Immunology*. [Online] 2013;20(9): 1360–1369. Available from: doi:10.1128/CVI.00090-13

47. Putonti C, Nowicki B, Shaffer M, Fofanov Y, Nowicki S. Where does *Neisseria* acquire foreign DNA from: an examination of the source of genomic and pathogenic islands and the evolution of the Neisseria genus. *Bmc Evolutionary Biology*. [Online] 2013;13(1): 184. Available from: doi:10.1186/1471-2148-13-184

48. Harrison OB, Bennett JS, Derrick JP, Maiden MCJ, Bayliss CD. Distribution and diversity of the haemoglobin-haptoglobin iron-acquisition systems in pathogenic and non-pathogenic *Neisseria*. *Microbiology*. [Online] 2013;159(Pt 9): 1920–1930. Available from: doi:10.1099/mic.0.068874-0

49. Harrison OB, Maiden MC, Rokbi B. Distribution of transferrin binding protein B gene (*tbpB*) variants among Neisseria species. *BMC microbiology*. [Online] BioMed Central; 2008;8(1): 66. Available from: doi:10.1186/1471-2180-8-66

50. Clemence MEA, Maiden MCJ, Harrison OB. Characterization of capsule genes in non-pathogenic *Neisseria* species. *Microbial Genomics*. [Online] Microbiology Society; 2018;11: 589. Available from: doi:10.1099/mgen.0.000208

51. Diallo K, Trotter C, Timbine Y, Tamboura B, Sow SO, Issaka B, et al. Pharyngeal carriage of *Neisseria* species in the African meningitis belt. *Journal of Infection*. [Online] 2016;72(6): 667–677. Available from: doi:10.1016/j.jinf.2016.03.010

52. Bennett JS, Jolley KA, Earle SG, Corton C, Bentley SD, Parkhill J, et al. A genomic approach to bacterial taxonomy: an examination and proposed reclassification of species within the genus *Neisseria*. *Microbiology*. [Online] Microbiology Society; 2012;158(Pt 6): 1570–1580. Available from: doi:10.1099/mic.0.056077-0

53. Esposito S, Principi N. Vaccine profile of 4CMenB: a four-component *Neisseria meningitidis* serogroup B vaccine. *Expert Review of Vaccines*. [Online] 2014;13(2): 193–202. Available from: doi:10.1586/14760584.2014.874949

54. Xie O, Pollard AJ, Mueller JE, Norheim G. Emergence of serogroup X meningococcal disease in Africa: Need for a vaccine. *Vaccine*. [Online] 2013;31(27): 2852–2861. Available from: doi:10.1016/j.vaccine.2013.04.036

55. Pelton SI. The Global Evolution of Meningococcal Epidemiology Following the Introduction of Meningococcal Vaccines. *Journal of Adolescent Health*. [Online] Elsevier; 2016;59(2): S3–S11. Available from: doi:10.1016/j.jadohealth.2016.04.012

56. Knol MJ, Hahné SJM, Lucidarme J, Campbell H, de Melker HE, Gray SJ, et al. Temporal associations between national outbreaks of meningococcal serogroup W and C disease in the Netherlands and England: an observational cohort... - PubMed - NCBI. *The Lancet Public Health*. [Online] 2017;2(10): e473–e482. Available from: doi:10.1016/S2468-2667(17)30157-3

57. Garrido-Estepa M, León-Gómez I, Herruzo R, Cano R. Changes in meningococcal C epidemiology and vaccine effectiveness after vaccine introduction and schedule modification. *Vaccine*. [Online] 2014;32(22): 2604–2609. Available from: doi:10.1016/j.vaccine.2014.03.010

58. Hepkema H, Pouwels KB, van der Ende A, Westra TA, Postma MJ. Meningococcal Serogroup A, C, W135 and Y Conjugated Vaccine: A Cost-Effectiveness Analysis in the Netherlands. Borrow R (ed.) *PloS one*. [Online] 2013;8(5): e65036. Available from: doi:10.1371/journal.pone.0065036

59. Christensen H, May M, Bowen L, Hickman M, Trotter CL. Meningococcal carriage by age: a systematic review and meta-analysis. *The Lancet infectious diseases*. [Online] 2010;10(12): 853–861. Available from: doi:10.1016/S1473-3099(10)70251-6

60. Vernikos G, Medini D. Bexsero® chronicle. *Pathogens and global health*. [Online] 2014;108(7): 305–316. Available from: doi:10.1179/2047773214Y.0000000162

61. Panatto D, Amicizia D, Lai PL, Cristina ML, Domnich A, Gasparini R. New versus old meningococcal group B vaccines: how the new ones may benefit infants & toddlers. *The Indian journal of medical research*. 2013;138(6): 835–846.

62. Häyrinen J, Jennings H, Raff HV, Rougon G, Hanai N, Gerardy-Schahn R, et al. Antibodies to polysialic acid and its N-propyl derivative: binding properties and interaction with human embryonal brain glycopeptides. *The Journal of infectious diseases*. [Online] 1995;171(6): 1481–1490. Available from: doi:10.2307/30135570

63. van de Waterbeemd B, Mommen GPM, Pennings JLA, Eppink MH, Wijffels RH, van der Pol LA, et al. Quantitative Proteomics Reveals Distinct Differences in the Protein Content of Outer Membrane Vesicle Vaccines. *Journal of Proteome Research*. [Online] 2013;12(4): 1898–1908. Available from: doi:10.1021/pr301208g

64. Vaughan TE, Skipp PJ, O'Connor CD, Hudson MJ, Vipond R, Elmore MJ, et al. Proteomic analysis of Neisseria lactamica and N eisseria meningitidis outer membrane vesicle vaccine antigens. *Vaccine*. [Online] 2006;24(25): 5277–5293. Available from: doi:10.1016/j.vaccine.2006.03.013

65. Oster P, Lennon D, O'Hallahan J, Mulholland K, Reid S, Martin D. MeNZB (TM): a safe and highly immunogenic tailor-made vaccine against the New Zealand *Neisseria meningitidis* serogroup B disease epidemic strain. *Vaccine*. [Online] 2005;23(17-18): 2191–2196. Available from: doi:10.1016/j.vaccine.2005.01.063

66. van der Voort ER, van der Ley P, van der Biezen J, George S, Tunnela O, van Dijken H, et al. Specificity of human bactericidal antibodies against PorA P1.7,16 induced with a hexavalent meningococcal outer membrane vesicle vaccine. *Infection and Immunity*. 1996;64(7): 2745–2751.

67. Holst J, Feiring B, Fuglesang JE, Høiby EA, Nokleby H, Aaberge IS, et al. Serum bactericidal activity correlates with the vaccine efficacy of outer membrane vesicle vaccines against *Neisseria meningitidis* serogroup B disease. *Vaccine*. 2003;21(7-8): 734–737.

68. Boslego J, Garcia J, Cruz C, Zollinger W, Brandt B, Ruiz S, et al. Efficacy, safety, and immunogenicity of a meningococcal group B (15:P1.3) outer membrane protein vaccine

in Iquique, Chile. Chilean National Committee for Meningococcal Disease. *Vaccine*. 1995;13(9): 821–829.

69. Sierra GV, Campa HC, Varcacel NM, Garcia IL, Izquierdo PL, Sotolongo PF, et al. Vaccine against group B *Neisseria meningitidis*: protection trial and mass vaccination results in Cuba. *NIPH annals*. 1991;14(2): 195–207–discussion208–10.

70. Caron F, Châtelet du IP, Leroy J-P, Ruckly C, Blanchard M, Bohic N, et al. From tailor-made to ready-to-wear meningococcal B vaccines: longitudinal study of a clonal meningococcal B outbreak. *The Lancet infectious diseases*. [Online] 2011;11(6): 455–463. Available from: doi:10.1016/S1473-3099(11)70027-5

71. Michalik M, Djahanshiri B, Leo JC, Linke D. Reverse Vaccinology: The Pathway from Genomes and Epitope Predictions to Tailored Recombinant Vaccines. *Methods in molecular biology (Clifton, N.J.)*. [Online] 2016;1403: 87–106. Available from: doi:10.1007/978-1-4939-3387-7_4

72. Serruto D, Bottomley MJ, Ram S, Giuliani MM, Rappuoli R. The new multicomponent vaccine against meningococcal serogroup B, 4CMenB: immunological, functional and structural characterization of the antigens. - PubMed - NCBI. *Vaccine*. 2012;30: B87–B97.

73. Snape MD, Dawson T, Oster P, Evans A, John TM, Ohene-Kena B, et al. Immunogenicity of Two Investigational Serogroup B Meningococcal Vaccines in the First Year of Life A Randomized Comparative Trial. *Pediatric Infectious Disease Journal*. [Online] 2010;29(11): E71–E79. Available from: doi:10.1097/INF.0b013e3181f59f6d

74. Snape MD, Saroey P, John TM, Robinson H, Kelly S, Gossger N, et al. Persistence of bactericidal antibodies following early infant vaccination with a serogroup B meningococcal vaccine and immunogenicity of a preschool booster dose. *CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne*. [Online] Canadian Medical Association; 2013;185(15): E715–E724. Available from: doi:10.1503/cmaj.130257

75. Santolaya ME, O'Ryan M, Valenzuela MT, Prado V, Vergara RF, Muñoz A, et al. Persistence of antibodies in adolescents 18-24 months after immunization with one, two, or three doses of 4CMenB meningococcal serogroup B vaccine. *Human Vaccines & Immunotherapeutics*. 2013;9(11): 2304–2310.

76. Health DO. JCVI position statement on use of Bexsero® meningococcal B vaccine in the UK. 2014;: 1–13.

77. Medini D, Stella M, Wassil J. MATS: Global coverage estimates for 4CMenB, a novel multicomponent meningococcal B vaccine. *Vaccine*. [Online] 2015;33(23): 2629–2636. Available from: doi:10.1016/j.vaccine.2015.04.015

78. Abad R, Biolchi A, Moschioni M, Giuliani MM, Pizza M, Vázquez JA. A large portion of MATS negative meningococcal strains from Spain are killed by sera from adolescents and infants immunized with 4CMenB. Plotkin SA (ed.) *Clinical and Vaccine Immunology*. American Society for Microbiology; 2015;22(4): CVI.00669–14–360.

79. Frosi G, Biolchi A, Sapio Lo M, Rigat F, Gilchrist S, Lucidarme J, et al. Bactericidal antibody against a representative epidemiological meningococcal serogroup B panel confirms

that MATS underestimates 4CMenB vaccine strain coverage. *Vaccine*. [Online] 2013;31(43): 4968–4974. Available from: doi:10.1016/j.vaccine.2013.08.006

80. Parikh SR, Newbold L, Slater S, Stella M, Moschioni M, Lucidarme J, et al. Meningococcal serogroup B strain coverage of the multicomponent 4CMenB vaccine with corresponding regional distribution and clinical characteristics in England, Wales, and Northern Ireland, 2007-08 and 2014-15: a qualitative and quantitative assessment. *The Lancet infectious diseases*. [Online] 2017;17(7): 754–762. Available from: doi:10.1016/S1473-3099(17)30170-6

81. Abad R, Medina V, Stella M, Boccadifuoco G, Comanducci M, Bambini S, et al. Predicted Strain Coverage of a New Meningococcal Multicomponent Vaccine (4CMenB) in Spain: Analysis of the Differences with Other European Countries. Manganelli R (ed.) *PloS one*. [Online] Public Library of Science; 2016;11(3): e0150721. Available from: doi:10.1371/journal.pone.0150721

82. Parikh SR, Andrews NJ, Beebeejaun K, Campbell H, Ribeiro S, Ward C, et al. Effectiveness and impact of a reduced infant schedule of 4CMenB vaccine against group B meningococcal disease in England: a national observational cohort study. *The Lancet*. [Online] 2016;388(10061): 2775–2782. Available from: doi:10.1016/S0140-6736(16)31921-3

83. Christensen H, Trotter CL, Hickman M, Edmunds WJ. Re-evaluating cost effectiveness of universal meningitis vaccination (Bexsero) in England: modelling study. *British Medical Journal (Abstracts)*. [Online] BMJ Publishing Group Ltd; 2014;349: g5725–g5725. Available from: doi:10.1136/bmj.g5725

84. Christensen H, Trotter CL. Modelling the cost-effectiveness of catch-up 'MenB' (Bexsero) vaccination in England. *Vaccine*. [Online] 2017;35(2): 208–211. Available from: doi:10.1016/j.vaccine.2016.11.076

85. Abad R, Medina V, Fariñas MDC, Martínez-Martínez L, Bambini S, Dari A, et al. Potential impact of the 4CMenB vaccine on oropharyngeal carriage of *Neisseria meningitidis*. *Journal of Infection*. [Online] W.B. Saunders; 2017;75(6): 511–520. Available from: doi:10.1016/j.jinf.2017.09.021

86. Vipond C, Suker J, Jones C, Tang C, Feavers IM, Wheeler JX. Proteomic analysis of a meningococcal outer membrane vesicle vaccine prepared from the group B strain NZ98/254. *PROTEOMICS*. [Online] 2006;6(11): 3400–3413. Available from: doi:10.1002/pmic.200500821

87. Cooper LV, Boukary RM, Aseffa A, Mihret W, Collard J-M, Daugla D, et al. Investigation of correlates of protection against pharyngeal carriage of *Neisseria meningitidis* genogroups W and Y in the African meningitis belt. Hozbor DF (ed.) *PloS one*. [Online] 2017;12(8): e0182575. Available from: doi:10.1371/journal.pone.0182575

88. Baker M, McNicholas A, Garrett N, Jones N, Stewart J, Koberstein V, et al. Household crowding a major risk factor for epidemic meningococcal disease in Auckland children. *The Pediatric Infectious Disease Journal*. 2000;19(10): 983–990.

89. Basta NE, Berthe A, Keita M, Onwuchekwa U, Tamboura B, Traore A, et al. Meningococcal carriage within households in the african meningitis belt: a longitudinal

pilot study. *Journal of Infection*. [Online] 2017. Available from: doi:10.1016/j.jinf.2017.11.006

90. Iro MA, Snape MD, Voysey M, Jawad S, Finn A, Heath PT, et al. Persistence of bactericidal antibodies following booster vaccination with 4CMenB at 12, 18 or 24 months and immunogenicity of a fifth dose administered at 4years of age-a phase 3 extension to a randomised controlled trial. *Vaccine*. [Online] 2017;35(2): 395–402. Available from: doi:10.1016/j.vaccine.2016.11.009

91. Martínez-Porchas M, Villalpando-Canchola E, Vargas-Albores F. Significant loss of sensitivity and specificity in the taxonomic classification occurs when short 16S rRNA gene sequences are used. *Heliyon*. [Online] Elsevier; 2016;2(9): e00170. Available from: doi:10.1016/j.heliyon.2016.e00170

92. Sim K, Cox MJ, Wopereis H, Martin R, Knol J, Li M-S, et al. Improved Detection of Bifidobacteria with Optimised 16S rRNA-Gene Based Pyrosequencing. Ahmed N (ed.) *PloS one*. [Online] Public Library of Science; 2012;7(3): –e32543. Available from: doi:10.1371/journal.pone.0032543

93. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, et al. QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*. [Online] 2010;7(5): 335–336. Available from: doi:10.1038/nmeth.f.303

94. McMurdie PJ, Holmes S. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. Watson M (ed.) *PloS one*. [Online] Public Library of Science; 2013;8(4): e61217. Available from: doi:10.1371/journal.pone.0061217

95. Gaspar JM, Thomas WK. FlowClus: efficiently filtering and denoising pyrosequenced amplicons. *BMC bioinformatics*. [Online] 2015;16(1): 457. Available from: doi:10.1186/s12859-015-0532-1

96. Edgar RC. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*. [Online] 2010;26(19): 2460–2461. Available from: doi:10.1093/bioinformatics/btq461

97. Katoh K, Misawa K, Kuma K-I, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic acids research*. Oxford University Press; 2002;30(14): 3059–3066.

98. Price MN, Dehal PS, Arkin AP. FastTree 2--approximately maximum-likelihood trees for large alignments. Poon AFY (ed.) *PloS one*. [Online] Public Library of Science; 2010;5(3): e9490. Available from: doi:10.1371/journal.pone.0009490

99. Eren AM, Morrison HG, Lescault PJ, Reveillaud J, Vineis JH, Sogin ML. Minimum entropy decomposition: Unsupervised oligotyping for sensitive partitioning of high-throughput marker gene sequences. *Isme Journal*. [Online] 2014;9(4): 968–979. Available from: doi:10.1038/ismej.2014.195

100. Sedgwick P. Multiple hypothesis testing and Bonferroni's correction. *British Medical Journal (Abstracts)*. [Online] British Medical Journal Publishing Group; 2014;349(oct20 3): g6284–g6284. Available from: doi:10.1136/bmj.g6284

101. Bacteriologists SOA, Bergey DH. *Bergey's Manual of Systematic Bacteriology*. 2004.

102. Maiden MCJ, van Rensburg MJJ, Bray JE, Earle SG, Ford SA, Jolley KA, et al. MLST revisited: the gene-by-gene approach to bacterial genomics. *Nature Reviews Microbiology*. [Online] Nature Publishing Group; 2013;11(10): 728–736. Available from: doi:10.1038/nrmicro3093

103. Pritchard L, Glover RH, Humphris S, Elphinstone JG, Toth IK. Genomics and taxonomy in diagnostics for food security: soft-rotting enterobacterial plant pathogens. *Analytical Methods*. [Online] Royal Society of Chemistry; 2016;8(1): 12–24. Available from: doi:10.1039/C5AY02550H

104. Xie CH, Yokota A. Phylogenetic analysis of Alysiella and related genera of *Neisseriaceae*: Proposal of *Alysiella crassa comb. nov*., *Conchiformibium steedae gen. nov., comb. nov., Conchiformibium kuhniae sp nov* and *Bergeriella denitrificans gen. nov., comb. nov*. *Journal of General and Applied Microbiology*. 2005;51(1): 1–10.

105. Kimura M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of molecular evolution*. 1980;16(2): 111–120.

106. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *Journal of molecular biology*. [Online] 1990;215(3): 403–410. Available from: doi:10.1016/S0022-2836(05)80360-2

107. Biesbroek G, Tsivtsivadze E, Sanders EAM, Montijn R, Veenhoven RH, Keijser BJF, et al. Early respiratory microbiota composition determines bacterial succession patterns and respiratory health in children. *American Journal of Respiratory and Critical Care Medicine*. [Online] 2014;190(11): 1283–1292. Available from: doi:10.1164/rccm.201407-1240OC

108. Long PA, Sly LI, PHAM AV, DAVIS GHG. Characterization of *Morococcus cerebrosus* gen. nov., sp. nov. and Comparison with Neisseria mucosa. *International Journal of Systematic and Evolutionary Microbiology*. [Online] Microbiology Society; 1981;31(3): 294–301. Available from: doi:10.1099/00207713-31-3-294

109. Huse SM, Ye Y, Zhou Y, Fodor AA. A Core Human Microbiome as Viewed through 16S rRNA Sequence Clusters. Ahmed N (ed.) *PloS one*. [Online] 2012;7(6): e34242. Available from: doi:10.1371/journal.pone.0034242

110. Richter M, Rossello-Mora R. Shifting the genomic gold standard for the prokaryotic species definition. *Proceedings of the National Academy of Sciences*. [Online] 2009;106(45): 19126–19131. Available from: doi:10.1073/pnas.0906412106

111. Colston SM, Fullmer MS, Beka L, Lamy B, Gogarten JP, Graf J. Bioinformatic genome comparisons for taxonomic and phylogenetic assignments using Aeromonas as a test case. *mBio*. [Online] American Society for Microbiology; 2014;5(6): e02136. Available from: doi:10.1128/mBio.02136-14

112. MenAfriCar consortium. The Diversity of Meningococcal Carriage Across the African Meningitis Belt and the Impact of Vaccination With a Group A Meningococcal Conjugate Vaccine. *The Journal of infectious diseases*. [Online] 2015;212(8): 1298–1307. Available from: doi:10.1093/infdis/jiv211

113. Barth KR, Isabella VM, Clark VL. Biochemical and genomic analysis of the denitrification pathway within the genus Neisseria. *Microbiology*. [Online] 2009;155(12): 4093–4103. Available from: doi:10.1099/mic.0.032961-0

114. Bennett JS, Jolley KA, Maiden MCJ. Genome sequence analyses show that *Neisseria oralis* is the same species as *'Neisseria mucosa* var. heidelbergensis'. *International Journal of Systematic and Evolutionary Microbiology*. [Online] 2013;63(Pt 10): 3920–3926. Available from: doi:10.1099/ijs.0.052431-0

115. Gold R, Goldschneider I, Lepow ML, Draper TF, Randolph M. Carriage of *Neisseria meningitidis* and *Neisseria lactamica* in Infants and Children. *The Journal of infectious diseases*. [Online] Oxford University Press; 1978;137(2): 112–121. Available from: doi:10.1093/infdis/137.2.112

116. Kremastinou J, Tzanakaki G, Velonakis E, Voyiatzi A, Nickolaou A, Elton RA, et al. Carriage of *Neisseria meningitidis* and *Neisseria lactamica* among ethnic Greek school children from Russian immigrant families in Athens. *FEMS immunology and medical microbiology*. 1999;23(1): 13–20.

117. Kremastinou J, Tzanakaki G, Levidiotou S, Markou F, Themeli E, Voyiatzi A, et al. Carriage of *Neisseria meningitidis* and *Neisseria lactamica* in northern Greece. *FEMS immunology and medical microbiology*. [Online] Wiley/Blackwell (10.1111); 2003;39(1): 23–29. Available from: doi:10.1016/S0928-8244(03)00174-3

118. Cartwright KAV, Stuart JM, Jones DM, Noah ND. The Stonehouse Survey - Nasopharyngeal Carriage of Meningococci and *Neisseria lactamica*. *Epidemiology and Infection*. [Online] 1987;99(3): 591–601. Available from: doi:10.2307/3863308

119. Saez-Nieto JA, Dominguez JR, Monton JL, Cristobal P, Fenoll A, Vazquez J, et al. Carriage of *Neisseria meningitidis* and *Neisseria lactamica* in a school population during an epidemic period in Spain. *The Journal of hygiene*. Cambridge University Press; 1985;94(3): 279–288.

120. Bakir M, Yagci A, Ulger N, Akbenlioglu C, Ilki A, Soyletir G. Asymtomatic carriage of *Neisseria meningitidis* and *Neisseria lactamica* in relation to *Streptococcus pneumoniae* and *Haemophilus influenzae* colonization in healthy children: Apropos of 1400 children sampled. *European Journal of Epidemiology*. [Online] Kluwer Academic Publishers; 2001;17(11): 1015–1018. Available from: doi:10.1023/A:1020021109462

121. Evans CM, Pratt CB, Matheson M, Vaughan TE, Findlow J, Borrow R, et al. Nasopharyngeal Colonization by *Neisseria lactamica* and Induction of Protective Immunity against *Neisseria meningitidis*. *Clinical Infectious Diseases*. [Online] Oxford University Press; 2011;52(1): 70–77. Available from: doi:10.1093/cid/ciq065

122. He J, Tu Q, Ge Y, Qin Y, Cui B, Hu X, et al. Taxonomic and Functional Analyses of the Supragingival Microbiome from Caries-Affected and Caries-Free Hosts. *Microbial Ecology*. [Online] 5 ed. Springer US; 2017;75(2): 543–554. Available from: doi:10.1007/s00248-017-1056-1

123. Goleva E, Jackson LP, Harris JK, Robertson CE, Sutherland ER, Hall CF, et al. The Effects of Airway Microbiome on Corticosteroid Responsiveness in Asthma. *American Journal of Respiratory and Critical Care Medicine*. [Online] American Thoracic Society; 2013;188(10): 1193–1201. Available from: doi:10.1164/rccm.201304-0775OC

124. Guerrero-Preston R, Godoy-Vitorino F, Jedlicka A, Rodríguez-Hilario A, González H, Bondy J, et al. 16S rRNA amplicon sequencing identifies microbiota associated with oral cancer, human papilloma virus infection and surgical treatment. *Oncotarget*. [Online] 2016;7(32). Available from: doi:10.18632/oncotarget.9710

125. Kulski JK. Next-Generation Sequencing — An Overview of the History, Tools, and 'Omic' Applications. *Next Generation Sequencing - Advances, Applications and Challenges*. InTech; 2016.

126. Buermans HPJ, Dunnen den JT. Next generation sequencing technology: Advances and applications. *Biochimica et biophysica acta*. [Online] 2014;1842(10): 1932–1941. Available from: doi:10.1016/j.bbadis.2014.06.015

127. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews: Genetics*. [Online] Nature Publishing Group; 2016;17(6): 333–351. Available from: doi:10.1038/nrg.2016.49

128. Miya M, Sato Y, Fukunaga T, Sado T, Poulsen JY, Sato K, et al. MiFish, a set of universal PCR primers for metabarcoding environmental DNA from fishes: detection of more than 230 subtropical marine species. *Royal Society Open Science*. [Online] 2015;2(7): 150088. Available from: doi:10.1098/rsos.150088

129. Cornman RS, Otto CRV, Iwanowicz D, Pettis JS. Taxonomic Characterization of Honey Bee (*Apis mellifera*) Pollen Foraging Based on Non-Overlapping Paired-End Sequencing of Nuclear Ribosomal Loci. Jan E (ed.) *PloS one*. [Online] 2015;10(12): e0145365. Available from: doi:10.1371/journal.pone.0145365

130. Logan G, Freimanis GL, King DJ, Valdazo-Gonzalez B, Bachanek-Bankowska K, Sanderson ND, et al. A universal protocol to generate consensus level genome sequences for foot-and-mouth disease virus and other positive-sense polyadenylated RNA viruses using the Illumina MiSeq. *BMC genomics*. [Online] BioMed Central; 2014;15(1). Available from: doi:10.1186/1471-2164-15-828

131. Fadrosh DW, Ma B, Gajer P, Sengamalay N, Ott S, Brotman RM, et al. An improved dual-indexing approach for multiplexed 16S rRNA gene sequencing on the Illumina MiSeq platform. *Microbiome*. [Online] 2014;2(1): 6–6. Available from: doi:10.1186/2049-2618-2-6

132. Rhodes J, Beale MA, Fisher MC. Illuminating Choices for Library Prep: A Comparison of Library Preparation Methods for Whole Genome Sequencing of *Cryptococcus neoformans* Using Illumina HiSeq. Nielsen K (ed.) *PloS one*. [Online] Public Library of Science; 2014;9(11): e113501. Available from: doi:10.1371/journal.pone.0113501

133. Claesson MJ, Wang Q, O'Sullivan O, Greene-Diniz R, Cole JR, Ross RP, et al. Comparison of two next-generation sequencing technologies for resolving highly complex microbiota composition using tandem variable 16S rRNA gene regions. *Nucleic acids research*. [Online] 2nd ed. 2010;38(22): e200–e200. Available from: doi:10.1093/nar/gkq873

134. Luo C, Tsementzi D, Kyrpides N, Read T, Konstantinidis KT. Direct Comparisons of Illumina vs. Roche 454 Sequencing Technologies on the Same Microbial Community DNA Sample. Rodriguez-Valera F (ed.) *PloS one*. [Online] 2012;7(2): e30087. Available from: doi:10.1371/journal.pone.0030087

135. Mitra A, Skrzypczak M, Ginalski K, Rowicka M. Strategies for achieving high sequencing accuracy for low diversity samples and avoiding sample bleeding using illumina platform. Oudejans C (ed.) *PloS one*. [Online] 2015;10(4): e0120520. Available from: doi:10.1371/journal.pone.0120520

136. Shokralla S, Porter TM, Gibson JF, Dobosz R, Janzen DH, Hallwachs W, et al. Massively parallel multiplex DNA sequencing for specimen identification using an Illumina MiSeq platform. *Scientific Reports*. [Online] 2015;5. Available from: doi:10.1038/srep09687

137. Shore S, Henderson JM, Lebedev A, Salcedo MP, Zon G, McCaffrey AP, et al. Small RNA Library Preparation Method for Next-Generation Sequencing Using Chemical Modifications to Prevent Adapter Dimer Formation. *PloS one*. [Online] Public Library of Science; 2016;11(11): e0167009. Available from: doi:10.1371/journal.pone.0167009

138. Wu L, Wen C, Qin Y, Yin H, Tu Q, Van Nostrand JD, et al. Phasing amplicon sequencing on Illumina Miseq for robust environmental microbial community analysis. *BMC microbiology*. [Online] 2015;15(1): 125. Available from: doi:10.1186/s12866-015-0450-4

139. Meyer M, Kircher M. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harbor Protocols*. [Online] 2010;2010(6): pdb.prot5448–pdb.prot5448. Available from: doi:10.1101/pdb.prot5448

140. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, et al. Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities. *Applied and environmental microbiology*. [Online] 2009;75(23): 7537–7541. Available from: doi:10.1128/AEM.01541-09

141. Edgar RC. Updating the 97% identity threshold for 16S ribosomal RNA OTUs. Valencia A (ed.) *Bioinformatics*. [Online] 2018;16: 412. Available from: doi:10.1093/bioinformatics/bty113

142. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*. 2011;17(1): pp.10–pp.12.

143. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic acids research*. [Online] Oxford University Press; 2013;41(Database issue): D590–D596. Available from: doi:10.1093/nar/gks1219

144. DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, et al. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Applied and environmental microbiology*. [Online] American Society for Microbiology; 2006;72(7): 5069–5072. Available from: doi:10.1128/AEM.03006-05

145. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods*. [Online] 2016;13(7): 581–583. Available from: doi:10.1038/nmeth.3869

146. Eren AM, Borisy GG, Huse SM, Mark Welch JL. Oligotyping analysis of the human oral microbiome. *Proceedings of the National Academy of Sciences of the United States of America*. [Online] 2014;111(28): E2875–E2884. Available from: doi:10.1073/pnas.1409644111

147. Greenland S, Mansournia MA, Altman DG. Sparse data bias: a problem hiding in plain sight. *British Medical Journal (Abstracts)*. [Online] British Medical Journal Publishing Group; 2016;353. Available from: doi:10.1136/bmj.i1981

148. Balzer S, Malde K, Lanzén A, Sharma A, Jonassen I. Characteristics of 454 pyrosequencing data—enabling realistic simulation with flowsim. *Bioinformatics*. [Online] Oxford University Press; 2010;26(18): i420–i425. Available from: doi:10.1093/bioinformatics/btq365

149. Quince C, Lanzén A, Curtis TP, Davenport RJ, Hall N, Head IM, et al. Accurate determination of microbial diversity from 454 pyrosequencing data. *Nature Methods*. [Online] Nature Publishing Group; 2009;6(9): 639–641. Available from: doi:10.1038/nmeth.1361

150. Lozupone CA, Knight R. Species divergence and the measurement of microbial diversity. *FEMS microbiology reviews*. [Online] 2008;32(4): 557–578. Available from: doi:10.1111/j.1574-6976.2008.00111.x

151. Menke S, Wasimuddin, Meier M, Melzheimer JR, Mfune JKE, Heinrich S, et al. Oligotyping reveals differences between gut microbiomes of free-ranging sympatric Namibian carnivores (*Acinonyx jubatus*, *Canis mesomelas*) on a bacterial species-like level. *Frontiers in microbiology*. [Online] 2014;5: e51521. Available from: doi:10.3389/fmicb.2014.00526

152. Schmidt VT, Reveillaud J, Zettler E, Mincer TJ, Murphy L, Amaral-Zettler LA. Oligotyping reveals community level habitat selection within the genus Vibrio. *Frontiers in microbiology*. [Online] 2014;5(205): 361. Available from: doi:10.3389/fmicb.2014.00563

153. Girard C, Tromas N, Amyot M, Shapiro BJ. Gut Microbiome of the Canadian Arctic Inuit. Krajmalnik-Brown R (ed.) *mSphere*. [Online] 2017;2(1): e00297–16. Available from: doi:10.1128/mSphere.00297-16

154. Cardenas PA, Cooper PJ, Cox MJ, Chico M, Arias C, Moffatt MF, et al. Upper Airways Microbiota in Antibiotic-Naive Wheezing and Healthy Infants from the Tropics of Rural Ecuador. Krauss-Etschmann S (ed.) *PloS one*. [Online] Public Library of Science; 2012;7(10): –e46803. Available from: doi:10.1371/journal.pone.0046803

155. Teo SM, Mok D, Pham K, Kusel M, Serralha M, Troy N, et al. The Infant Nasopharyngeal Microbiome Impacts Severity of Lower Respiratory Infection and Risk of Asthma Development. *Cell Host & Microbe*. [Online] 2015;17(5): 704–715. Available from: doi:10.1016/j.chom.2015.03.008

156. Arrieta M-C, Stiemsma LT, Dimitriu PA, Thorson L, Russell S, Yurist-Doutsch S, et al. Early infancy microbial and metabolic alterations affect risk of childhood asthma. *Science Translational Medicine*. [Online] 2015;7(307): 307ra152–307ra152. Available from: doi:10.1126/scitranslmed.aab2271

157. Prevaes SMPJ, de Steenhuijsen Piters WAA, de Winter-de Groot KM, Janssens HM, Tramper-Stranders GA, Chu MLJN, et al. Concordance between upper and lower airway microbiota in infants with cystic fibrosis. *European Respiratory Journal*. [Online] 2017;49(3): 1602235. Available from: doi:10.1183/13993003.02235-2016

158. Bosch AATM, Levin E, van Houten MA, Hasrat R, Kalkman G, Biesbroek G, et al. Development of Upper Respiratory Tract Microbiota in Infancy is Affected by Mode of Delivery. *EBioMedicine*. [Online] 2016;9: 336–345. Available from: doi:10.1016/j.ebiom.2016.05.031

159. Biesbroek G, Bosch AATM, Wang X, Keijser BJF, Veenhoven RH, Sanders EAM, et al. The impact of breastfeeding on nasopharyngeal microbial communities in infants. *American Journal of Respiratory and Critical Care Medicine*. [Online] 2014;190(3): 298–308. Available from: doi:10.1164/rccm.201401-0073OC

160. Feazel LM, Santorico SA, Robertson CE, Bashraheil M, Scott JAG, Frank DN, et al. Effects of Vaccination with 10-Valent Pneumococcal Non-Typeable Haemophilus influenza Protein D Conjugate Vaccine (PHiD-CV) on the Nasopharyngeal Microbiome of Kenyan Toddlers. Borrow R (ed.) *PloS one*. [Online] 2015;10(6): e0128064. Available from: doi:10.1371/journal.pone.0128064

161. Biesbroek G, Wang X, Keijser BJF, Eijkemans RMJ, Trzciński K, Rots NY, et al. Seven-valent pneumococcal conjugate vaccine and nasopharyngeal microbiota in healthy children. *Emerging infectious diseases*. [Online] 2014;20(2): 201–210. Available from: doi:10.3201/eid2002.131220

162. Oksanen J, Blanchet FG, package RKR, 2011. vegan: Community ecology package. *worldagroforestry.org*

163. Paulson JN, Stine OC, Bravo HC, Pop M. Differential abundance analysis for microbial marker-gene surveys. *Nature Methods*. [Online] Nature Research; 2013;10(12): 1200–1202. Available from: doi:10.1038/nmeth.2658

164. McIver LJ, Abu-Ali G, Franzosa EA, Schwager R, Morgan XC, Waldron L, et al. bioBakery: a meta'omic analysis environment. *Bioinformatics*. [Online] 2018;34(7): 1235–1237. Available from: doi:10.1093/bioinformatics/btx754

165. Warton DI, Wright ST, Wang Y. Distance-based multivariate analyses confound location and dispersion effects. *Methods in Ecology and Evolution*. [Online] Blackwell Publishing Ltd; 2012;3(1): 89–101. Available from: doi:10.1111/j.2041-210X.2011.00127.x

166. Cardenas PA, Cooper PJ, Cox MJ, Chico M, Arias C, Moffatt MF, et al. Upper Airways Microbiota in Antibiotic-Naïve Wheezing and Healthy Infants from the Tropics of Rural Ecuador. Krauss-Etschmann S (ed.) *PloS one*. [Online] Public Library of Science; 2012;7(10): e46803. Available from: doi:10.1371/journal.pone.0046803

167. Tsolakos N, Lie K, Bolstad K, Maslen S, Kristiansen PA, Høiby EA, et al. Characterization of meningococcal serogroup B outer membrane vesicle vaccines from strain 44/76 after growth in different media. *Vaccine*. [Online] 2010;28(18): 3211–3218. Available from: doi:10.1016/j.vaccine.2010.02.023

168. Williams JN, Weynants V, Poolman JT, Heckels JE, Christodoulides M. Immuno-proteomic analysis of human immune responses to experimental *Neisseria meningitidis* outer membrane vesicle vaccines identifies potential cross-reactive antigens. *Vaccine*. [Online] 2014;32(11): 1280–1286. Available from: doi:10.1016/j.vaccine.2013.12.070

169. Sadarangani M, Pollard AJ. Can we control all-cause meningococcal disease in Europe? *Clinical Microbiology and Infection*. [Online] 2016;22: S103–S112. Available from: doi:10.1016/j.cmi.2016.03.006

170. McNamara LA, Shumate AM, Johnsen P, MacNeil JR, Patel M, Bhavsar T, et al. First Use of a Serogroup B Meningococcal Vaccine in the US in Response to a University Outbreak. *Pediatrics*. [Online] 2015;135(5): 798–804. Available from: doi:10.1542/peds.2014-4015

171. Sadarangani M, Sell T, Iro MA, Snape MD, Voysey M, Finn A, et al. Persistence of immunity after vaccination with a capsular group B meningococcal vaccine in 3 different toddler schedules. *CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne*. [Online] Canadian Medical Association Journal; 2017;189(41): E1276–E1285. Available from: doi:10.1503/cmaj.161288

172. Troncoso G, Sánchez S, Criado MT, Ferreirós CM. Analysis of *Neisseria lactamica* antigens putatively implicated in acquisition of natural immunity to Neisseria meningitidis. *FEMS immunology and medical microbiology*. [Online] Oxford University Press; 2002;34(1): 9–15. Available from: doi:10.1111/j.1574-695X.2002.tb00597.x

173. Sheikhi R, Amin M, Rostami S, Shoja S, Ebrahimi N. Oropharyngeal Colonization With *Neisseria lactamica*, Other Nonpathogenic *Neisseria* Species and *Moraxella catarrhalis* Among Young Healthy Children in Ahvaz, Iran. *Jundishapur Journal of Microbiology*. [Online] Kowsar; 2015;8(3). Available from: doi:10.5812/jjm.14813

174. Peterson ME, Mile R, Li Y, Nair H, Kyaw MH. Meningococcal carriage in high-risk settings: A systematic review. *International journal of infectious diseases : IJID : official publication of the International Society for Infectious Diseases*. [Online] 2018. Available from: doi:10.1016/j.ijid.2018.05.022

175. Zhang Y, Han SW, Cox LM, Li H. A multivariate distance-based analytic framework for microbial interdependence association test in longitudinal study. *Genetic Epidemiology*. [Online] Wiley-Blackwell; 2017;41(8): 769–778. Available from: doi:10.1002/gepi.22065

176. Hall MW, Rohwer RR, Perrie J, McMahon KD, Beiko RG. Ananke: temporal clustering reveals ecological dynamics of microbial communities. *PeerJ*. [Online] 2017;5(1): e3812. Available from: doi:10.7717/peerj.3812

177. Brown RG, Marchesi JR, Lee YS, Smith A, Lehne B, Kindinger LM, et al. Vaginal dysbiosis increases risk of preterm fetal membrane rupture, neonatal sepsis and is exacerbated by erythromycin. *BMC medicine*. [Online] BioMed Central; 2018;16(1): 9. Available from: doi:10.1186/s12916-017-0999-x

178. Gu C. Smoothing Spline ANOVA Models: R Package gss. *Journal of Statistical Software*. [Online] 2014;58(1): 1–25. Available from: doi:10.18637/jss.v058.i05

179. Ownby DR, Johnson CC, Peterson EL. Exposure to dogs and cats in the first year of life and risk of allergic sensitization at 6 to 7 years of age. *JAMA*. 2002;288(8): 963–972.

180. Fitzpatrick PE, Salmon RL, Hunter PR, Roberts RJ, Palmer SR. Risk factors for carriage of *Neisseria meningitidis* during an outbreak in Wales. *Emerging infectious diseases*. [Online] Centers for Disease Control and Prevention; 2000;6(1): 65–69. Available from: doi:10.3201/eid0601.000112

181. Hjuler T, Wohlfahrt J, Simonsen J, Kaltoft MS, Koch A, Kamper-Jørgensen M, et al. Perinatal and crowding-related risk factors for invasive pneumococcal disease in infants and young children: a population-based case-control study. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America*. [Online] 2007;44(8): 1051–1056. Available from: doi:10.1086/512814

182. Levine OS, Farley M, Harrison LH, Lefkowitz L, McGeer A, Schwartz B. Risk factors for invasive pneumococcal disease in children: a population-based case-control study in North America. *Pediatrics*. 1999;103(3): E28.

183. Lundbo LF, Benfield T. Risk factors for community-acquired bacterial meningitis. *Infectious Diseases*. [Online] 2017;49(6): 433–444. Available from: doi:10.1080/23744235.2017.1285046

184. Schmidt MA. *Characterization of the Molecular Epidemiology of Neisseria Meningitidis and Investigation of Two Potential Risk Factors Associated with Invasive Meningococcal Disease*. ProQuest; 2009. 1 p.

185. Stuart JM, Cartwright KA, Robinson PM, Noah ND. Effect of smoking on meningococcal carriage. *Lancet*. 1989;2(8665): 723–725.

186. Lee C-C, Middaugh NA, Howie SRC, Ezzati M. Association of secondhand smoke exposure with pediatric invasive bacterial disease and bacterial carriage: a systematic review and meta-analysis. Lanphear BP (ed.) *PLoS medicine*. [Online] 2010;7(12): e1000374. Available from: doi:10.1371/journal.pmed.1000374

187. Oh C, Lee K, Cheong Y, Lee S-W, Park S-Y, Song C-S, et al. Comparison of the Oral Microbiomes of Canines and Their Owners Using Next-Generation Sequencing. White BA (ed.) *PloS one*. [Online] Public Library of Science; 2015;10(7): e0131468. Available from: doi:10.1371/journal.pone.0131468

188. Tun HM, Konya T, Takaro TK, Brook JR, Chari R, Field CJ, et al. Exposure to household furry pets influences the gut microbiota of infant at 3-4 months following various birth scenarios. *Microbiome*. [Online] 2017;5(1): 40. Available from: doi:10.1186/s40168-017-0254-x

189. Ali O, Aseffa A, Omer AB, Lema T, Demissie TM, Tekletsion Y, et al. Household transmission of *Neisseria meningitidis* in the African meningitis belt: a longitudinal cohort study. *Lancet Global Health*. [Online] 2016;4(12): E989–E995. Available from: doi:10.1016/S2214-109X(16)30244-3

190. Wyllie AL, Pannekoek Y, Bovenkerk S, van Engelsdorp Gastelaars J, Ferwerda B, van de Beek D, et al. Sequencing of the variable region of *rpsB* to discriminate between Streptococcus pneumoniae and other streptococcal species. *Open Biology*. [Online] 2017;7(9): 170074. Available from: doi:10.1098/rsob.170074

191. Wroblewski D, Cole J, McGinnis J, Perez M, Wilson H, Mingle LA, et al. *Neisseria dumasiana sp. nov.* from human sputum and a dog's mouth. *International Journal of Systematic and Evolutionary Microbiology*. [Online] 2017;67(11): 4304–4310. Available from: doi:10.1099/ijsem.0.002148

192. Yi H, Chun J. *Neisseria weaveri* Andersen et al. 1993 is a later heterotypic synonym of *Neisseria weaveri* Holmes et al. 1993. *International Journal of Systematic and*

Evolutionary Microbiology. [Online] 2015;65(Pt 2): 463–464. Available from: doi:10.1099/ijs.0.070664-0

193. Hanage WP, Fraser C, Spratt BG. Fuzzy species among recombinogenic bacteria. *BMC biology*. [Online] 2005;3(1): 6. Available from: doi:10.1186/1741-7007-3-6

194. Urbanczyk H, Ogura Y, Hayashi T. Contrasting inter- and intraspecies recombination patterns in the 'Harveyi clade' vibrio collected over large spatial and temporal scales. *Genome biology and evolution*. [Online] 2014;7(1): 71–80. Available from: doi:10.1093/gbe/evu269

195. Weel JF, Hopman CT, van Putten JP. Bacterial entry and intracellular processing of *Neisseria gonorrhoeae* in epithelial cells: immunomorphological evidence for alterations in the major outer membrane protein P.IB. *The Journal of experimental medicine*. The Rockefeller University Press; 1991;174(3): 705–715.

196. Bartram AK, Lynch MDJ, Stearns JC, Moreno-Hagelsieb G, Neufeld JD. Generation of Multimillion-Sequence 16S rRNA Gene Libraries from Complex Microbial Communities by Assembling Paired-End Illumina Reads. *Applied and environmental microbiology*. [Online] American Society for Microbiology (ASM); 2011;77(11): 3846–3852. Available from: doi:10.1128/AEM.02772-10

197. Sinclair D, Préziosi M-P, John TJ, Greenwood B. The epidemiology of meningococcal disease in India. *Tropical Medicine & International Health*. [Online] John Wiley & Sons, Ltd (10.1111); 2010;15(12): 1421–1435. Available from: doi:10.1111/j.1365-3156.2010.02660.x

Appendix 1: R environment

R version 3.4.4 (2018-03-15)

Platform: x86_64-apple-darwin15.6.0 (64-bit)

Running under: macOS High Sierra 10.13.6

Matrix products: default

BLAS:
/System/Library/Frameworks/Accelerate.framework/Versions/A/Frameworks/vecLib.framework/Versions/A/libBLAS.dylib

LAPACK: /Library/Frameworks/R.framework/Versions/3.4/Resources/lib/libRlapack.dylib

locale:

[1] en_GB.UTF-8/en_GB.UTF-8/en_GB.UTF-8/C/en_GB.UTF-8/en_GB.UTF-8

attached base packages:

parallel	stats	graphics	grDevices
utils	datasets	methods	base

other attached packages:

cowplot_0.9.2	glmulti_1.0.7	rJava_0.9-9	reshape2_1.4.3	forcats_0.3.0	dplyr_0.7.4	purrr_0.2.4
readr_1.1.1	tidyr_0.8.0	tibble_1.4.2	ggplot2_2.2.1	tidyverse_1.2.1	stringr_1.3.1	dunn.test_1.3.5
phyloseq_1.22.3	vegan_2.5-1	lattice_0.20-35	permute_0.9-4	metagenomeSeq_1.20.1	RColorBrewer_1.1-2	glmnet_2.0-16
foreach_1.4.4	limma_3.34.9	Biobase_2.38.0	BiocGenerics_0.24.0	lmerTest_3.0-1	lme4_1.1-17	Matrix_1.2-14

loaded via a namespace (and not attached):

nlme_3.1-137	bitops_1.0-6	matrixStats_0.53.1	lubridate_1.7.4	httr_1.3.1	numDeriv_2016.8-1
tools_3.4.4	R6_2.2.2	KernSmooth_2.23-15	lazyeval_0.2.1	mgcv_1.8-23	colorspace_1.3-2
ade4_1.7-11	mnormt_1.5-5	curl_3.2	compiler_3.4.4	cli_1.0.0	rvest_0.3.2
xml2_1.2.0	caTools_1.17.1	scales_0.5.0	psych_1.8.4	foreign_0.8-70	minqa_1.2.4
rio_0.5.10	XVector_0.18.0	pkgconfig_2.0.1	rlang_0.2.0	readxl_1.1.0	rstudioapi_0.7
bindr_0.1.1	jsonlite_1.5	gtools_3.5.0	car_3.0-0	magrittr_1.5	biomformat_1.6.0
Rcpp_0.12.16	munsell_0.4.3	S4Vectors_0.16.0	ape_5.1	abind_1.4-5	stringi_1.2.2
yaml_2.1.19	carData_3.0-1	MASS_7.3-50	zlibbioc_1.24.0	rhdf5_2.22.0	gplots_3.0.1
plyr_1.8.4	grid_3.4.4	gdata_2.18.0	crayon_1.3.4	Biostrings_2.46.0	haven_1.1.1
splines_3.4.4	multtest_2.34.0	hms_0.4.2	knitr_1.20	pillar_1.2.2	igraph_1.2.1
codetools_0.2-15	stats4_3.4.4	glue_1.2.0	GUniFrac_1.1	data.table_1.10.4-3	modelr_0.1.2
nloptr_1.0.4	cellranger_1.1.0	gtable_0.2.0	assertthat_0.2.0	openxlsx_4.0.17	broom_0.4.4
survival_2.42-3	iterators_1.0.9	IRanges_2.12.0	bindrcpp_0.2.2	cluster_2.0.7-1	gss_2.1-8
RVAideMemoire_0.9-69-3					

Appendix 2: rplfV3 primer sequences

First step PCR primers

Name	Sequence
rplF_PCF_F	CTACACGACGCTCTTCCGATCTCAGTGACTGTTCCCGCTGGTGT
rplF_PCR_Fm	CTACACGACGCTCTTCCGATCTCAGTGACTGTTCCYGCTGGTGT
rplF_PCR_R	AGACGTGTGCTCTTCCGATCTAGGYTCAGGAGKWCGGAAHG

Second step PCR primers

rplF_P5_I1	AATGATACGGCGACCACCGAGATCTACACACTATCTGACACTCTTTCCCTACACGACGCTCTTCCGATCT
rplF_P5_I2	AATGATACGGCGACCACCGAGATCTACACTAGCGAGTACACTCTTTCCCTACACGACGCTCTTCCGATCT
rplF_P5_I3	AATGATACGGCGACCACCGAGATCTACACCGTGAGTGACACTCTTTCCCTACACGACGCTCTTCCGATCT
rplF_P5_I4	AATGATACGGCGACCACCGAGATCTACACGGATATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT
rplF_P5_I5	AATGATACGGCGACCACCGAGATCTACACCTACTATAACACTCTTTCCCTACACGACGCTCTTCCGATCT
rplF_P5_I6	AATGATACGGCGACCACCGAGATCTACACTACGAGACACACTCTTTCCCTACACGACGCTCTTCCGATCT
rplF_P5_I7	AATGATACGGCGACCACCGAGATCTACACACGTCTCGACACTCTTTCCCTACACGACGCTCTTCCGATCT
rplF_P5_I8	AATGATACGGCGACCACCGAGATCTACACGTCAGATAACACTCTTTCCCTACACGACGCTCTTCCGATCT
rplF_P7_I1	CAAGCAGAAGACGGCATACGAGATTAAGGCGAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
rplF_P7_I2	CAAGCAGAAGACGGCATACGAGATCGTACTAGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
rplF_P7_I3	CAAGCAGAAGACGGCATACGAGATAGGCAGAAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
rplF_P7_I4	CAAGCAGAAGACGGCATACGAGATTCCTGAGCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
rplF_P7_I5	CAAGCAGAAGACGGCATACGAGATGGACTCCTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
rplF_P7_I6	CAAGCAGAAGACGGCATACGAGATTAGGCATGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT

rplF_P7_I7	CAAGCAGAAGACGGCATACGAGATCTCTCTACGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
rplF_P7_I8	CAAGCAGAAGACGGCATACGAGATCGAGGCTGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
rplF_P7_I9	CAAGCAGAAGACGGCATACGAGATAAGAGGCAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
rplF_P7_I10	CAAGCAGAAGACGGCATACGAGATGTAGAGGAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
rplF_P7_I11	CAAGCAGAAGACGGCATACGAGATGCTCATGAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
rplF_P7_I12	CAAGCAGAAGACGGCATACGAGATATCTCAGGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT

Appendix 3: List of isolate sequences used in chapter 2

Isolates used for Neisseriaceae rplF sequence comparison

ACCESSION	ORGANISM
CP019644	*Vitreoscilla sp*. C1
CP022423	*Vitreoscilla filiformis*
FMAP01000006	*Snodgrassella sp*. R-53583
CP007446	*Snodgrassella alvi* wkB2
JAIK01000358	*Snodgrassella alvi* SCGC AB-598-O11
JAIL01000249	*Snodgrassella alvi* SCGC AB-598-O02
AVQL01000455	*Snodgrassella alvi* SCGC AB-598-J21
NXEN01000153	*Snodgrassella alvi*
NAHQ01000127	*Snodgrassella alvi*
NAHP01000071	*Snodgrassella alvi*
NAHO01000038	*Snodgrassella alvi*
NAHN01000064	*Snodgrassella alvi*
NAHM01000061	*Snodgrassella alvi*
NAHL01000083	*Snodgrassella alvi*
NAHK01000111	*Snodgrassella alvi*
NAHJ01000047	*Snodgrassella alvi*
NAHI01000066	*Snodgrassella alvi*
NAHH01000060	*Snodgrassella alvi*
NAHG01000147	*Snodgrassella alvi*
NAHF01000092	*Snodgrassella alvi*
NAHE01000091	*Snodgrassella alvi*
NAHD01000068	*Snodgrassella alvi*
NAHC01000084	*Snodgrassella alvi*
NAHB01000055	*Snodgrassella alvi*
NAHA01000088	*Snodgrassella alvi*
NAGZ01000122	*Snodgrassella alvi*
NAGY01000063	*Snodgrassella alvi*
NAGX01000210	*Snodgrassella alvi*
NAGW01000068	*Snodgrassella alvi*
MVDP01000063	*Snodgrassella alvi*
MEIX01000024	*Snodgrassella alvi*
MEIW01000114	*Snodgrassella alvi*
MEIV01000030	*Snodgrassella alvi*
MEIU01000059	*Snodgrassella alvi*

MEIT01000088	*Snodgrassella alvi*
MEIS01000122	*Snodgrassella alvi*
MEIR01000043	*Snodgrassella alvi*
MEIQ01000045	*Snodgrassella alvi*
MEIP01000001	*Snodgrassella alvi*
MEIO01000062	*Snodgrassella alvi*
MEIN01000006	*Snodgrassella alvi*
MEIM01000015	*Snodgrassella alvi*
MEIL01000016	*Snodgrassella alvi*
MEIK01000029	*Snodgrassella alvi*
MEIJ01000028	*Snodgrassella alvi*
MEII01000027	*Snodgrassella alvi*
MDVJ01000027	*Snodgrassella alvi*
MDVI01000028	*Snodgrassella alvi*
MDVH01000029	*Snodgrassella alvi*
MDVG01000068	*Snodgrassella alvi*
MDVF01000001	*Snodgrassella alvi*
MDVE01000034	*Snodgrassella alvi*
MDVD01000054	*Snodgrassella alvi*
MDVC01000107	*Snodgrassella alvi*
MDVB01000004	*Snodgrassella alvi*
MDVA01000026	*Snodgrassella alvi*
MDUZ01000109	*Snodgrassella alvi*
MDUY01000041	*Snodgrassella alvi*
JFZW01000004	*Snodgrassella alvi*
JFZV01000004	*Snodgrassella alvi*
NZ_JH815305	*Simonsiella muelleri* ATCC 29453
JH815305	*Simonsiella muelleri*
CP019448	*Simonsiella muelleri*
ADCY02000043	*Simonsiella muelleri*
CP024847	*Neisseriaceae* bacterium DSM 100970
JUFZ01000070	*Morococcus cerebrosus*
ACJW02000003	*Kingella oralis* ATCC 51147
NZ_LT841145	*Kingella negevensis*
FXUV02000076	*Kingella negevensis*
NZ_HG427488	*Kingella kingae* Vir5453
NZ_JH621344	*Kingella kingae* PYKK081
JH621344	*Kingella kingae*
AJGB01000062	*Kingella kingae*
NZ_HG427487	*Kingella kingae* PV1748
NZ_HG428640	*Kingella kingae* KK411
NZ_HG427448	*Kingella kingae* KK274
NZ_HG426454	*Kingella kingae* KK260

NZ_HG426455	*Kingella kingae* KK245
NZ_HG427447	*Kingella kingae* KK242
NZ_HG424701	*Kingella kingae* KK238
NZ_HG428422	*Kingella kingae* KK199
NZ_HG426452	*Kingella kingae* KK189
NZ_HG427203	*Kingella kingae* KK171
NZ_HG424981	*Kingella kingae* KK158
NZ_HG427200	*Kingella kingae* KK156
NZ_HG427199	*Kingella kingae* KK145
NZ_HG424700	*Kingella kingae* KK141
NZ_HG427764	*Kingella kingae* KK128
NZ_HG427486	*Kingella kingae* KK113
NZ_HG427485	*Kingella kingae* KK107
NZ_HG424696	*Kingella kingae* KK101
NZ_HG424703	*Kingella kingae* KK97
NZ_HG424702	*Kingella kingae* KK93
NZ_HG424697	*Kingella kingae* KK88
NZ_HG427198	*Kingella kingae* KK83
NZ_HG427197	*Kingella kingae* KK75
NZ_HG427763	*Kingella kingae* KK60
NZ_HG426961	*Kingella kingae* KK56
NZ_HG426962	*Kingella kingae* KK12
NZ_HG426718	*Kingella kingae* KK6
NZ_HG426717	*Kingella kingae* KK3
NZ_HG426453	*Kingella kingae* D7330
NZ_HG426451	*Kingella kingae* D2363
NZ_HG424704	*Kingella kingae* CC254
NZ_FR904249	*Kingella kingae* BB631
NZ_HG422736	*Kingella kingae* BB114
NZ_HG422735	*Kingella kingae* BB060
NZ_HG422584	*Kingella kingae* BB016
NZ_GL891963	*Kingella kingae* ATCC 23330
GL891963	*Kingella kingae*
AFHS01000058	*Kingella kingae*
NZ_HG421939	*Kingella kingae* AA417
NZ_HG421938	*Kingella kingae* AA255
NZ_HG421937	*Kingella kingae* AA105
NZ_HG421936	*Kingella kingae* AA068
NZ_JH768600	*Kingella kingae* 11220434
LN869922	*Kingella kingae*
NZ_GL870929	*Kingella denitrificans* ATCC 33394
GL870929	*Kingella denitrificans*
AEWV01000024	*Kingella denitrificans*

LXSQ01000003	*Eikenella sp.* NML130454
LXSP01000003	*Eikenella sp*. NML120348
LXSO01000004	*Eikenella sp.* NML080894
LXSN01000004	*Eikenella sp*. NML070372
LXST01000002	*Eikenella sp*. NML99-0057
LXSS01000010	*Eikenella sp*. NML97-A-109
LXSR01000015	*Eikenella sp*. NML96-A-049
LXSM01000017	*Eikenella sp.* NML03-A-027
LXSL01000032	*Eikenella sp*. NML02-A-017
LXSK01000019	*Eikenella sp*. NML01-A-086
NZ_KV820997	*Eikenella sp*. HMSC073A11
LTST01000064	*Eikenella sp*. HMSC073A11
KV820997	*Eikenella sp*. HMSC073A11
NZ_KV804547	*Eikenella sp*. HMSC071B05
LTGG01000050	*Eikenella sp*. HMSC071B05
KV804547	*Eikenella sp*. HMSC071B05
LTKM01000056	*Eikenella sp*. HMSC061C02
NZ_KI635563	*Eikenella corrodens* CC92I
KI635563	*Eikenella corrodens*
AZGQ01000010	*Eikenella corrodens*
NZ_EQ973320	*Eikenella corrodens*
EQ973320	*Eikenella corrodens*
ACEA01000034	*Eikenella corrodens*
NHRS01000001	*Eikenella corrodens*
LXSJ01000018	*Eikenella corrodens*
LXSI01000001	*Eikenella corrodens*
LXSH01000020	*Eikenella corrodens*
LXSG01000026	*Eikenella corrodens*
LXSF01000001	*Eikenella corrodens*
LXSE01000030	*Eikenella corrodens*
LT906482	*Eikenella corrodens*
LQQU01000029	*Crenobacter luteus*
LAQT01000038	*Amantichitinum ursilacus*
OCNF01000015	*Alysiella filiformis* DSM 16848

List of isolates used for upper respiratory tact commensals rplF sequence comparison

Accession	Organism	Accession	Organism
AFUJ01000053	*Veillonella sp.* oral taxon 780 str. F0422	**CAZL010000417**	*Prevotella sp*. CAG:1092
AENU01000007	*Veillonella sp*. oral taxon 158 str. F0412	**NOVE01000002**	*Prevotella sp*. 885
JAPX01000009	*Veillonella sp*. ICM51a	**LFQU01000023**	*Prevotella sp*. 109
ATCG01000022	*Veillonella s*p. HPA0037	**FQWA01000013**	*Prevotella scopos* JCM 17725
AZML01000415	*Veillonella sp*. DORA_B_18_19_23	**CP016204**	*Prevotella scopos* JCM 17725
LSDQ01000004	*Veillonella sp*. DNF00869	**AWGW01000006**	*Prevotella salivae* F0493
CAWP010000109	*Veillonella sp*. CAG:933	**AEQO01000115**	*Prevotella salivae* DSM 15606
AZYX01000008	*Veillonella sp*. AS16	**AMEP01000141**	*Prevotella saccharolytica* F0055
ALKL01000025	*Veillonella sp*. ACP1	**CP002006**	*Prevotella ruminicola* 23
AHAF01000003	*Veillonella seminalis* ACS-216-V-Col6b	**AWET01000004**	*Prevotella pleuritidis* F0068
ASKE01000015	*Veillonella parvula* HSIVP1	**AFPY01000099**	*Prevotella pallens* ATCC 700821
CP001820	*Veillonella parvula* DSM 2008	**ADGI01000056**	*Prevotella oulorum* F0390
ADFU01000013	*Veillonella parvula* ATCC 17745	**FUXK01000027**	*Prevotella oulorum*
AEXI01000014	*Veillonella parvula* ACS-068-V-Sch12	**ACUZ02000009**	*Prevotella oris* F0302
LSDP01000022	*Veillonella parvula*	**ATFN01000009**	*Prevotella oralis* HGA0225
LOJO01000001	*Veillonella parvula*	**AZJG01000012**	*Prevotella oralis* CC98A
JRNT01000042	*Veillonella montpellierensis* DNF00314	**AEPE02000005**	*Prevotella oralis* ATCC 33269
PPDD01000008	*Veillonella infantium*	**ADGJ01000006**	*Prevotella nigrescens* F0103
AZMJ01000527	*Veillonella dispar* DORA_11	**AZJH01000020**	*Prevotella nigrescens* CC14M
ACIK02000004	*Veillonella dispar* ATCC 17748	**AFPX01000029**	*Prevotella nigrescens* ATCC 33563
LSDO01000012	*Veillonella dispar*	**NHRW01000001**	*Prevotella nigrescens*
AMEX01000057	*Veillonella atypica* KON	**AEWX01000018**	*Prevotella multiformis* DSM 16608
AEDS01000068	*Veillonella atypica* ACS-134-V-Col7a	**AGWK01000046**	*Prevotella micans* F0438
AEDR01000028	*Veillonella atypica* ACS-049-V-Sch6	**JRNS01000507**	*Prevotella melaninogenica* DNF00666
LRQT01000021	*Veillonella atypica*	**CP002122**	*Prevotella melaninogenica* ATCC 25845
CP020566	*Veillonella atypica*	**CP022040**	*Prevotella melaninogenica*
AEKO01000002	*Streptococcus vestibularis* F0396	**AP018049**	*Prevotella melaninogenica*
AEVI01000004	*Streptococcus vestibularis* ATCC 49124	**AEEI01000061**	*Prevotella marshii* DSM 16973 = JCM 13450

CTEN01000008	*Streptococcus varani*	**AGEK01000028**	*Prevotella maculosa* OT 289
AHAA01000006	*Streptococcus urinalis* FB127-CNA-2	**JNGW01000022**	*Prevotella loescheii* DSM 19665 = JCM 12249 = ATCC 15930
AEUZ02000001	*Streptococcus urinalis* 2285-97	**FZNZ01000034**	*Prevotella jejuni*
JATH01000002	*Streptococcus uberis* S6261	**ATMK01000007**	*Prevotella intermedia* ZT
JANW01000004	*Streptococcus uberis* EF20/0145	**CP019300**	*Prevotella intermedia* ATCC 25611 = DSM 20706
JATJ01000002	*Streptococcus uberis* C9359	**CP003503**	*Prevotella intermedia 17*
JATG01000015	*Streptococcus uberis* C8329	**PGGD01000001**	*Prevotella intermedia*
JATA01000004	*Streptococcus uberis* C6344	**PESN01000001**	*Prevotella intermedia*
JATF01000003	*Streptococcus uberis* C5388	**PENH01000003**	*Prevotella intermedia*
JATI01000003	*Streptococcus uberis* C5072	**AP014925**	*Prevotella intermedia*
JATC01000001	*Streptococcus uberis* B362	**AP014597**	*Prevotella intermedia*
JATE01000004	*Streptococcus uberis* B190	**JRNJ01000062**	Prevotella histicola JCM 15637 = DNF00424
JATB01000006	*Streptococcus uberis* 6736	**AFXP01000001**	*Prevotella histicola* F0411
CP022435	*Streptococcus uberis*	**CP012074**	*Prevotella fusca* JCM 17724
AP014612	*Streptococcus troglodytae*	**CP013195**	*Prevotella enoeca*
CP000023	*Streptococcus thermophilus* LMG 18311	**AWUY01000040**	*Prevotella disiens* JCM 6334 = ATCC 29426
AGFN01000272	*Streptococcus thermophilus* CNCM I-1630	**AEDO01000051**	*Prevotella disiens* FB035-09AN
CBBT010000004	*Streptococcus thermophilus* CAG:236	**JRNR01000052**	*Prevotella disiens* DNF00882
AZTK01000057	*Streptococcus thermophilus* 1F8CT	**CP002589**	*Prevotella denticola* F0289
PHHK01000002	*Streptococcus thermophilus*	**JRNO01000017**	*Prevotella denticola* DNF00960
PHHJ01000002	*Streptococcus thermophilus*	**AEXO01000056**	*Prevotella denticola* CRIS 18C-A
CP006246	*Streptococcus suis* T15	**CP003368**	*Prevotella dentalis* DSM 3688
CP002640	*Streptococcus suis* SS12	**AFPW01000006**	*Prevotella dentalis* DSM 3688
CP003922	*Streptococcus suis* SC070731	**LRQG01000001**	*Prevotella corporis*
CP002465	*Streptococcus suis* JS14	**ACBX02000037**	*Prevotella copri* DSM 18205
CP002644	*Streptococcus suis* D12	**CBCH010000187**	*Prevotella copri* CAG:164
CP017142	*Streptococcus suis*	**NMPZ01000010**	*Prevotella copri*
CP015557	*Streptococcus suis*	**CZAM01000443**	*Prevotella copri*
CP012911	*Streptococcus suis*	**JRNP01000031**	*Prevotella buccalis* DNF00985
CP012731	*Streptococcus suis*	**JRNN01000095**	*Prevotella buccalis* DNF00853
CP011419	*Streptococcus suis*	**ADEG01000086**	*Prevotella buccalis* ATCC 35310
ASKH01000061	*Streptococcus sp.* HSISS3	**AEPD01000006**	*Prevotella buccae* ATCC 33574
ASKC01000164	*Streptococcus sp*. HSISS2	**FOEM01000017**	*Prevotella bryantii* B14
LTXE01000056	*Streptococcus sp*. HMSC072C09	**ADWO01000092**	*Prevotella bryantii* B14
LTXK01000045	*Streptococcus sp*. HMSC071H03	**NPJF01000020**	*Prevotella bryantii*

LTEU01000119	*Streptococcus sp.* HMSC071D03	**FNQQ01000008**	*Prevotella bryantii*
LTLE01000105	*Streptococcus sp*. HMSC070B10	**FNHC01000007**	*Prevotella bryantii*
LTLK01000076	*Streptococcus sp.* HMSC36C04	**ADFO01000009**	*Prevotella bivia* JCVIHMP010
LTYL01000086	Streptococcus *sp.* HMSC034F03	**JRNM01000047**	*Prevotella bivia* DNF00650
LTRW01000082	*Streptococcus sp*. HMSC034F02	**JRNQ01000017**	*Prevotella bivia* DNF00320
NTGM02000001	*Streptococcus sp*. FDAARGOS_256	**JRNF01000006**	*Prevotella bivia* DNF00188
CP020431	*Streptococcus sp.* FDAARGOS_192	**LTAG01000035**	*Prevotella bivia*
LVJM01000046	*Streptococcus sp*. CCUG 49591	**LRQF02000106**	*Prevotella bivia*
MOWR01000013	Streptoc*occus sp*. 'caviae'	**ACKS01000085**	*Prevotella bergensis* DSM 17361
ACOI02000011	*Streptococcus sp.* 2_1_36FAA	**AWEY01000011**	*Prevotella baroniae* F0067
AWVA01000047	*Streptococcus sobrinus* W1703	**JRNU01000042**	*Prevotella amnii* DNF00058
AOCE01000006	*Streptococcus sobrinus* DSM 20742 = ATCC 33478	**ADFQ01000036**	*Prevotella amnii* CRIS 21A-A
NPOU01000001	*Streptococcus sobrinus*	**LSDL01000011**	*Prevotella amnii*
JPEN01000097	*Streptococcus sinensis*	**FOIQ01000003**	*Prevotella aff*. ruminicola Tc2-24
AEXV01000016	*Streptococcus sanguinis* SK72	**19024_M10_240480**	*Neisseria species*
AFFO01000002	*Streptococcus sanguinis* SK49	**19024_M10_240480**	*Neisseria species*
PKHZ01000004	*Streptococcus salivarius*	**NKHK01000032**	*Moraxella sp*. VT-16-12
NXMB01000025	*Streptococcus salivarius*	**LTKA01000013**	*Moraxella sp*. HMSC061H09
NPDJ01000016	*Streptococcus salivarius*	**MUYV01000012**	*Moraxella porci* DSM 25326
LRQS02000008	*Streptococcus salivarius*	**MUYU01000029**	*Moraxella pluranimalium*
JYOY01000023	*Streptococcus salivarius*	**NCXF01000002**	*Moraxella osloensis*
AOCD01000015	*Streptococcus ratti* FA-1 = DSM 20564	**LZMT01000023**	*Moraxella osloensis*
AJTZ01000004	*Streptococcus ratti* FA-1 = DSM 20564	**LZDQ01000012**	*Moraxella osloensis*
CP000003	*Streptococcus pyogenes* MGAS10394	**LZDO01000075**	*Moraxella osloensis*
CP000056	*Streptococcus pyogenes* MGAS6180	**CP024443**	*Moraxella osloensis*
CP012045	*Streptococcus pyogenes*	**CP024185**	*Moraxella osloensis*
CP011535	*Streptococcus pyogenes*	**CP024180**	*Moraxella osloensis*
CP011068	*Streptococcus pyogenes*	**CP024176**	*Moraxella osloensis*
CP010449	*Streptococcus pyogenes*	**CP014234**	*Moraxella osloensis*
CP009612	*Streptococcus pyogenes*	**AP017381**	*Moraxella osloensis*
AP014596	*Streptococcus pyogenes*	**ANIN01000001**	*Moraxella macacae* 0408225
AENS01000008	*Streptococcus pseudoporcinus* SPIN 20026	**MUYT01000015**	*Moraxella lincolnii*
AEUY02000005	*Streptococcus pseudoporcinus* LQ 940-04	**MXAN01000055**	*Moraxella lacunata*

AJKE01000066	*Streptococcus pseudopneumoniae* SK674	**LZMS01000060**	*Moraxella lacunata*
MWSM01000007	*Streptococcus pseudopneumoniae* ATCC BAA-960 = CCUG 49455	**LZDR01000033**	*Moraxella lacunata*
AICS01000190	*Streptococcus pseudopneumoniae* ATCC BAA-960 = CCUG 49455	**MXAP01000099**	*Moraxella equi*
AYRO01000002	*Streptococcus pseudopneumoniae* 22725	**FTNU01000012**	*Moraxella cuniculi* DSM 21768
LJHI01000065	*Streptococcus pseudopneumoniae*	**MUYC01000010**	*Moraxella cuniculi*
HE983624	*Streptococcus pneumoniae* SPNA45	**MUXU01000018**	*Moraxella caviae*
CP000936	*Streptococcus pneumoniae* Hungary19A-6	**AMSO01000008**	*Moraxella catarrhalis* RH4
CP000936	*Streptococcus pneumoniae* Hungary19A-6	**AERL01000036**	*Moraxella catarrhalis* O35E
CP001033	*Streptococcus pneumoniae* CGSP14	**AERK01000003**	*Moraxella catarrhalis* CO72
FM211187	*Streptococcus pneumoniae* ATCC 700669	**AERJ01000003**	*Moraxella catarrhalis* BC8
CP002176	*Streptococcus pneumoniae* 670-6B	**AERI01000026**	*Moraxella catarrhalis* BC7
CP025536	*Streptococcus pluranimalium*	**AERH01000031**	*Moraxella catarrhalis* BC1
JSAP01000063	*Streptococcus phocae subsp. salmonis*	**CP002005**	*Moraxella catarrhalis* BBH18
LHQM01000001	*Streptococcus phocae*	**AERC01000010**	*Moraxella catarrhalis* 7169
AEVF01000001	*Streptococcus peroris* ATCC 700780	**AERE01000009**	*Moraxella catarrhalis* 103P14B1
LOCM01000004	*Streptococcus penaeicida*	**AEPC01000015**	*Moraxella catarrhalis* 101P30B1
LSRA01000066	*Streptococcus pasteurianus*	**AERF01000011**	*Moraxella catarrhalis* 46P47B1
ALWR01000004	*Streptococcus parauberis* KRS-02109	**AERG01000021**	*Moraxella catarrhalis* 12P80B1
ALYM01000010	*Streptococcus parauberis* KRS-02083	**CP007669**	*Moraxella catarrhalis*
NSGS01000010	*Streptococcus parauberis*	**MUXT01000009**	*Moraxella canis*
NSGR01000006	*Streptococcus parauberis*	**AOMT01000021**	*Moraxella bovoculi* 237
MLCP01000007	*Streptococcus parauberis*	**CP011381**	*Moraxella bovoculi*
LHAC01000005	*Streptococcus parauberis*	**CP011380**	*Moraxella bovoculi*
CP025420	*Streptococcus parauberis*	**CP011379**	*Moraxella bovoculi*
BAWT01000013	*Streptococcus parauberis*	**CP011378**	*Moraxella bovoculi*
AFUC01000002	*Streptococcus parasanguinis* SK236	**CP011377**	*Moraxella bovoculi*
AJMV01000078	*Streptococcus parasanguinis* F0449	**CP011376**	*Moraxella bovoculi*
AEKM01000005	*Streptococcus parasanguinis* F0405	**CP011374**	*Moraxella bovoculi*
AZMG01001391	*Streptococcus parasanguinis* DORA_23_24	**MUXV01000031**	*Moraxella bovis*
AZJD01000010	*Streptococcus parasanguinis* CC87K	**MXAO01000068**	*Moraxella atlantae*

AEVE01000058	*Streptococcus parasanguinis* ATCC 903	**LZNA01000023**	*Moraxella atlantae*
PKIC01000008	*Streptococcus parasanguinis*	**LZMZ01000030**	*Moraxella atlantae*
LQNY01000402	*Streptococcus parasanguinis*	**ALJP01000011**	*Haemophilus sputorum* HK 2154
JYPA01000005	*Streptococcus parasanguinis*	**AGRK01000030**	*Haemophilus sp*. oral taxon 851 str. F0397
AORU01000008	*Streptococcus oralis subsp. tigurinus* AZ_3a	**CP027235**	Haemophilus *sp.* oral taxon 036
NCUH01000009	*Streptococcus oralis subsp. tigurinus*	**LTVN01000001**	*Haemophilus sp*. HMSC073C03
LNVH01000002	*Streptococcus oralis subsp. tigurinus*	**LTOU01000012**	*Haemophilus sp*. HMSC071C11
LNVG01000001	*Streptococcus oralis subsp. tigurinus*	**LTFQ01000068**	*Haemophilus sp.* HMSC066A11
LNVF01000002	*Streptococcus oralis subsp. tigurinus*	**LZOX01000011**	*Haemophilus sp.* CCUG 66565
NCUV01000023	*Streptococcus oralis subsp. oralis*	**LDVZ01000009**	*Haemophilus sp*. C1
NCUK01000012	*Streptococcus oralis subsp. oralis*	**MDJC01000070**	*Haemophilus quentini*
NCUJ01000017	*Streptococcus oralis subsp. oralis*	**MDJB01000058**	*Haemophilus quentini*
PNFQ01000003	*Streptococcus oralis subsp. dentisani*	**MCII02000004**	*Haemophilus quentini*
PKIG01000001	*Streptococcus oralis subsp. dentisani*	**AFUV01000010**	*Haemophilus pittmaniae* HK 85
PKIF01000002	*Streptococcus oralis subsp. dentisani*	**LT906463**	*Haemophilus pittmaniae*
NCUW01000028	*Streptococcus oralis subsp. dentisani*	**AJMU01000021**	*Haemophilus paraphrohaemolyticus* HK411
ALJN01000022	*Streptococcus oralis* SK304	**MUXX01000004**	*Haemophilus paraphrohaemolyticus*
AFNM01000055	*Streptococcus oralis* SK255	**FQ312002**	*Haemophilus parainfluenzae* T3T1
AJKP01000051	*Streptococcus oralis* SK100	**AJTC01000028**	*Haemophilus parainfluenzae* HK2019
AJKO01000016	*Streptococcus oralis* SK10	**AJMW01000062**	*Haemophilus parainfluenzae* HK262
AEPO01000010	*Streptococcus oralis* ATCC 49296	**AFQS01000009**	*Haemophilus parainfluenzae* ATCC 33392
AEDW01000009	*Streptococcus oralis* ATCC 35037	**AEWU01000008**	*Haemophilus parainfluenzae* ATCC 33392
ADMV01000003	*Streptococcus oralis* ATCC 35037	**MPJK01000017**	*Haemophilus parainfluenzae*
MUXP01000005	*Streptococcus oralis*	**MPJJ01000007**	*Haemophilus parainfluenzae*
MBDM01000008	*Streptococcus oralis*	**MAQD01000005**	*Haemophilus parainfluenzae*
LWCF01000001	*Streptococcus oralis*	**LZMY01000012**	*Haemophilus parainfluenzae*
LQZP01000238	*Streptococcus oralis*	**LZMX01000028**	*Haemophilus parainfluenzae*
LQZE01000056	*Streptococcus oralis*	**LFXN01000002**	*Haemophilus parainfluenzae*
LQZB01000079	*Streptococcus oralis*	**AJSW01000011**	*Haemophilus parahaemolyticus* HK385
LQRI01000151	*Streptococcus oralis*	**MUXY01000007**	*Haemophilus parahaemolyticus*

LQOG01000032	*Streptococcus oralis*	**MUYA01000013**	*Haemophilus paracuniculus*
LQOC01000007	*Streptococcus oralis*	**L42023**	*Haemophilus influenzae* Rd KW20
LQOB01000281	*Streptococcus oralis*	**ACSM01000001**	*Haemophilus influenzae* RdAW
LQNZ01000118	*Streptococcus oralis*	**AAZE01000003**	*Haemophilus influenzae* R3021
LQNX01000067	*Streptococcus oralis*	**CP002277**	*Haemophilus influenzae* R2866
AHSO01000019	*Streptococcus mutans* W6	**CP002276**	*Haemophilus influenzae* R2846
CP007016	*Streptococcus mutans* UA159-FR	**AAZI01000001**	*Haemophilus influenzae* PittII
AHSB01000014	*Streptococcus mutans* U138	**AAZH01000004**	*Haemophilus influenzae* PittHH
AGWE01000011	*Streptococcus mutans* U2B	**CP000672**	*Haemophilus influenzae* PittGG
AHSU01000005	*Streptococcus mutans* U2A	**CP000671**	*Haemophilus influenzae* PittEE
AHSE01000031	*Streptococcus mutans* T4	**AAZG01000004**	*Haemophilus influenzae* PittAA
AHST01000015	*Streptococcus mutans* ST6	**ACSL01000001**	*Haemophilus influenzae* NT127
AHTA01000012	*Streptococcus mutans* 66-2A	**CP005967**	*Haemophilus influenzae* KR494
AHTE01000008	*Streptococcus mutans* 24	**ABFC01001211**	*Haemophilus influenzae* HK1212
AHSZ01000034	*Streptococcus mutans* 21	**FQ670204**	*Haemophilus influenzae* F3047
AHRR01000036	*Streptococcus mutans* 15VF2	**FQ670178**	*Haemophilus influenzae* F3031
AHRJ01000003	*Streptococcus mutans* 15JP3	**CP007805**	*Haemophilus influenzae* CGSHiCZ412602
AHSY01000026	*Streptococcus mutans* 14D	**JXCI01000014**	*Haemophilus influenzae* 60294N1
AHRT01000064	*Streptococcus mutans* 11VS1	**FQ312006**	*Haemophilus influenzae* 10810
BDOS01000001	*Streptococcus mutans*	**CP009610**	*Haemophilus influenzae*
AFQV01000006	*Streptococcus mitis* SK1080	**CP007472**	*Haemophilus influenzae*
AICR01000017	*Streptococcus mitis* SK616	**CP007471**	*Haemophilus influenzae*
AEDV01000006	*Streptococcus mitis* SK597	**CP007470**	*Haemophilus influenzae*
AJJL01000076	*Streptococcus mitis* SK579	**AFQR01000015**	*Haemophilus haemolyticus* M21639
AICU01000075	*Streptococcus mitis* SK575	**AFQQ01000001**	*Haemophilus haemolyticus* M21621
AFUF01000020	*Streptococcus mitis* SK569	**AFQP01000022**	*Haemophilus* haemolyticus M21127
AEDU01000019	*Streptococcus mitis* SK564	**AFQO01000016**	*Haemophilus haemolyticus* M19501
AEDT01000032	*Streptococcus mitis* SK321	**AFQN01000015**	*Haemophilus haemolyticus* M19107
AEDX01000018	*Streptococcus mitis* NCTC 12261	**AJSV01000006**	*Haemophilus haemolyticus* HK386
AFUB01000055	*Streptococcus mitis* bv. 2 str. SK95	**LZOY01000008**	*Haemophilus haemolyticus*
ATAB01000001	*Streptococcus mitis* 29/42	**LZDL01000027**	*Haemophilus haemolyticus*
AYRS01000002	*Streptococcus mitis* 27/7	**LYCK01000013**	*Haemophilus haemolyticus*
AYRR01000009	*Streptococcus mitis* 21/39	**LCTK01000003**	*Haemophilus haemolyticus*
ATAA01000005	*Streptococcus mitis* 18/56	**LCTJ01000063**	*Haemophilus haemolyticus*
AQTU01000008	*Streptococcus mitis* 13/39	**LCTI01000044**	*Haemophilus haemolyticus*

PKIE01000002	*Streptococcus mitis*	**LCTH01000033**	*Haemophilus haemolyticus*
LQOA01000042	*Streptococcus mitis*	**MUXZ01000001**	*Haemophilus haemoglobinophilus*
HE613569	*Streptococcus macedonicus* ACA-DC 198	**AFBC01000021**	*Haemophilus aegyptius* ATCC 11116
PKIB01000016	*Streptococcus macedonicus*	**LZPH01000031**	*Haemophilus aegyptius*
PEBN01000013	*Streptococcus macedonicus*	**LAYS01000081**	*Corynebacterium xerosis*
PEBM01000006	*Streptococcus macedonicus*	**FWFI01000029**	*Corynebacterium xerosis*
PEBL01000022	*Streptococcus macedonicus*	**CP004353**	*Corynebacterium vitaeruminis* DSM 20294
JNCV01000005	*Streptococcus macedonicus*	**CP002917**	*Corynebacterium variabile* DSM 44702
FLZS01000001	*Streptococcus macedonicus*	**FAUH01000001**	*Corynebacterium variabile*
AEUW02000001	*Streptococcus macacae* NCTC 11558	**CP011546**	*Corynebacterium uterequi*
CP003025	*Streptococcus lutetiensis* 033	**CP009215**	*Corynebacterium ureicelerivorans*
LQOI01000136	*Streptococcus lutetiensis*	**CP004085**	*Corynebacterium urealyticum* DSM 7111
AFXN01000014	*Streptococcus intermedius* F0395	**AM942444**	*Corynebacterium urealyticum* DSM 7109
ANFT01000022	*Streptococcus intermedius* BA1	**LT906481**	*Corynebacterium urealyticum*
PNRP01000004	*Streptococcus intermedius*	**AYUJ01000002**	*Corynebacterium ulcerans* NCTC 12077
PNRI01000006	*Streptococcus intermedius*	**CP011913**	*Corynebacterium ulcerans* FRC58
PNRH01000004	*Streptococcus intermedius*	**CP009622**	*Corynebacterium ulcerans* FRC11
AOCT02000048	*Streptococcus iniae* IUSA1	**CP002791**	*Corynebacterium ulcerans* BR-AD22
AMOO01000001	*Streptococcus iniae* 9117	**CP002790**	*Corynebacterium ulcerans* 809
MNAC01000014	*Streptococcus iniae*	**AP012284**	*Corynebacterium ulcerans* 0102
LXQF01000103	*Streptococcus iniae*	**MIOS01000030**	*Corynebacterium ulcerans*
LSNB01000040	*Streptococcus iniae*	**LT906443**	*Corynebacterium ulcerans*
CP024843	*Streptococcus iniae*	**LJVH01000015**	*Corynebacterium ulcerans*
CP022392	*Streptococcus iniae*	**LGSY02000015**	*Corynebacterium ulcerans*
CP017952	*Streptococcus iniae*	**LGSX01000004**	*Corynebacterium ulcerans*
AEDY01000032	*Streptococcus infantis* SK1302	**JZUT01000001**	*Corynebacterium ulcerans*
AFNN01000024	*Streptococcus infantis* SK1076	**JZUS01000001**	*Corynebacterium ulcerans*
AJTA01000017	*Streptococcus infantis* ATCC 700779	**CP021417**	*Corynebacterium ulcerans*
LQZF01000165	*Streptococcus infantis*	**CP011095**	*Corynebacterium ulcerans*
ABJK02000013	*Streptococcus infantarius subsp. infantarius* ATCC BAA-102	**CP010818**	*Corynebacterium ulcerans*
AEUX02000003	*Streptococcus ictaluri* 707-05	**CP009716**	*Corynebacterium ulcerans*
JRYN01000008	*Streptococcus hongkongensis*	**CP009583**	*Corynebacterium ulcerans*
LQWV01000003	*Streptococcus gordonii*	**CP009500**	*Corynebacterium ulcerans*

LQRC01000246	*Streptococcus gordonii*	**JRNG01000020**	*Corynebacterium tuscaniense* DNF00037
LOBS01000052	*Streptococcus gordonii*	**ACVP01000015**	*Corynebacterium tuberculostearicum* SK141
JYOZ01000002	*Streptococcus gordonii*	**CP011545**	*Corynebacterium testudinoris*
JYGN01000004	*Streptococcus gordonii*	**CP003696**	*Corynebacterium terpenotabidum* Y-11
CP017295	*Streptococcus gordonii*	**ACGE01000074**	*Corynebacterium striatum* ATCC 6940
AEEM01000002	*Streptococcus gallolyticus subsp. gallolyticus* TX20005	**LTBH01000024**	*Corynebacterium striatum*
MAMV01000010	*Streptococcus gallolyticus subsp. gallolyticus*	**LTBG01000006**	*Corynebacterium striatum*
LXFC01000013	*Streptococcus gallolyticus subsp. gallolyticus*	**LTBF01000036**	*Corynebacterium striatum*
JYKU01000001	*Streptococcus gallolyticus subsp. gallolyticus*	**LAYR01000018**	*Corynebacterium striatum*
LQXV01000407	*Streptococcus gallolyticus*	**CTEG01000029**	*Corynebacterium striatum*
LQOF01000407	*Streptococcus gallolyticus*	**CTEG01000012**	*Corynebacterium striatum*
FPBN01000010	*Streptococcus gallolyticus*	**CP024932**	*Corynebacterium striatum*
FOLZ01000007	*Streptococcus gallolyticus*	**CP024932**	*Corynebacterium striatum*
CWFG01000002	*Streptococcus equi subsp. equi*	**CP024931**	*Corynebacterium striatum*
CWCZ01000001	*Streptococcus equi subsp. equi*	**CP024931**	*Corynebacterium striatum*
CWBZ01000001	*Streptococcus equi subsp. equi*	**CP021252**	*Corynebacterium striatum*
CWBV01000001	*Streptococcus equi subsp. equi*	**CP019963**	*Corynebacterium stationis*
CWBT01000002	*Streptococcus equi subsp. equi*	**CP016326**	*Corynebacterium stationis*
CWBE01000001	*Streptococcus equi subsp. equi*	**CP014279**	*Corynebacterium stationis*
CVZV01000001	*Streptococcus equi subsp. equi*	**CP009251**	*Corynebacterium stationis*
CVZE01000002	*Streptococcus equi subsp. equi*	**FOPJ01000001**	*Corynebacterium spheniscorum*
CVZD01000002	*Streptococcus equi subsp. equi*	**CP009248**	*Corynebacterium sphenisci* DSM 44792
CVZC01000002	*Streptococcus equi subsp. equi*	**NSGK01000002**	*Corynebacterium sp*. NML 150383
CVYP01000001	*Streptococcus equi subsp. equi*	**NSGL01000004**	*Corynebacterium sp*. NML 120412
CVYG01000001	*Streptococcus equi subsp. equi*	**LTZQ01000018**	*Corynebacterium sp.* HMSC11H10
CVXW01000002	*Streptococcus equi subsp. equi*	**LWNV01000053**	*Corynebacterium sp.* HMSC08C04
CVXQ01000001	*Streptococcus equi subsp. equi*	**LWOC01000031**	*Corynebacterium sp.* HMSC06G04
CP021972	*Streptococcus equi subsp. equi*	**LWOE01000010**	*Corynebacterium sp.* HMSC06D04
AUZH01000006	*Streptococcus equinus* JB1	**LWOF01000047**	*Corynebacterium sp.* HMSC06C06
AEEL01000001	*Streptococcus equinus* ATCC 700338	**LWOJ01000050**	*Corynebacterium sp.* HMSC05D08
AUZG01000009	*Streptococcus equinus* ATCC 33317	**AGEM01000009**	*Corynebacterium sp*. HFH0082
AEVB01000041	*Streptococcus equinus* ATCC 9812	**LZMN01000084**	*Corynebacterium sp*. EPI-003-04-2554_SCH2473622

JPGC01000005	*Streptococcus equinus*	**LSCZ01000023**	*Corynebacterium sp*. DNF00584
FZRA01000010	*Streptococcus equinus*	**MKKI01000001**	*Corynebacterium sp.* CNJ-954
FOYC01000009	*Streptococcus equinus*	**LSRB01000075**	*Corynebacterium sp*. CMW7794
FOOP01000009	*Streptococcus equinus*	**CP008913**	*Corynebacterium sp.* ATCC 6931
FODM01000004	*Streptococcus equinus*	**CP010827**	*Corynebacterium singulare*
FNZW01000008	*Streptococcus equinus*	**LTEB01000014**	*Corynebacterium simulans*
FNZS01000007	*Streptococcus equinus*	**LTEB01000012**	*Corynebacterium simulans*
FNYU01000010	*Streptococcus equinus*	**CP014635**	*Corynebacterium simulans*
FNKC01000005	*Streptococcus equinus*	**CP014634**	*Corynebacterium simulans*
FNJX01000006	*Streptococcus equinus*	**CP014634**	*Corynebacterium simulans*
FNGX01000009	*Streptococcus equinus*	**CP012342**	*Corynebacterium riegelii*
FMTH01000008	*Streptococcus equinus*	**CP002857**	*Corynebacterium resistens* DSM 45100
CP020438	*Streptococcus equinus*	**PDJF01000001**	*Corynebacterium renale*
AFIN01000214	*Streptococ*cus dysgalactiae subsp. equisimilis SK1249	**ATBY01000017**	*Corynebacterium pyruviciproducens* ATCC BAA-1742
MCRR01000007	*Streptococcus dysgalactiae subsp. equisimilis* AKSDE4288	**CP002924**	*Corynebacterium pseudotuberculosis* PAT10
NBUZ01000056	*Streptococcus dysgalactiae subsp. equisimilis*	**CP003385**	*Corynebacterium pseudotuberculosis* P54B96
MCRQ01000002	*Streptococcus dysgalactiae subsp. equisimilis*	**CP002251**	*Corynebacterium pseudotuberculosis* I19
MCRP01000064	*Streptococcus dysgalactiae subsp. equisimilis*	**CP002097**	*Corynebacterium pseudotuberculosis* FRC41
MCRO01000005	*Streptococcus dysgalactiae subsp. equisimilis*	**CP003652**	*Corynebacterium pseudotuberculosis* Cp162
MCRN01000018	*Streptococcus dysgalactiae subsp. equisimilis*	**CP003061**	*Corynebacterium pseudotuberculosis* CIP 52.97
MAUA01000047	*Streptococcus dysgalactiae subsp. equisimilis*	**CP001829**	*Corynebacterium pseudotuberculosis* C231
MATZ01000045	*Streptococcus dysgalactiae subsp. equisimilis*	**CP001809**	*Corynebacterium pseudotuberculosis* 1002
MATY01000014	*Streptococcus dysgalactiae subsp. equisimilis*	**CP003077**	*Corynebacterium pseudotuberculosis* 316
MATX01000038	*Streptococcus dysgalactiae subsp. equisimilis*	**CP003407**	*Corynebacterium pseudotuberculosis* 267
MATW01000009	*Streptococcus dysgalactiae subsp. equisimilis*	**CP003540**	*Corynebacterium pseudotuberculosis* 258
MATV01000041	*Streptococcus dysgalactiae subsp. equisimilis*	**CP003062**	*Corynebacterium pseudotuberculosis* 42/02-A
LAKV01000029	*Streptococcus dysgalactiae subsp. equisimilis*	**CP003421**	*Corynebacterium pseudotuberculosis* 31
LAKU01000033	*Streptococcus dysgalactiae subsp. equisimilis*	**CP003152**	*Corynebacterium pseudotuberculosis* 3/99-5
LAKT01000031	*Streptococcus dysgalactiae subsp. equisimilis*	**CP003082**	*Corynebacterium pseudotuberculosis* 1/06-A
LAKS01000037	*Streptococcus dysgalactiae subsp. equisimilis*	**ABYQ02000004**	*Corynebacterium pseudogenitalium* ATCC 33035

FWEH01000029	*Streptococcus dysgalactiae subsp. equisimilis*	**AVFF01000046**	*Corynebacterium pseudodiphtheriticum* 090104
AEKN01000017	*Streptococcus downei* F0415	**FXAR01000004**	*Corynebacterium pollutisoli*
MSJM01000010	*Streptococcus cuniculi*	**CP009249**	*Corynebacterium phocae*
AFUE01000002	*Streptococcus cristatus* ATCC 51100	**LKST01000002**	*Corynebacterium oculi*
AEVC01000021	*Streptococcus cristatus* ATCC 51100	**LT629700**	*Corynebacterium mycetoides*
LQRD01000077	*Streptococcus cristatus*	**CP011542**	*Corynebacterium mustelae*
JYGK01000001	*Streptococcus cristatus*	**JSEF01000006**	*Corynebacterium minutissimum*
JYGJ01000002	*Streptococcus cristatus*	**ACEB01000004**	*Corynebacterium matruchotii* ATCC 33806
AEUV02000002	*Streptococcus criceti* HS-6	**ACSH02000002**	*Corynebacterium matruchotii* ATCC 14266
BASX01000016	*Streptoc*occus constellatus subsp. *pharyngis* SK1060 = CCUG 46377	**PJAF01000011**	*Corynebacterium mastitidis*
AFUP01000009	*Streptococcus constellatus subsp. pharyngis SK1060* = CCUG 46377	**CP003924**	*Corynebacterium maris* DSM 45190
AFUP01000009	*Streptococcus constellatus subsp. pharyngis SK1060* = CCUG 46377	**CP007790**	*Corynebacterium marinum* DSM 44953
CP003859	*Streptococcus constellatus subsp. pharyngis* C1050	**LKEV01000001**	*Corynebacterium lowii*
CP003840	*Streptococcus constellatus subsp. pharyngis* C818	**ACHJ01000167**	*Corynebacterium lipophiloflavum* DSM 44291
CP003800	*Streptococcus constellatus subsp. pharyngis* C232	**CP006841**	*Corynebacterium lactis* RW2-5
AICQ01000046	*Streptococcus constellatus subsp. constellatus* SK53	**CP011312**	*Corynebacterium kutscheri*
LQXU01000085	*Streptococcus constellatus*	**CP001620**	*Corynebacterium kroppenstedtii* DSM 44385
JWIY01000004	*Streptococcus constellatus*	**LSCY01000014**	*Corynebacterium kroppenstedtii*
BEWZ01000033	*Streptococcus canis*	**NGUZ01000002**	*Corynebacterium kefirresidentii*
LZDD01000001	*Streptococcus bovimastitidis*	**CR931997**	*Corynebacterium jeikeium K411*
MSPT01000020	*Streptococcus azizii*	**ACYW01000060**	*Corynebacterium jeikeium* ATCC 43734
MSPS01000019	*Streptococcus azizii*	**MTGK01000080**	*Corynebacterium jeikeium*
MSPR01000027	*Streptococcus azizii*	**FMTZ01000017**	*Corynebacterium jeikeium*
AFUD01000015	*Streptococcus australis* ATCC 700641	**CP022054**	*Corynebacterium jeikeium*
AEQR01000003	*Streptococcus australis* ATCC 700641	**LT906467**	*Corynebacterium imitans*
BASY01000002	*Streptococcus anginosus* T5	**CP009211**	*Corynebacterium imitans*
AICP01000029	*Streptococcus anginosus subsp. whileyi* CCUG 39159	**CP005286**	*Corynebacterium humireducens* NBRC 106098 = DSM 45392
ALJO01000004	*Streptococcus anginosus* SK1138	**CP003697**	*Corynebacterium halotolerans* YIM 70093 = DSM 44683

NARD01000002	*Streptococcus anginosus* SK52 = DSM 20563	**NQMQ01000017**	*Corynebacterium hadale*
AFIM01000034	*Streptococcus anginosus* SK52 = DSM 20563	**CP006842**	*Corynebacterium glyciniphilum* AJ 3170
AECT01000038	*Streptococcus anginosus* F0211	**CP004062**	*Corynebacterium glutamicum* ZL-6
AZMF01000137	*Streptococcus anginosus* DORA_7	**APJJ01000009**	*Corynebacterium glutamicum* ZL-2
CP003860	*Streptococcus anginosus* C1051	**AKXP01000012**	*Corynebacterium glutamicum* Z188
CP003861	*Streptococcus anginosus* C238	**CP004048**	*Corynebacterium glutamicum* SCgG2
BASW01000017	*Streptococcus anginosus* 1505	**CP004047**	*Corynebacterium glutamicum* SCgG1
ADME01000011	*Streptococcus anginosus* 1_2_62CV	**AFYA01000004**	*Corynebacterium glutamicum* S9114
PVTC01000001	*Streptococcus anginosus*	**CP005959**	*Corynebacterium glutamicum* MB001
PVTB01000007	*Streptococcus anginosus*	**HE802067**	*Corynebacterium glutamicum* K051
PVTA01000001	*Streptococcus anginosus*	**CP022614**	*Corynebacterium glutamicum* ATCC 14067
PVSZ01000003	Streptococcus anginosus	**AGQQ02000002**	*Corynebacterium glutamicum* ATCC 14067
PVSY01000001	*Streptococcus anginosus*	**BX927149**	*Corynebacterium glutamicum* ATCC 13032
CP012719	*Streptococcus anginosus*	**ABYP01000002**	*Corynebacterium glucuronolyticum* ATCC 51867
ALSI01000039	*Streptococcus agalactiae str. Gottschalk* 1003A	**ACHF01000028**	*Corynebacterium glucuronolyticum* ATCC 51866
ANEY01000016	*Streptococcus agalactiae str. Gottschalk* 1002A	**FWWS01000035**	*Corynebacterium glucuronolyticum*
ALRQ01000025	*Streptococcus agalactiae* STIR-CD-23	**CP019688**	*Corynebacterium glaucum*
ALRP01000023	*Streptococcus agalactiae* STIR-CD-22	**ACLJ02000001**	*Corynebacterium genitalium* ATCC 33030
ALRO01000036	*Streptococcus agalactiae* STIR-CD-21	**CP009247**	*Corynebacterium frankenforstense* DSM 45800
MCQP01000013	*Streptococcus agalactiae* serogroup III	**CP009246**	*Corynebacterium flavescens*
MCQO01000013	*Streptococcus agalactiae* serogroup III	**CP007156**	*Corynebacterium falsenii* DSM 44353
MCQN01000013	*Streptococcus agalactiae* serogroup III	**CP011541**	*Corynebacterium epidermidicanis*
ANEH01000016	*Streptococcus agalactiae* MRI Z1-199	**BA000035**	*Corynebacterium efficiens* YS-314
ANEG01000015	*Streptococcus agalactiae* MRI Z1-198	**ACLI01000059**	*Corynebacterium efficiens* YS-314
ALRU01000064	*Streptococcus agalactiae* LMG 14608	**AMEM01000044**	*Corynebacterium durum* F0235
ALSH01000063	*Streptococcus agalactiae* LDS 628	**CP006764**	*Corynebacterium doosanense* CAU 212 = DSM 45436
ALUP01000009	*Streptococcus agalactiae* GB00922	**CP003217**	*Corynebacterium diphtheriae* VA01
ALUO01000015	*Streptococcus agalactiae* GB00914	**AUZO01000007**	*Corynebacterium diphtheriae* str. Aberdeen

ALQH01000040	*Streptococcus agalactiae* FSL S3-337	**CP003216**	*Corynebacterium diphtheriae* PW8
ALQC01000019	*Streptococcus agalactiae* FSL S3-277	**CP003208**	Corynebacterium diphtheriae INCA 402
ANCV01000055	*Streptococcus agalactiae* FSL S3-268	**CP003215**	*Corynebacterium diphtheriae* HC04
ALQP01000024	*Streptococcus agalactiae* CCUG 37738	**CP003214**	*Corynebacterium diphtheriae* HC03
ALQO01000037	*Streptococcus agalactiae* CCUG 37737	**CP003213**	*Corynebacterium diphtheriae* HC02
ANDE01000035	*Streptococcus agalactiae* CCUG 37736	**CP003212**	*Corynebacterium diphtheriae* HC01
ALRG01000030	*Streptococcus agalactiae* BSU178	**AUZN01000033**	*Corynebacterium diphtheriae* DSM 43988
ANEF01000026	*Streptococcus agalactiae* BSU174	**CP003211**	*Corynebacterium diphtheriae* CDCE 8392
ALRL01000071	*Streptococcus agalactiae* BSU167	**CP003210**	*Corynebacterium diphtheriae* C7 (beta)
ALRI01000018	*Streptococcus agalactiae* BSU165	**JAQO01000005**	*Corynebacterium diphtheriae* bv. mitis str. ISS 3319
MSJL01000042	*Streptococcus acidominimus*	**LSVQ01000028**	*Corynebacterium diphtheriae* bv. *mitis*
LT906454	*Streptococcus acidominimus*	**LSVN01000064**	*Corynebacterium diphtheriae* bv. mit*is*
NDFM01000001	*Rothia sp*. Olga	**LSVL01000034**	*Corynebacterium diphtheriae bv. mitis*
LTKQ01000001	*Rothia sp*. HMSC078H08	**AJVH01000008**	*Corynebacterium diphtheriae bv. intermedius* str. NCTC 5011
LTHR01000038	*Rothia sp.* HMSC075F09	**LTAS01000034**	*Corynebacterium diphtheriae bv. intermedius*
LTXN01000022	*Rothia sp.* HMSC073B08	**JAQQ01000008**	*Corynebacterium diphtheriae bv. gravis* str. ISS 4749
LTKI01000182	*Rothia sp*. HMSC071F11	**JAQP01000009**	*Corynebacterium diphtheriae bv. gravis* str. ISS 4746
LTVT01000023	*Rothia sp.* HMSC071C12	**JAQN01000011**	*Corynebacterium diphtheriae bv. gravis* str. ISS 4060
LTKR01000032	*Rothia sp.* HMSC071B01	**LSYP01000030**	*Corynebacterium diphtheriae bv. gravis*
LTIJ01000064	*Rothia sp*. HMSC069D01	**LTAR01000094**	*Corynebacterium diphtheriae bv. belfanti*
LTXR01000076	*Rothia sp*. HMSC069C04	**LSWA01000044**	*Corynebacterium diphtheriae bv. belfanti*
LTUK01000025	*Rothia sp*. HMSC069C01	**LSVX01000048**	*Corynebacterium diphtheriae bv. belfanti*
LTWL01000056	*Rothia sp*. HMSC068E02	**LSVP01000039**	*Corynebacterium diphtheriae bv. belfanti*
LTYQ01000020	*Rothia sp*. HMSC067H10	**CP003209**	*Corynebacterium diphtheriae* BH8
LTTT01000025	*Rothia sp*. HMSC066H02	**CP003207**	*Corynebacterium diphtheriae* 241
LTUU01000043	*Rothia sp*. HMSC066G07	**CP003206**	*Corynebacterium diphtheriae* 31A
LTXD01000004	*Rothia sp*. HMSC066G02	**OANN01000091**	*Corynebacterium diphtheriae*
LTFZ01000054	*Rothia sp.* HMSC065G12	**LJXS01000001**	*Corynebacterium diphtheriae*
LTTM01000011	*Rothia sp.* HMSC065D09	**JZUJ01000001**	*Corynebacterium diphtheriae*

LTTG01000006	*Rothia sp.* HMSC065D02	**JRUZ01000003**	*Corynebacterium diphtheriae*
LTPL01000068	*Rothia sp*. HMSC065C03	**CP020410**	*Corynebacterium diphtheriae*
LTPN01000061	*Rothia sp.* HMSC064F07	**CP018331**	*Corynebacterium diphtheriae*
LTJO01000032	*Rothia sp*. HMSC064D08	**BX248355**	*Corynebacterium diphtheriae*
LTHF01000056	*Rothia sp*. HMSC062H08	**CP009220**	*Corynebacterium deserti* GIMN1.010
LTWE01000055	*Rothia sp*. HMSC061E04	**FOGQ01000002**	*Corynebacterium cystitidis* DSM 20524
LTQJ01000029	*Rothia sp*. HMSC061D12	**LT906473**	*Corynebacterium cystitidis*
LTSI01000050	*Rothia sp*. HMSC061C12	**AQPS01000037**	*Corynebacterium crenatum* MT
LTJT01000049	*Rothia sp.* HMSC058E10	**CAFW01000086**	*Corynebacterium casei* UCMA 3821
LWNC01000018	*Rothia sp*. HMSC08A08	**CP004350**	*Corynebacterium casei* LMG S-19264
LXWF01000004	*Rothia nasimurium*	**CP027001**	*Corynebacterium camporealensis*
ACSB01000005	*Rothia mucilaginosa* M508	**CP011311**	*Corynebacterium camporealensis*
AP011540	*Rothia mucilaginosa* DY-18	**CP004354**	*Corynebacterium callunae* DSM 20147
AP011540	*Rothia mucilaginosa* DY-18	**JRVJ01000004**	*Corynebacterium auriscanis*
ACVO01000015	*Rothia mucilaginosa* ATCC 25296	**CP001601**	*Corynebacterium aurimucosum* ATCC 700975
CP023510	*Rothia mucilaginosa*	**CP008944**	*Corynebacterium atypicum*
AP014938	*Rothia mucilaginosa*	**CP006365**	*Corynebacterium argentoratense* DSM 44202
CP002280	*Rothia dentocariosa* ATCC 17931	**CP009245**	*Corynebacterium aquilae* DSM 44791
PDEV01000001	*Rothia dentocariosa*	**FTOF01000005**	*Corynebacterium appendicis* CIP 107643
NCWU01000004	*Rothia dentocariosa*	**ABZU01000002**	*Corynebacterium amycolatum* SK46
JQHE01000010	*Rothia dentocariosa*	**MTPT01000034**	*Corynebacterium amycolatum*
JPVS01000007	*Rothia dentocariosa*	**MIFV01000023**	*Corynebacterium amycolatum*
AJJQ01000041	*Rothia aeria* F0474	**CP009244**	*Corynebacterium ammoniagenes* DSM 20306
AXZG01000054	*Rothia aeria* F0184	**ADNS01000017**	*Corynebacterium ammoniagenes* DSM 20306
AP017895	*Rothia aeria*	**CP019705**	*Corynebacterium ammoniagenes*
ACVA01000049	*Prevotella veroralis* F0319	**LXGG01000018**	*Corynebacterium afermentans subsp. afermentans*
JRPQ01000020	*Prevotella timonensis* S9-PR14	**FTMH01000012**	*Corynebacterium afermentans*
ADEF01000044	*Prevotella timonensis* CRIS 5C-B1	**AEED01000060**	*Corynebacterium accolens* ATCC 49726
AFZZ01000130	*Prevotella stercorea* DSM 18206	**ACGD01000026**	*Corynebacterium accolens* ATCC 49725
CBIT010000137	*Prevotella stercorea* CAG:629	**NWBP01000006**	*Corynebacterium accolens*
FOWK01000001	*Prevotella sp.* tf2-5	**NWBP01000001**	*Corynebacterium accolens*
FOCK01000008	*Prevotella sp*. ne3005	**AE017143**	[Haemophilus] ducreyi 35000HP

ALJQ01000043	*Prevotella sp.* MSX73	**MUYD01000012**	[Haemophilus] ducreyi
FNUF01000006	*Prevotella sp*. lc2012	**FOJC01000008**	[Haemophilus] ducreyi
FOHY01000001	*Prevotella sp*. kh1p2	**CP022037**	[Haemophilus] ducreyi
JACD01000090	*Prevotella sp*. ICM33	**CP015434**	[Haemophilus] ducreyi
AWUZ01000120	*Prevotella sp.* F0091	**CP011219**	[Haemophilus] ducreyi
LSDM01000034	*Prevotella sp.* DNF00663	**CP011218**	[Haemophilus] ducreyi
CBBW010000013	*Prevotella sp.* CAG:5226	**CP004046**	[Brevibacterium] flavum ZL-1
CBCU010000014	*Prevotella sp*. CAG:1320	**CP011309**	[Brevibacterium] flavum
CAYW010000220	*Prevotella sp.* CAG:1185		

List of Neisseria isolates used in both analyses

ACCESSION	ORGANISM
MTBN01000007	*Neisseria animaloris*
KY010812	*Neisseria arctica*
MTBL01000003	*Neisseria canis*
MF480373	*Neisseria cinerea*
MF480375	*Neisseria cinerea*
MF480376	*Neisseria cinerea*
ACDY02000004	*Neisseria cinerea*
MTBO01000005	*Neisseria dentiae*
MTAA01000007	*Neisseria dumasiana*
MF480386	*Neisseria elongata*
ADBF01000050	*Neisseria elongata subsp. glycolytica*
CP007726	*Neisseria elongata subsp. glycolytica*
MF480371	*Neisseria elongata subsp. nitroreducens*
LAEJ01000033	*Neisseria flavescens*
LAEK01000032	*Neisseria flavescens*
ATPI01000409	*Neisseria gonorrhoeae*
ATPJ01000009	*Neisseria gonorrhoeae*
ATPK01000069	*Neisseria gonorrhoeae*
ATPM01000379	*Neisseria gonorrhoeae*
ATPN01000440	*Neisseria gonorrhoeae*
ATPO01000380	*Neisseria gonorrhoeae*
ATPP01000483	*Neisseria gonorrhoeae*
ATPQ01000383	*Neisseria gonorrhoeae*
ATPR01000175	*Neisseria gonorrhoeae*
ATPS01000195	*Neisseria gonorrhoeae*
ATPU01000099	*Neisseria gonorrhoeae*

ATPV01000043	*Neisseria gonorrhoeae*
ATPW01000100	*Neisseria gonorrhoeae*
ATPX01000442	*Neisseria gonorrhoeae*
ATPY01000012	*Neisseria gonorrhoeae*
ATPY01000079	*Neisseria gonorrhoeae*
ATPZ01000212	*Neisseria gonorrhoeae*
ATQA01000116	*Neisseria gonorrhoeae*
EQ973035	*Neisseria gonorrhoeae*
EQ973111	*Neisseria gonorrhoeae*
CP019894	*Neisseria lactamica*
MF480361	*Neisseria lactamica*
MF480364	*Neisseria lactamica*
MF480387	*Neisseria lactamica*
FN995097	*Neisseria lactamica*
ACEQ02000002	*Neisseria lactamica*
JOVI01000002	*Neisseria lactamica*
KN046803	*Neisseria lactamica*
CACL01000009	*Neisseria lactamica*
AFQE01000001	*Neisseria macacae*
GL985618	*Neisseria macacae*
APUY01000013	*Neisseria meningitidis*
AVNU01000016	*Neisseria meningitidis*
APUU01000016	*Neisseria meningitidis*
APVD01000011	*Neisseria meningitidis*
AVNV01000026	*Neisseria meningitidis*
ANSD01000002	*Neisseria meningitidis*
ANSC01000002	*Neisseria meningitidis*
CP002424	*Neisseria meningitidis*
CVTF01000079	*Neisseria meningitidis*
FR774048	*Neisseria meningitidis*
AL157959	*Neisseria meningitidis*
JQHF01000015	*Neisseria mucosa*
MF480362	*Neisseria mucosa*
MF480380	*Neisseria mucosa*
MF480381	*Neisseria mucosa*
KT997727	*Neisseria musculi*
KT997728	*Neisseria musculi*
KT997729	*Neisseria musculi*
KT997730	*Neisseria musculi*
MF480370	*Neisseria polysaccharea*
MF480385	*Neisseria polysaccharea*
MF480389	*Neisseria polysaccharea*
ADBE01000009	*Neisseria polysaccharea*

AGAY01000048	*Neisseria shayeganii*
JH164926	*Neisseria shayeganii*
MF480368	*Neisseria sicca*
MF480388	*Neisseria sicca*
ACKO02000016	*Neisseria sicca*
AJMT01000096	*Neisseria sicca*
MF480366	*Neisseria sp.*
MF480372	*Neisseria sp.*
LGZA01000019	*Neisseria sp.*
LGYH01000005	*Neisseria sp.*
MF540366	*Neisseria sp.*
MF488965	*Neisseria sp.*
MF540364	*Neisseria sp.*
CP023429	*Neisseria sp.*
MF540365	*Neisseria sp.*
CP022278	*Neisseria sp.*
MF488964	*Neisseria sp.*
ACWS01000113	*Neisseria sp.*
JZWO01000245	*Neisseria sp.*
LTWU01000004	*Neisseria sp.*
LTNC01000052	*Neisseria sp.*
LTVF01000165	*Neisseria sp.*
CP022527	*Neisseria sp.*
MF480369	*Neisseria subflava*
MF480374	*Neisseria subflava*
MF480382	*Neisseria subflava*
AFWR01000028	*Neisseria weaveri*
AFWQ01000001	*Neisseria weaveri*
MG697220	*Neisseria zalophi*
LT906434	*Neisseria zoodegmatis*
MTBM01000006	*Neisseria zoodegmatis*