Development and Evaluation of a Core Genome MLST Schema for *Haemophilus influenzae*

By

Mariam Iskander

A Thesis Submitted to the Faculty of Graduate Studies of

The University of Manitoba

In partial fulfillment of the requirements of the degree of

MASTER OF SCIENCE

Department of Medical Microbiology and Infectious Diseases

University of Manitoba

Winnipeg

Abstract

Haemophilus influenzae is a human pathogen that can cause disease in young children and the elderly. While there are several typing methods used for *H. influenzae*, serotyping and multi locus sequence typing (MLST) are the two most commonly used methods. The antigenic properties of the polysaccharide capsule surrounding some *H. influenzae* are used to classify the encapsulated strains into six serotypes (a-f), whereas non-encapsulated strains are considered non-typeable (NTHi). Historically, *H. influenzae* serotype b (Hib) has been the leading cause of morbidity and mortality worldwide. The introduction of a Hib conjugate vaccine in the 1990s drastically reduced the incidence of Hib disease. In the following years however, serotype f (Hif) has emerged as the most dominant serotype in the general population while serotype a (Hia) has emerged in the indigenous populations of North America. Since the Hib vaccine does not protect against non-Hib strains, the rising rates of disease warrants investigation into the development of vaccines for other *H. influenzae* serotypes. Developing an effective vaccine for serotypeable strains requires an understanding of its population structure; however, the population structure of *H. influenzae* is currently unclear. Although the 7-gene MLST is commonly used in laboratories worldwide, advances in genome sequencing can be used to provide a vastly more detailed understanding of the population structure of *H. influenzae.* This study investigates the utility of a core genome MLST scheme (cgMLST) as a potential extended MLST scheme for *H. influenzae* typing.

A total of 314 genomes were used to design a cgMLST schema. Minimum spanning trees were generated based on the 7-gene MLST, the ribosomal protein MLST (rMLST) and cgMLST schemas, and all three schemas were evaluated for concordance using Simpson's index of

diversity, the adjusted Rand coefficient and the adjusted Wallace coefficient. A single nucleotide variant (SNV) analysis was performed, and a SNV-based phylogeny was used to compare the concordance of all three methods.

The cgMLST schema contained a total of 980 loci, and partitioned the *H. influenzae* genomes into 204 partitions. The cgMLST schema was shown to have higher discriminatory power compared to the 7-gene MLST and rMLST schemas. Additionally, the cgMLST was found to have the highest level of concordance to the SNV-based phylogeny. The results of this study indicate possible capsular switching or loss among *H. influenzae*. Overall, the cgMLST schema provides higher discriminatory power over the classical 7-gene MLST and the rMLST schemas. A 7-gene MLST schema is considered the gold standard in *H. influenzae* typing, however, with the lowering cost of sequencing, whole genome sequencing-based typing methods should be used. The cgMLST has strong potential to replace the 7-gene MLST scheme as a typing method for *H. influenzae*.

Acknowledgements

First, I'd like to thank my advisor Dr. Gary Van Domselaar for his guidance and support throughout my undergraduate and graduate education. It has been a privilege to study under his supervision and absorb so much knowledge that will help me in my career. I'd also like to thank my committee members who have tremendously helped me during my Master's degree. I'd like to thank Dr. Morag Graham for her kindness, support and the amazing notes she always gave me after our meetings; Dr. George Zhanel for his advice and teachings that have made me a better writer; and Dr. Richard Sparling for his extensive knowledge and for giving me a unique perspective on my project. I would like to thank Dr. Raymond Tsang for his expertise, and teaching me so much about *H. influenzae.*

I would like to thank the entire Bioinformatics team at the National Microbiology Laboratory, who made it enjoyable to work in the supercube. I would especially like to thank Philip Mabon and Aaron Petkau for their endless patience, training and support over the past 5 years, and for never turning me away when I had stupid questions. I also like to thank Dr. Natalie Knox for her support, and always helping me with my research. I'd also like to thank Kristy Hayden for her co-operation and help with wet-lab work that I never would have been able to do without her.

I would like to thank Dr. Eduardo Taboada for his advice and expertise. I did not think it possible to learn as much as I did during a 1-hour teleconference. I'd also like to thank Dillon Barker for helping me get started on my project, and sharing his scripts for the cluster stability analysis. I would like to thank Mickael Silva for his amazing program and helping me interpret my results.

I would like to thank the NML Genomics Core for their help with sequencing, and Genome Quebec for helping me close the Hia genome.

Finally, I would like to thank my parents, Salah and Suzan, for always encouraging me to better myself. I would not have been here without your endless love and support. I'd also like to thank my husband Michael for his support, always being patient and listening to my endless rants. And I'd like to thank my brother Mark for always making me laugh and brightening my day.

Dedication

To my heroes: Mom and Dad

Table of Contents

List of Abbreviations

BLAST	Basic Local Alignment Search Tool
agMLST	Accessory Genome MLST
bp	Base Pair
CDS	Coding Sequence
cgMLST	Core Genome Multilocus Sequence Typing
chewBBACA	Comprehensive and Highly Efficient Workflow Blast Score Ratio Based Allele Calling Algorithm
cps	Capsule Polysaccharide Synthesis
DLV	Double Locus Variant
EMBL	European Molecular Biology Laboratory
FLASH	Fast Length Adjustment of Short Reads
Hia	*Haemophilus influenzae* serotype a
Hib	*Haemophilus influenzae* Serotype b
Hic	*Haemophilus influenzae* Serotype c
Hid	*Haemophilus influenzae* Serotype d
Hie	*Haemophilus influenzae* Serotype e
Hif	*Haemophilus influenzae* Serotype f
iToL	Interactive Tree of Life
LOS	Lipooligosaccharide
LPS	Lipopolysaccharide
MLST	Multilocus Sequence Typing
NCBI	National Center for Biotechnology Information
nLV	N Locus Variant
NML	National Microbiology Laboratory
NTHi	Non-typable *Haemophilus influenzae*
PCR	Polymerase Chain Reaction
pgMLST	Pan-genome MLST
pgMLST	Pan-genome Multilocus Sequence Typing
Prodigal	PROkaryotic DYnamic Programming Genefinding ALgorithm
PRP	Polyribosylribitol phosphate
rMLST	Ribosomal MLST
SAST	Slide agglutination serotyping
SLV	Single locus variant
SNVPhyl	Single Nucleotide Variant Phylogenomics
SPAdes	St. Petersburg Genome Assembler
SRA	Sequence Read Archive
ST	Sequence Type
wgMLST	Whole genome MLST

List of Tables

List of Figures

Chapter 1 Introduction

1.1 General Introduction

Haemophilus influenzae, first described by Pfeiffer in 1892 (Pfeiffer 1892), is a human pathogen that has been causing outbreaks for centuries. At first, it was thought to be the causative agent of the 1918 Influenza pandemic that killed between 20 and 50 million people. For this reason, it was erroneously given the name *influenzae* in 1922 (Kristensen 1922). It is now known that viral influenza was to blame, however *H. influenzae* is thought to have had a key role in this pandemic by causing secondary infections (Morris, Cleary, and Clarke 2017).

Decades later, serotype b (Hib), one of the seven subtypes in *H. influenzae,* continued to cause outbreaks globally, mostly affecting young children, the elderly, and the immunocompromised. A vaccine was developed and introduced in the 1990's that radically reduced the incidence rate of Hib disease. However, with that newfound protection came an increase in disease incidence caused by other *H. influenzae* subtypes. While there are efforts to develop vaccines that protect against all *H. influenzae* subtypes, not much is known about the population structure and diversity of this pathogen, which can greatly aid in the selection of an effective vaccination strategy. Although there are many methods used to study *H. influenzae,* not a lot of methods utilize new advances in sequencing technologies, despite the current trend to modernize pathogen typing and surveillance of other public health priority pathogens with these new sequencing technologies (Zhou et al. 2017; Moura et al. 2016; Kingry et al. 2016). In the current study, a novel typing scheme for *H. influenzae* was developed and evaluated as a potential tool for studying the population structure of *H. influenzae*, and monitoring its spread in populations.

1.2 Bacteriology

1.2.1 Features and Growth Requirements

Haemophilus influenzae is a coccobacillus belonging to the Phylum Proteobacteria, Class Gammaproteobacteria, Order Pasteurellales, Family Pasteurellaceae. It is a Gram-negative, non-motile, facultative anaerobe. All *H. influenzae* have a cell wall with a cytoplasmic membrane, a periplasm, and an outer membrane, and in some strains, a polysaccharide capsule. *H. influenzae* strains lacking a capsule are not detectable using serological methods, and are referred to as non-typeable (NTHi). In contrast, strains with a polysaccharide capsule typically present antigens on the capsular surface, which are used to group them into 6 serotypes (a-f), and are referred to as Hia, Hib, Hic, Hid, Hie and Hif.

The word "*Haemophilus*" is a Greek word meaning "blood-loving" to reflect the bacteria's requirement of the X and V factors to grow. Many *Haemophilus spp.* only require the X factor hemin, which is an organic compound found in the red blood cells. In addition to the X factor, *H. influenzae* also requires the V factor, nicotinamide adenine dinucleotide (NAD) for growth (Evans, Smith, and Wicken 1974). Interestingly, in environments lacking the V factor, *H. influenzae* can grow in satellite colonies in the presence of other NAD-producing organisms such as *Staphylococcus* spp. and some species of fungi (Hirschmann and Everett 1979; Evans and Smith 1972).

In most Gram-negative bacteria, the lipopolysaccharide (LPS) is a major component of the outer membrane, which is comprised of an inner core anchored to lipid A and an outer core with the O antigen, a highly variable polysaccharide. Rather than having an LPS, *H. influenzae* belongs to a group of bacteria that have a lipooligosaccharide (LOS), which lack the O antigen.

The LOS composition is highly variable due to phase variation that occurs in the LOS biosynthesis genes (Swords, Jones, and Apicella 2003; Hardy, Tudor, and Geme 2003; Bayliss, Field, and Moxon 2001). In addition, the surface structures on the *H. influenzae* cell can undergo antigenic variation, either due to mutations or horizontal gene transfer (Hardy, Tudor, and Geme 2003; Bayliss, Field, and Moxon 2001). This awards *H. influenzae* an advantage in evading the immune system. For example, some NTHi strains produce a LOS that mimics a human glycolipid, camouflaging the bacteria from the immune system (Moran, Prendergast, and Appelmelk 1996).

1.2.2 The Polysaccharide Capsule

Many species of bacteria have an outer capsule made of polysaccharides, prominent examples include *Streptococcus pneumoniae*, *Neisseria meningitidis*, *Escherichia coli*, *Salmonella typhi*, and *Haemophilus influenzae*. The capsular locus has been found to be highly conserved among encapsulated bacteria, and it is thought to have a crucial role in the pathogenicity and virulence of these species (Roberts 1996; Zwahlen et al. 1989). There are many functions to the capsule, such as prevention from desiccation (Roberson and Firestone 1992), adhering to each other and to the environment (Wang et al. 2015), and providing a defense mechanism against the host's immune response (Roberts 1996; Cress et al. 2014; Hallström and Riesbeck 2010).

Not all *H. influenzae* strains have an outer capsule. The presence of the polysaccharide capsule was first described in 1931 by Margaret Pittman, when she, and other scientists, observed different phenotypic characteristics of *H. influenzae* in culture. She observed that some strains are encapsulated and appeared smooth, large and opaque while other strains appeared to

lack a capsule, and were shown to be rough, small, and more transparent than the encapsulated strains. Nowadays, the capsule is recognized as an important virulence factor in *H. influenzae,* and has been thoroughly studied.

The capsule itself is composed of monosaccharides joined together by phosphodiester linkages, and it is attached to the cell through covalent bonds (Kuo et al. 1985; Cress et al. 2014). Due to the presence of multiple hydroxyl groups in the monosaccharides, two molecules can link together in different configurations, making it possible for *H. influenzae* to produce six different polysaccharide capsules, a-f (Roberts 1996; Crisel et al. 1975). Capsular biosynthesis begins in the cytoplasm where the sugars are synthesized, polymerized, then transported outside of the cell via an ATP-binding cassette (ABC) transporter (Follens et al. 2001; Roberts 1996).

The capsular biosynthesis genes in *H. influenzae* reside in the capsule polysaccharide synthesis operon – or *cps* operon. The capsular genes are homologous with those found in *E. coli* and *N. meningitidis* (Frosch et al. 1991; Thiên Trí Lâm et al. 2011). The *cps* locus consists of three regions; regions I and III are common to all six serotypes, whereas region II is serotype specific. Region I contains four genes, *bexA, bexB, bexC* and *bexD,* which are responsible for encoding an ATP-binding cassette transporter for exporting the capsular sugars outside the cell (J. S. Kroll et al. 1990). Region III has two genes, *hcsA* and *hcsB,* which encode proteins that play a role the post-translational modification and the export of the polysaccharide from the periplasm to the cell surface (Sukupolvi-Petty, Grass, and St. Geme 2006). Region II, flanked by regions I and III, typically has four to eight genes depending on the serotype. This region harbours the serotype-specific genes, and is responsible for the expression of the polysaccharides (J. S. Kroll, Loynds, and Moxon 1991; Van Eldere et al. 1995; Follens et al. 2001).

While most serotypeable *H. influenzae* have only one copy of the *cps* operon, between 55% and 80% of serotype b (Hib) strains were observed to have a partial tandem duplication – and in some cases, up to six copies – of the *cps* locus (Cerquetti, Cardines, Giufrè, et al. 2006; Cerquetti et al. 2005; Cerquetti, Cardines, Giufre, et al. 2006). The number of copies of the *cps* region has been found to be proportional to the size of the capsule, which in turn has been linked to increased virulence (Corn et al. 1993; Susan K Hoiseth, Moxon, and Silver 1986; J. Simon Kroll, Moxon, and Loynds 1993). The duplicated 17 kb segments are flanked by insertion sequence *1016* (IS*1016*) elements (J. S. Kroll et al. 1990; Zwahlen et al. 1989). In this duplication event, one copy of the *cps* region was truncated, with a 1.2 kb deletion in the IS*1016* and *bexA* gene. The presence of IS*1016* presents potential for recombination in the *cps* region. In fact, spontaneous loss of capsule in Hib strains has been reported if all copies of the *bexA* genes have been disrupted (J. Simon Kroll, Hopkins, and Moxon 1988; S. K. Hoiseth, Connelly, and Moxon 1985; Cintra and Takagi 2015).

In recent years, there has been an increase in hypervirulent Hia strains. It was found that the increased virulence of some Hia strains was associated with a duplication-deletion of the IS*1016* and *bexA* gene, like the mutation found in Hib. While only one copy of the *cps* region was sufficient to cause invasive disease, the duplicated *cps* region has been associated with high mortality rate, especially among children (Kapogiannis et al. 2005; Lima et al. 2010; Ulanova and Tsang 2014). The emergence of this mutation indicates the potential of Hia to cause global outbreaks, similar to those caused by Hib in the pre-vaccine era (Adderson et al. 2001).

1.3 Typing Methods

Traditionally, identification of *H. influenzae* was done through observation of the colony's morphology, testing for growth requirement using the X and V factors, and testing for the presence or absence of the polysaccharide capsule (Price et al. 2015). Now, the most commonly used methods for studying *H. influenzae* are serotyping and multilocus sequence typing (MLST). In addition to MLST and serotyping, other methods are also used: biotyping, which tests the bacteria's ability to produce indole, urease and ornithine; outer membrane protein (OMP) typing, which types bacteria based on the properties of the major outer membrane proteins; multilocus enzyme electrophoresis (MLEE) which identifies variants of 17 housekeeping enzymes; ribotyping (fingerprinting using ribosomal RNA gene enzyme restriction); and pulsed-field gel electrophoresis (PFGE) (fingerprinting using genomic DNA enzyme restriction) (Ulanova and Tsang 2014; Hardy, Tudor, and Geme 2003).

1.3.1 Multilocus Sequence Typing

Multilocus sequence typing (MLST) is a typing method that has been extensively used in bacterial research, and is considered the gold standard for bacterial typing (Maiden et al. 1998). This method relies on the detection of variants in a set of housekeeping genes—typically between 2 and 13 genes—chosen for their stability within the genome. Gene variants, called *alleles* are shared on an online database, making it possible for scientists to compare their data in a standardized way (Maiden et al. 1998). The MLST scheme most commonly used for *H. influenzae* has seven genes, called *loci*, which are summarized in Table 1 (Meats et al. 2003). Each of the seven loci are amplified using polymerase chain reaction (PCR) primers designed by Meats et al. (2003) and sequenced. Each newly sequenced allele is compared to a public MLST database and assigned an allele number, or if a novel allele is found, a curator can assign a

number and add it to the database. The database of loci, their alleles, and their nomenclature is called a *schema.* The allele numbers for all seven loci define an MLST profile, and each unique combination of alleles is assigned a sequence type (ST) number (Sullivan, Diggle, and Clarke 2005). These steps can be done using the Bacterial Isolate Genome Sequence Database (BIGSdb) software (Jolley and Maiden 2010), which is hosted by the University of Oxford on the PubMLST website (https://pubmlst.org/).

Table 1 Housekeeping genes used in the MLST scheme for *Haemophilus influenzae*

Locus	**Gene product**	**Gene length**
adk	Adenylate kinase	477
atpG	ATP synthase F1 subunit gamma	447
frdB	Fumarate reductase iron-sulfur protein	489
fucK	Fuculokinase	345
mdh	Malate dehydrogenase	405
pgi	Glucose-6-phosphate isomerase	468
recA	RecA protein	426

Allele profile data can be clustered and visualized using the global optimal eBURST (goeBURST) algorithm that has been implemented in Phyloviz v2.0 (Nascimento et al. 2017). The algorithm builds a minimum spanning tree—a tree that minimizes the summed distance of all links—using the goeBURST algorithm, which clusters isolates into clonal complexes based on the number of allelic differences between profiles. A great advantage of the goeBURST algorithm is the ability to choose the locus variant level, or the maximum number of allelic mismatches that are used to cluster isolates (e.g. single locus variants (SLV), double locus variants (DLV), …, nLV). The resulting minimum spanning tree will link all clonal isolates, based on the locus variant level used.

1.3.2 Serotyping

Serotyping is a method of identifying bacterial strains based on their surface antigens. Traditionally, *H. influenzae* serotyping was done through slide agglutination serotyping (SAST), using rabbit anti-capsular antiserum (Shively et al. 1981). Six separate antisera, one for each serotype, are used to test each strain, and the reaction to the antisera is observed. However, SAST has been shown to produce inconsistent results. For instance, some Hib strains do not produce a capsule if the *bexA* gene has been disrupted. In this case, SAST cannot distinguish between those serotype b capsule-deficient mutants (Hib^- strains) and NTHi (LaClaire et al. 2003; Satola et al. 2007).

Serotyping by PCR has been more widely used in the laboratory. Primers for the *cps* genes, including the serotype-specific genes, are used to perform PCR, and products are identified by performing electrophoresis (Falla et al., 1994; Ulanova & Tsang, 2014). The PCR serotyping method has been shown to have higher sensitivity and specificity compared to SAST, and is considered a gold standard for serotyping *H. influenzae* (LaClaire et al. 2003). Although PCR can detect capsule-deficient serotypeable strains, this method cannot differentiate between strains that have retained a disrupted copy of the *cps* operon from stains that have completely lost the operon. Alternate methods for serotyping *H. influenzae* have been suggested, such as using the housekeeping gene *pgi*, however these methods are not commonly used (Anyanwu et al. 2003).

1.4 Epidemiology and Pathogenesis

Haemophilus influenzae has a worldwide distribution, with humans being the only reservoir. Historically, Hib had been a significant cause of global outbreaks (Ulanova and Tsang 2009; Dworkin, Park, and Borchardt 2007) In the late 1980's, a conjugate vaccine was developed for

Hib, and was introduced to most developed nations by the 1990s. Since then, a dramatic decline in Hib disease has been observed. In Canada, an average of 700 cases of Hib disease were reported per year before the introduction of the Hib vaccine, whereas the number decreased to 26 cases per year from 2008 to 2012 (Desai et al. 2014). In the United States, the annual number of Hib cases declined on the same scale, from 20,000 cases in the pre-vaccine years, to 2,562 cases from 2003 to 2010 (Hamborsky, Kroger, and Wolfe 2015). The effectiveness of the Hib vaccine is highlighted by observing an increase in Hib disease incidence in under-vaccinated communities, such as the Amish communities in Pennsylvania and Missouri (Myers et al. 2017; Fry et al. 2001).

As the rate of Hib disease decreased, capsular replacement was reported globally. In the general population in the United States, NTHi and Hif strains have been found to be the most prominent cause of disease (Livorsi et al. 2012). Similar trends were also reported in Canada (Shuel et al. 2011; Sill et al. 2007), England and Wales (Hargreaves et al. 1996), Alaska (Perdue et al. 2000), Europe (Ladhani et al. 2010; Resman et al. 2011; R. Whittaker et al. 2017), and Australia (Staples, Graham, and Jennison 2017; Cleland et al. 2017). On the other hand, increase in disease incidence due to Hia has been reported in the Indigenous communities in Northern Canada (Tsang et al. 2016, 2017; Brown et al. 2009; Kelly et al. 2011), United States (Millar et al. 2005), and Alaska (Bruce et al. 2013).

Haemophilus influenzae is a commensal organism that is commonly found in the upper respiratory system in healthy humans (Beck, Young, and Huffnagle 2012). Carriage begins in childhood, where approximately 20% of children under the age of 12 months carry NTHi. Once they reach adulthood, approximately 50% of adults carry NTHi in their nasopharynx, and 6.6% carry typeable strains. Typically, the carriage is transient, and an adult will only carry an *H.*

influenzae strain for a few weeks or months at a time (Howard, Dunkin, and W 1988; Spinola et al. 1986; High, Fan, and Schwartzman 2015; Hardy, Tudor, and Geme 2003). *H. influenzae* is mostly transmitted among people in crowded areas, particularly in schools or overcrowded daycares (Schumacher et al. 2012; Murphy et al. 2009).

Despite the high non-symptomatic carriage rate, *H. influenzae* is an opportunistic pathogen, and can cause disease in vulnerable populations. Toddlers have the highest risk of disease, since the maternal antibodies are no longer protective after 12 months, and their immune system does not mature until around 36 months (Aubrey and Tang 2003). The elderly and immuno-compromised are also at an elevated risk of *H. influenzae* infection. Serotypeable strains typically cause invasive disease, such as meningitis, epiglottitis, cellulitis, septicemia, pneumonia and septic arthritis, while NTHi strains tend to cause non-invasive disease such as bronchitis, sinusitis, and otitis media. In recent years, however, invasive disease caused by NTHi has emerged, and has been reported globally (Van Eldere et al. 2014; Langereis and De Jonge 2015; T. T. Lâm et al. 2016; Tsang et al. 2016).

There are three steps in the development of invasive disease. Colonization, passing the epithelial-blood barrier into the blood stream, and passing the blood-brain barrier into the central nervous system (CNS). *Haemophilus influenza* is a successful colonizer of the nasopharyngeal mucosa, which is the first line of defense against pathogens. Invasive *H. influenzae* can penetrate the mucosal layer, disrupt tight junctions of the respiratory epithelial cells, and gain access to the host's blood stream, which can lead to septicemia (Wilson 1991; High, Fan, and Schwartzman 2015). Once in the blood stream, *H. influenzae* can cross the blood-brain barrier, causing meningitis, which has 3% mortality rate, and 15-30% chance of causing permanent damage (Parisi and Martinez 2014; High, Fan, and Schwartzman 2015).

Non-invasive *H. influenzae* disease, on the other hand, often occurs in the presence of factors that can reduce the integrity of the mucosa and respiratory epithelium. Risk factors include infection with respiratory pathogens, such as viral influenzae, chronic obstructive pulmonary disease and cystic fibrosis (le Roux and Zar 2017; Sriram et al. 2017; Morris, Cleary, and Clarke 2017). There are also reports of elevated rates of *H. influenzae* disease among patients infected with the human immunodeficiency virus, or those who have sickle cell anemia or cancer (Allali et al. 2016; High, Fan, and Schwartzman 2015).

1.5 *H. influenzae* Population Structure

While the population structure of *H. influenzae* is largely unknown, it has been the target of study for many years. The serotypeable strains are more clonal, compared to NTHi, which form a more diverse group (E J Feil et al. 2001). For instance, MLEE analysis reveals that serotypes c, d, e, and f form a monophyletic group (Musser et al. 1988; Meats et al. 2003). Additionally, serotypeable strains can be clustered in two groups: division I, which contains Hia, Hib, Hic, Hid and Hie, and division II, which contains Hia, Hib and Hif. (Ulanova and Tsang 2014; Hardy, Tudor, and Geme 2003).

The NTHi population has been found to be a highly recombinogenic population, with much greater diversity than the serotypeable strains (Cody et al. 2003). For instance, capsular loss among Hib strains has also been documented, and it has been suggested that some NTHi strains may be Hib$^-$. In a Finnish study performing Southern Blot analysis, 31% of NTHi stains studied hybridized with the Hib *cps* locus, 15% of which had IS*1016*. These results suggest that a subgroup within the NTHi population may have descended from encapsulated strains that lost the capsule, and are more closely related to serotypeable strains than to other NTHi (St. Geme et

al. 1994). *Haemophilus haemolyticus* also has a documented presence in the NTHi population, due to the high phenotypic similarity to *H. influenzae* (Nørskov-Lauritsen 2009; Price et al. 2015).

1.6 Hib Conjugate Vaccine

The first attempt to protect against Hib used a polyribosylribitol phosphate (PRP) vaccine. In the early 1970's, the inadequacy of this vaccine became evident, as it did not deliver a strong immune response, especially in children (Peltola et al. 1977). The polysaccharide antigens are T-cell independent, and therefore stimulate the B-cells in adults (Barrett et al. 1992). However, since children under the age of 18 months do not yet have mature B-cells, the polysaccharide vaccine failed to generate an immune response (Timens et al. 1989). Since the PRP vaccine did not protect toddlers, the most vulnerable to Hib disease, alternative vaccines were developed.

To increase the effectiveness of the vaccine, the PRP was conjugated with other proteins that can mount a larger immune response in humans. The first Hib conjugate vaccine (PRP-D), the Hib polysaccharide was conjugated with the diphtheria toxoid as a protein carrier. Since the diphtheria toxoid is T-cell dependent, this vaccine resulted in an enhanced immunogenic response to vaccines, even in young children (Mäkelä and Käyhty 2002; Hamborsky, Kroger, and Wolfe 2015). Currently, the PRP-D vaccine is not commonly used, and has been replaced by PRP-T, which uses the tetanus toxoid as a carrier protein, or the PRP-OMP, which uses the outer membrane protein of *Neisseria meningitidis* serogroup B. The current recommended vaccination schedule begins at two months of age with multiple doses, and a booster shot at 12-15 months. (Hamborsky, Kroger, and Wolfe 2015). In addition to the monovalent conjugate vaccines, there are also several combination vaccines that include the Hib conjugate vaccine: HepB-Hib-PRP-

OMP (hepatitis B and Hib), HibMenCY (Hib, *N. meningitidis* serotype C, and Y-tetanus toxoid) and DTaP-IPV-Hib-PRP-T (diphtheria, tetanus, acellular pertussis, poliovirus, and Hib) (Capeding et al. 2008; Habermehl et al. 2010; Hamborsky, Kroger, and Wolfe 2015).

1.7 Whole Genome Sequencing Approaches

1.7.1 Sequencing Technologies

Advancements in sequencing technology have made sequencing an affordable tool that is now used in laboratories worldwide. Current sequencing technologies fall into two broad classes: short read sequencers, of which the Illumina line of platforms are most popular; and long read sequencers, largely represented by the PacBio line of sequencers. The Illumina sequencing platforms use a sequence-by-synthesis approach; the DNA sequence is "read" as it is being synthesized by polymerases. Library preparation involves shearing the DNA into small fragments, and ligating adapters that are necessary for the sequencing process. The DNA fragments are amplified using a process called bridge PCR. The fragments can be sequenced one of three ways: 1) single-end, where fragments are read from only one end; 2) paired-end, where fragments are read from both ends, and the reads can either be overlapping or separated by a short sequence (typically used to extend the effective size of the read); 3) mate-pair, where fragments are read from both ends, and the reads are separated by a long sequence (used to order and orient disconnected sequences). Coloured fluorescent nucleotides, one colour for each of the four bases, are used in the sequencing process. As the polymerase incorporates a fluorescent nucleotide into a DNA strand, the newly liberated fluorescent molecule is excited and emits a signal that is captured by sensors. This produces raw sequences that can then be assembled with other overlapping reads into draft genome sequence data (Park et al. 2016).

PacBio single molecule real time (SMRT) sequencing is a highly sensitive method that does not require DNA amplification. Libraries are prepared by shearing the DNA into long fragments (~20 kb, but can be much higher), and sequencing is performed using a SMRT cell, which contains thousands of wells, called zero-mode waveguides (ZMW). A single polymerase is attached to the bottom of each ZMW, where a highly sensitive sensor is positioned. As with Illumina sequencing, fluorescent nucleotides are used, with a distinct colour for each of the four bases. As the polymerase incorporates nucleotide into a nascent DNA strand, the fluorescent molecules are cleaved, and the light emitted from them is captured and recorded. Unlike Illumina sequencing, PacBio sequencing results in very long reads, averaging over 10 kb, with some reads over 60 kb. The long reads produced by PacBio can often close a genome, depending on the size of the organism. However, the error rate in PacBio sequencing is the high, ~11-15%, compared to the lower error rate of Illumina, ~1% (Rhoads and Au 2015; Lin and Liao 2015). The high accuracy of short read sequencing is sometimes used to correct the errors in long reads, thus facilitating the closing and finishing of a genome.

1.7.2 Extended MLST Schemas

Extended MLST schemas use the same concept as the current 7-gene MLST scheme, but instead of seven genes, a much larger number of loci are used. Several types of schemas can be defined based on the type of information used, such as ribosomal protein (rMLST), whole genome (wgMLST), pan-genome (pgMLST), core genome (cgMLST) or accessory genome MLST (agMLST) schemas. In recent years, extended schemas have been used more frequently to study bacterial species such as *Salmonella* spp. (Yoshida et al. 2016), *Listeria monocytogenes* (Moura et al. 2016), *Klebsiella pneumoniae* (Zhou et al. 2017), among others. Commercial software can be used to define extended schemas, such as the BioNumerics wgMLST plugin

(Applied Maths, Belgium), or Ridom SeqSphere+ (Ridom, Germany). Open source software options include a combination of software, such as Roary (Page et al. 2015) and BIGSdb (Jolley and Maiden 2010), or using a complete workflow such as chewBBACA (Silva et al. 2017).

There are three general steps to defining an extended schema: 1) defining the loci, 2) allele calling, and 3) filtering the schema for quality and robustness. There are two approaches for defining the loci: the whole genome approach, and the pan-genome approach. The whole genome approach uses all genes in a (preferably high quality, closed, and curated) reference genome or a small set of reference genomes that represent the population under study. However, only using a few genomes can risk excluding accessory genes that may contain important discriminatory information from the final schema. The pan-genome approach uses genes from all genomes in the dataset. While this will ensure that all genes, including accessory genes, are included in the schema, there is a risk of including low quality loci and missing loci, which can may confound the typing and analysis, into the final schema.

Once the wgMLST or pgMLST loci are defined, allele calling will detect allelic variants for each locus in each genome in the dataset. Optionally, low quality genomes that are missing most loci, and low-quality loci missing from most genomes, can be removed from the dataset. The loci, together with the allelic profiles, will constitute the wgMLST or pgMLST schema, depending on the approach used when defining the loci. The cgMLST schemas are defined by selecting loci that are present in 100% of the genomes, and agMLST schemas are defined by using loci found in only a subset of strains, and not represented in the core genome.

Once a schema is defined, isolate relationships can be visualized by generating a goeBURST minimum spanning tree, as described in section 1.3.1 (Francisco et al. 2009; Nascimento et al.

2017). Finally, sequence types (STs) can be defined by clustering allelic profiles, and assigning ST numbers. Isolate metadata and epidemiological data are typically used to help define STs. Cluster stability analysis can also aid in defining a schema by calculating the Shannon index and adjusted Wallace coefficient for all possible goeBURST cluster levels, and choosing a threshold where the clusters begin to stabilize.

Currently, there isn't a single standardized database dedicated for storing wg/pg/cg/ag MLST schemas. Extended schemas are supported by BIGSdb (Jolley and Maiden 2010) on the PubMLST website. Several species already have cgMLSTs defined and available on PubMLST, such as *Campylobacter jejuni* and *Neisseria meningitidis.* Another database for storing cgMLST schemas is a server hosted by Ridom (http://www.cgmlst.org/), which currently has cgMLST schemas for 9 bacterial species.

1.8 Rationale and Objectives

The epidemiology of *H. influenzae* is complex, and continually evolving. With the success of the Hib conjugate vaccine, non-Hib *H. influenzae* has emerged worldwide. As similarities between Hib and Hia appear, concern over Hia's outbreak potential increases. Additionally, the increasing number of invasive disease caused by NTHi has been on the rise on a global scale. These trends confirm the need for more research and development into new vaccines *H. influenzae*. The population structure of *H. influenzae* is currently not very clear. Common typing schemes that are routinely used to study *H. influenzae* have not changed despite the availability of new sequencing technologies. In fact, the misidentification of *H. haemolyticus* as NTHi has been reported, and currently used methods are poor at distinguishing between the two species (Murphy et al. 2007; Price et al. 2015). There are many typing methods that exist for the

studying of *H. influenzae*, they exploit only a small subset of the available molecular information, and some of them are tedious and time consuming.

Instead of only using a handful of genes to type a bacterium with a complex population structure, the entire genome could be utilized, especially since the price of sequencing has dramatically decreased over the past decade. A cgMLST schema has an immense potential to replace some of the typing schemes used in *H. influenzae* research, thus simplifying the process. In the current study, a core genome MLST scheme is designed using both in-house sequences and publicly available *H. influenzae* genomes, and evaluated as a potential new typing tool.

1.8.1 Objectives

There are four objectives to this work:

1) Develop a core genome MLST schema using both typeable and non-typeable *H. influenzae* strains to discern the population structure of the organism
2) Compare the discriminatory power and concordance of the cgMLST schema with the 7-gene MLST schema and the ribosomal protein-based MLST schema
3) Evaluate the concordance of MLST-based trees against phylogenetic trees built using single nucleotide variant data
4) Determine the level of recombination in the *cps* region of *H. influenzae* serotype a to determine the suitability of the Hia capsule as a vaccine target.

1.8.2 Hypotheses

In this thesis, I hypothesize that:

1) The cgMLST schema has a higher discriminatory ability compared to both the 7-gene MLST scheme and the ribosomal MLST scheme
2) The cgMLST schema can better discriminate between serotypes, compared to the 7-gene MLST schema and the ribosomal MLST schema
3) Phylogeny generated using the cgMLST schema is congruent with the SNV-based phylogeny
4) The *cps* region in *H. influenzae* serotype a has little to no recombination

Chapter 2 Methods

2.1 Bacterial Isolates

A total of 83 serotypeable *Haemophilus influenzae* isolates (Table 2) were grown on brain heart infusion (BHI) agar, incubated at 37 °C in 5% CO_2. Bacteria was suspended in 0.2 ml of ddH_2O and heated to 100 °C for 15 minutes. The crude DNA preps were stored at -20 °C. DNA was extracted using either the Qiagen DNeasy Blood and Tissue Kit (Qiagen, Valencia, California, USA) or the MasterPure DNA Purification Kit (Epicentre Technologies, Madison, Wisconsin, USA). Amplified DNA was purified using Agencourt AMPure XP Kit (Beckman Coulter, Brea, CA, USA), and samples were quantified using the Qubit 2.0 Fluorometer (Invitrogen, Carlsbad, California, USA).

Table 2 The number of isolates in each serotype group selected for sequencing

Serotype	# isolates
a	37
b	11
c	10
d	5
e	10
f	10

2.2 Whole Genome Sequencing

A paired-end sequencing library was prepared using TruSeq Nano DNA HT Library Preparation Kit (Illumina, San Diego, California, USA) and sequencing was performed on the Illumina MiSeq platform (Illumina, San Diego, California, USA) at the NML Genomics Core. To establish a reference genome for *H. influenzae* serotype a, a putative vaccine candidate strain 11-139 was selected to be sequenced and closed. Genomic DNA was extracted using Qiagen DNeasy blood and tissue kit (Qiagen, Valencia, California, USA), and a large insert library (20 kb) was prepared using manufacturers' recommendations. Whole sequencing was performed on

a PacBio RSII platform (Pacific Biosciences, Menlo Park, California, USA) at Génome Québec, using a single-molecule real-time cell.

2.3 Publicly Available Data

H. influenzae genomes were downloaded from public databases. A total of 406 paired-end reads were downloaded from the NCBI Sequence Read Archive (SRA). The SRA is a public database hosted by the National Center for Biotechnology Information, which allows individual laboratories to upload and share data publicly. The SRA database holds raw sequencing reads from Illumina, Ion Torren and PacBio (Leinonen, Sugawara, and Shumway 2011). Additionally, 133 whole genomes were downloaded from the Pathosystem Resource Integration Center (PATRIC) database. PATRIC is a project hosted by the University of Chicago that collects reference genomes from public data, and provides tools for biomedical research (Wattam et al. 2017).

2.4 WGS Assembly

The PacBio sequenced genome was assembled using the Hierarchical Genome Assembly Process (HGAP) and polished using Quiver (Chin et al. 2013). In-house generated Illumina reads, as well as paired-end reads downloaded from public databases were assembled using the Galaxy Workflow Spades assemblies with FLASH v1.5 (Bioinformatics Core, NML). The workflow first uses Fast Length Adjustment of Short reads (FLASH) v1.3.0 to merge short paired-end reads. All reads are then assembled using SPAdes v.3.9, using auto k-mer selection. Finally, the contigs are filtered using Filter SPAdes repeats tool, which detects and removes small contigs under 5 kb in length, and contigs with coverage less than 33% of the overall genome coverage.

2.5 Seven-gene Multi-Locus Sequence Typing

The multilocus sequence typing scheme for *H. influenzae* includes the seven housekeeping genes: *adk, atpG, frdB, fucK, mdh, pgi,* and *recA.* Primers were prepared for the for the seven sequences as described by Meats et al. (2003). The DNA was amplified using PCR, prepared, and sequenced using the Illumina sequencing platform. Each sequence was assigned an allele number, and sequence types (ST) were identified using BIGSdb (Jolley and Maiden 2010) on the PubMLST website (https://pubmlst.org/hinfluenzae/).

All whole genome sequences downloaded from public sources were sequence typed using Torsten Seemann's mlst script v2.8 (https://github.com/tseemann/mlst). The program uses MLST schemas from PubMLST, downloaded on January 30th, 2017. All genomes were scanned with BLAST using the *hinfluenzae* MLST scheme. Alleles for each of the seven loci were assigned allelic numbers and each strain was assigned a ST based on the allelic profile. Genomes with missing or incomplete alleles were not assigned a sequence type. Additionally, any novel alleles or sequence types were assigned new numbers and included in the database.

2.6 Serotyping

2.6.1 Slide Agglutination and PCR

All NML strains were serotyped by slide agglutination (Difco, Oakville, Ontario, Canada; Denka Seiken, Tokyo, Japan). Results from the slide agglutination test were supported using PCR. Primers for region II serotype-specific capsular polysaccharide synthesis (cps) genes were prepared, and PCR was performed as previously described by Falla *et al.* (1994). The PCR

products were detected using gel electrophoresis, and compared against a positive control for each of the six serotypes. Strains with no detectable product were considered non-typeable.

2.6.2 *cps* Alignment Serotyping

Serotype specific genes of all six *H. influenzae* serotypes were downloaded from NCBI, and summarized in Table 3. A BLAST database was created using all *cps* genes, using BLAST+ v2.6.0 (Camacho et al. 2009). All genome sequences against were aligned against the *cps* genes database using BLASTn. A serotype was assigned for each strain based on the BLAST report's top hit. A minimum percent identity of 95% and hit length of 95% were used to evaluate top hits. Strains that returned no results were considered non-typeable.

2.6.3 MLST Serotyping

Isolate data was downloaded from PubMLST (https://pubmlst.org/hinfluenzae/) on August 10th, 2017, including isolate IDs, sequence types (ST), and serotypes of all isolates. First, isolate data was used to build a dictionary of serotype and ST pairs using a python script. For each ST, all isolates were queried, and serotypes were paired with the corresponding ST. For each ST, only one serotype was used to build the dictionary. In cases were a ST matched with more than one serotype, the serotype with the largest number of isolates was chosen. All strains were assigned a serotype based on the ST-serotype pairs dictionary. All strains with ST not found in the dictionary were considered non-serotypeable.

2.7 Quality Control

Before assembling reads, all FASTQ files containing fewer than 100,000 reads were eliminated from the dataset. Once assembled, the Galaxy assemblystats tool v1.0.1 was used to

Table 3 NCBI accession numbers of GenBank sequences, and locations of *H. influenzae* serotype specific genes used in the *cps* alignment serotyping method.

			Location	
Serotype	**Gene**	**Accession**	**Start**	**End**
a	acs1	Z37515.2	270	1694
	acs2	Z37515.2	1712	2827
	acs3	Z37515.2	2838	5207
	acs4	Z37515.2	5221	5577
b	bcs1	AF549213.1	4060	5484
	bcs2	AF549213.1	5502	6617
	bcs3	AF549213.1	6632	10279
	bcs4	AF549213.1	10296	12119
c	ccsA	HM770876.1	840	1778
	ccsB	HM770876.1	1789	5460
	ccsC	HM770876.1	5471	6700
	ccsD	HM770876.1	7473	7841
d	dcsA	HM770877.1	695	1819
	dcsB	HM770877.1	1837	3099
	dcsC	HM770877.1	3183	6125
	dcsD	HM770877.1	6138	8105
	dcsE	HM770877.1	8109	9098
	dcs1	HQ424464.1	264	1388
	dcs2	HQ424464.1	1406	2668
	dcs3	HQ424464.1	2752	5694
	dcs4	HQ424464.1	5707	7674
	dcs5	HQ424464.1	7678	8667
e	ecs1	FM882247.1	4351	5475
	ecs2	FM882247.1	5493	6758
	ecs3	FM882247.1	6803	9778
	ecs4	FM882247.1	9775	11235
	ecs5	FM882247.1	11251	11973
	ecs6	FM882247.1	11982	12359
	ecs7	FM882247.1	12356	13915
	ecs8	FM882247.1	13926	14903
f	fcs1	AF549211.1	7744	8838
	fcs2	AF549211.1	8847	11510
	fcs3	AF549211.1	11503	12642

produce assembly statistics such as N50 contig length, total number of bases, number of contigs, etc. N50 contig length values were plotted in order to determine a threshold value to remove low quality genomes. To reduce the chances of incorporating contaminated sequences into the analysis, genomes with total sizes less than 90% (1.62 Mb) or more than 110% (2.07Mb) of the expected genome size of 1.8 Mb were removed. Finally, genomes with any partial or missing MLST genes were also removed from the dataset.

Since genomes in this study were downloaded from multiple sources, some redundant genomes existed in the dataset. Custom python scripts were used to identify and remove duplicate genomes by matching NCBI biosample accession numbers. Additionally, some downloaded genomes were artificially transformed in a laboratory. A python script was used to query NCBI biosample pages for keywords to identify and remove transformed genomes from the dataset. Finally, the set of publicly available genomes that could not be assigned any serotypes were also removed from the dataset.

2.8 Schema Creation

2.8.1 Core Genome MLST

2.8.1.1 PgMLST Schema Creation

The pan-genome loci used in this study were defined using chewBBACA, (Comprehensive and Highly Efficient Workflow: BSR-Based Allele Calling Algorithm, v1.0; Silva et al. 2017). The program consists of a series of python scripts that first define either a whole genome or pan-genome MLST schema, performs allele calling using a set of target genomes, and filters the

schema by removing low quality genomes and loci to define a core genome MLST schema (Figure 1).

ChewBBACA scripts defined coding sequences (CDSs) for each genome with Prodigal (Hyatt et al. 2010), using a supplied *H. influenzae* training file. CDSs were filtered by discarding small sequences contained within larger sequences, as well as CDSs less than 200 bp in size. An all-against-all BLASTP was performed, and BLAST Score Ratios (BSR) are calculated by dividing the query-reference raw BLAST score with the BLAST score of the reference BLASTed against itself. Genes with a BSR of 0.60 or higher are considered homologs and grouped together under the same locus.

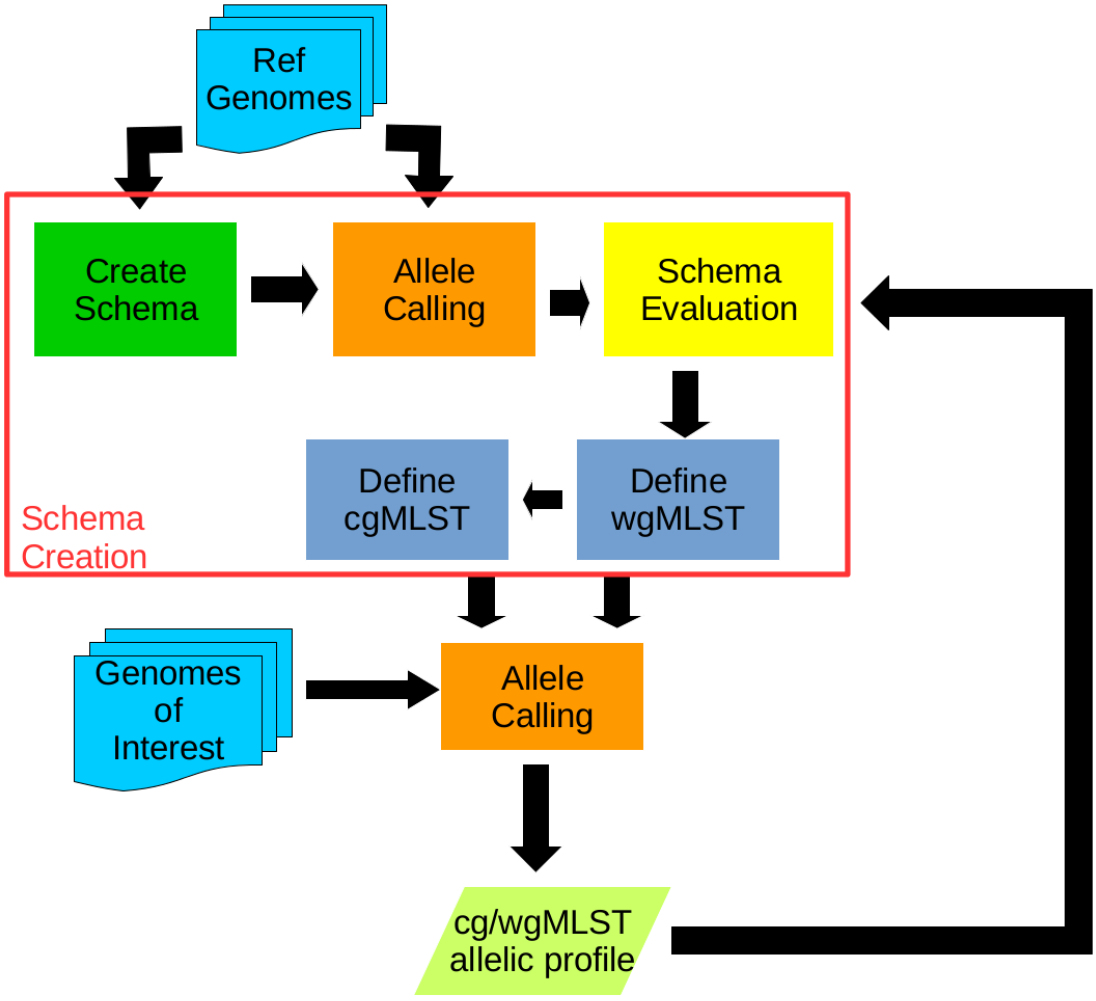
2.8.1.2 Allele calling

Alleles in a set of target genomes were detected using chewBBACA's AlleleCall script. Prodigal is used to detect CDSs in the query genomes, which are in turn BLASTed against the pgMLST schema described in Section 2.8.1.1. In this approach, only CDSs are considered, and any newly identified alleles are automatically added to the database. Paralogous genes are detected and reported. Once all alleles have been called, an evaluation of the schema quality was produced using the SchemaEvaluator script.

2.8.1.3 cgMLST schema definition

The TestGenomeQuality script was used to detect low quality draft genomes in the pgMLST schema. The algorithm detects genomes with missing loci that are found in at least 95% of all other genomes using an adjustable exclusion threshold for the maximum number of allowable missing loci. The script ran with a maximum of 12 iterations, testing exclusion thresholds from 0 to 125 missing loci, with step size of 5.

Figure 1 ChewBBACA workflow for defining a cgMLST schema.


Ref Genomes
Create Schema
Allele Calling
Schema Evaluation
Define cgMLST
Define wgMLST
Schema Creation
Genomes of Interest
Allele Calling
cg/wgMLST allelic profile

The final cgMLST schema was defined using the ExtractCgMLST script, which filters the pgMLST schema and removes low quality genomes and loci. All paralogous genes and low-quality genomes—both of which were detected in sections 2.8.1.2 and 2.7, respectively—were removed from the dataset. Additionally, the script accepts a cut-off parameter, *p*, to remove low quality loci. To create a true cgMLST schema, *p* was set to 1.0, meaning that all loci in the final schema must be present in 100% of the genomes.

2.8.1.4 Cluster stability analysis

The cgMLST schema was used to calculate globally optimal eBURST (goeBURST) (Edward J. Feil et al. 2004) clusters using Phyloviz (Nascimento et al. 2017), and all possible locus variant levels (nLV) were exported. Cluster stability analysis was used to determine the optimal threshold for clustering the cgMLST schema into core genome sequence types (CT). The algorithm calculated the adjusted Wallace coefficient (Severiano, Pinto, et al. 2011) and Shannon index (Keylock 2005) for neighbouring cluster thresholds. The adjusted Wallace coefficients and Shannon indices were plotted against the goeBURST nLV levels, along with the total number of clusters and percentage of singleton clusters at each threshold.

2.8.2 Ribosomal MLST

Ribosomal MLST loci were downloaded from PubMLST (https://pubmlst.org/hinfluenzae/), and prepared as a database using chewBBACA's PrepExternalSchema script. Alleles were called for 304 genomes that were in the final cgMLST schema, as described in section 2.8.1.2. To define a final rMLST schema, no genomes were selected for removal, and a *p* value of 1.0 was used to filter low quality alleles. Clusters at the single locus variant (SLV) level were generated

using Phyloviz (Nascimento et al. 2017), and ribosomal sequence types (rST) were assigned in sequential order for all profiles.

2.8.3 Minimum Spanning Trees

Isolate relationships were visualized using profile data from the three generated MLST schemas. Full minimum spanning trees for the 7-gene MLST scheme, the rMLST scheme and the cgMLST scheme were generated using Phyloviz Online (Ribeiro-Gonçalves et al. 2016). Isolate data was used to annotate tree nodes using serotype information. Both the 7-gene MLST and rMLST schemas were visualized using SLV, whereas the cgMLST schema was generated using locus variants level 25.

2.9 Typing Schemes Congruence Analysis

Congruence analysis was performed using Comparing Partitions (http://www.comparingpartitions.info). This web tool calculates a variety of indices and coefficients that are most commonly used in the analysis and evaluation of typing schemes. In this study, Simpson's index of diversity, adjusted Rand coefficient and the adjusted Wallace coefficient were calculated to discern the discriminatory power and congruence of serotyping, the 7-gene MLST, rMLST and cgMLST schemes.

2.9.1 Simpson's Index of Diversity

The Simpson's index of diversity (*D*) (Hunter and Gaston 1988) was used to test the discriminatory power of typing methods. *D* can be calculated for each typing method as follows:

$$D = 1 - \frac{1}{N(N-1)} \sum_{i=1}^{S} n_i(n_i - 1)$$

where *N* is the total number of strains, *S* is the number of partitions in the typing scheme, and n_i is the number of strains in the ith partition. The index calculates the probability of two strains chosen at random belonging to two different partitions. A 95% confidence interval was calculated as described by Grundmann et al. (2001) as follows:

$$CI = \left[D - 2\sqrt{\sigma^2}, D + 2\sqrt{\sigma^2}\right]$$

and the variance (σ^2) is calculated as follows:

$$\sigma^2 = \frac{4}{n}\left[\sum \pi_i^3 - \left(\sum \pi_i^2\right)^2\right]$$

where π_i is the frequency of strains in the ith partition (e.g. n_i/N).. Since the Simpson's index is measured using a sample of *H. influenzae* genomes, the index may not be representative of the *H. influenzae* population due to sample variability. Calculating the confidence intervals aims to test the reliability of the index. In addition to the CI calculation described above, the the 95% confidence intervals were also calculated using the jackknife approach (Severiano, Carriço, et al. 2011). The jackknife approach is a resampling approach that leaves out one strain in each resample and calculates a confidence interval. The jackknife approach performs better than other resampling techniques and CI calculations, hence it was used to estimate the CI for all coefficients used in this study.

2.9.2 Adjusted Rand Coefficient

The Rand coefficient evaluates the concordance between two clustering methods (Rand 1971), however, it does not consider that two entities can be clustered together by chance. The

adjusted Rand coefficient addresses this issue, and was used in this study in a pairwise manner to compare all four typing schemes (Hubert and Arabie 1985). The adjusted Rand coefficient is calculated by first generating a mismatch matrix from contingency tables (e.g. Table 4).

The adjusted Rand coefficient is calculated as follows:

$$AR = \frac{a + d - n_c}{a + b + c + d - n_c}$$

and,

$$n_c = \frac{N(N^2 + 1) - (N + 1)\sum n_i^2 - (N + 1)\sum n_j^2 + \sum\sum \frac{n_{ij}^2}{N}}{2(N - 1)}$$

where N is the total number of strains, n_i is the number of strains in the ith cluster of typing scheme *A*, and n_j is the number of strains in the jth cluster of typing scheme *B*. The jackknife resampling approach was used to calculate the 95% confidence intervals, and corresponding p-values.

Table 4 A mismatch matrix between two typing schemes, *A* and *B*. Table adapted from Hubert and Arabie (1985)

		Partition *B*		
Partition *A*	Number of pairs	In the same cluster	In different clusters	Sums
	In the same cluster	*a*	*b*	*a* + *b*
	In different clusters	*c*	*d*	*c* + *d*
	Sums	*a* + *c*	*b* + *d*	*M*

2.9.3 Adjusted Wallace Coefficient

The Wallace coefficient is used to evaluate the congruence between two typing methods in a bi-directional manner (Wallace 1983). If two strains cluster together in the same partition using

typing method *A*, the coefficient represents the probability that the two strains will cluster in the same partition using typing scheme *B*. Given a mismatch matrix (e.g. Table 4), the Wallace coefficient for two typing schemes, *A* and *B*, is calculated as follows:

$$W_{A\to B} = \frac{a}{a+b} \quad , \quad W_{B\to A} = \frac{a}{a+c}$$

where *a*, *b*, and *c* are entries in the mismatch table. The adjusted Wallace coefficient, proposed by Pinto, Melo-Cristino, and Ramirez (2008), improves this method by calculating the expected Wallace coefficient (W_i) assuming that classifications are independent, and by calculating a confidence interval. The adjusted Wallace coefficient is calculated as follows:

$$AW_{A\to B} = \frac{W_{A\to B} - W_{i(A\to B)}}{1 - W_{i(A\to B)}}$$

where W_i is the expected Wallace coefficient. W_i is calculated using *D*, Simpson's index of diversity for typing method B ($W_{i(A\to B)} = 1 - D_B$). A 95% confidence interval for the adjusted Wallace coefficient was calculated as described by Pinto, Melo-Cristino, and Ramirez (2008).

$$CI = \left[W_{A\to B} - 2\sqrt{\sigma_{A\to B}}\,, W_{A\to B} + 2\sqrt{\sigma_{A\to B}}\right]$$

and,

$$\sigma_{A\to B} = \frac{\sum_{i=1}^{r}\left(a_i(a_i-1)\right)^2 \sigma_{SID_{B,A_i}}}{\left(\sum_{i=1}^{r} a_i(a_i-1)\right)^2}$$

Where a_i is the sum of row i, and $\sigma_{SID_{B,A_i}}$ is the variance of Simpson's index of diversity of B within cluster A_i. The jackknife approach was also used to calculate the 95% confidence intervals and corresponding *p* values.

2.9.4 Hypotheses

H_0: The 7-gene MLST, rMLST and the cgMLST schemas can all have the same discriminatory power as typing methods for *H. influenzae.*

H_1: The cgMLST schema has a higher discriminatory power in the typing of *H. influenzae* compared to the 7-gene MLST and rMLST typing schemas.

2.10 Single Nucleotide Variant Analysis

2.10.1 *H. influenzae* Population Structure

Single nucleotide variants (SNVs) were detected in all genomes used in this study using the Galaxy SNVPhyl v1.0.1b workflow (Petkau et al. 2017), with the NTHi strain 2019 (accession CP008740) as a reference. SNVPhyl was run using a minimum coverage of 15 X, a minimum mean mapping quality of 15, SNV abundance ratio of 0.75 and a SNV density threshold of 25 over a 400 bp search window. The phylogenetic tree was visualized using EMBL's Interactive Tree of Life (iToL) website (Letunic and Bork 2016). Tree leaf nodes were coloured based on each strain's serotypes, and additional tracks were added outside of the tree to visualize the 7-gene MLST, rMLST, and cgMLST clusters.

Neighbour joining trees were generated using the whole genome SNV data, the cgMLST, rMLST and 7-gene MLST schemas. Either the SNV table or the MLST profiles were used to calculate distance matrices using a custom R script, and neighbour joining trees were built using the R package Ape (Paradis, Claude, and Strimmer 2004). Trees were ladderized and plotted – using ladderize() and plot(), respectively, and tree tips were coloured using serotype data. To compare tree topographies, the SNV tree was compared against each of the MLST trees, and

tanglegrams were generated using Dendroscope v3.539 (Huson and Scornavacca 2012). Final tanglegrams were visualized using Ape's cophyplot.

2.10.2 Recombination Detection in the *cps* Region

A phylogenetic tree was produced by running the Galaxy SNVPhyl workflow v1.0.1 on reads from the 35 Canadian Hia strains and 1 Hic strain, using the Hia reference NML-Hia-1 (accession number CP017811.1). The phylogeny based on whole genome sequences was produced using a minimum coverage of 30, minimum mean mapping quality of 30, SNV abundance ratio of 0.75 and a SNV density threshold of 50 over a 500 bp search window. To produce a phylogenetic tree based only on the *cps* region of Hia, first, the full *cps* region sequence was extracted from the Hia reference, NML-Hia-1 (1470003:1501642). Hia reads were then mapped against the *cps* reference sequence with bowtie2 (Langmead and Salzberg 2012). Reads that mapped to the *cps* region were extracted using samtools v1.5 (Li et al. 2009) and seqtk v1.2 (https://github.com/lh3/seqtk). Phylogeny based on the *cps* region was produced with SNVPhyl using a minimum coverage of 30, minimum mean mapping quality of 30, SNV abundance ratio of 0.75 and a SNV density threshold turned off. Both the whole genome and the *cps* region trees were rooted using the Hic strain as an outgroup. To visualize the differences in tree topologies, a tanglegram was produced using Dendroscope v3.5.9, and the trees were drawn using a custom R script using cophyplot() from the Ape package (Paradis, Claude, and Strimmer 2004).

Chapter 3 Results

3.1 Whole Genome Sequencing

3.2 Data Selection

All 621 genomes used in this dataset were either sequenced in-house or downloaded from public sources, 488 of which were pair-end reads. A small number (55) FASTQ files were judged invalid due to having different number of reads in the forward and reverse directions and were removed from the dataset. The assembly of the 433 read sets was conducted using a Galaxy SPAdes pipeline. A group of 47 read sets failed to assemble using the SPAdes pipeline running FLASH, but were successfully assembled using a modified pipeline that skips the FLASH step. A small number of read sets (11) failed to assemble and were removed from the dataset, leaving 555 *H. influenzae* draft genomes for further analysis.

3.3 Quality Control

To ensure the design of a high quality cgMLST schema is designed, a set of stringent quality criteria were used to eliminate low quality genomes from the dataset. All N50 contig lengths were scatterplotted (Figure 2). The distribution of N50 contig lengths has a strong negative skew, caused by a group of 18 closed genomes downloaded from PATRIC with N50 contig length values exceeding 1.8 million bp. Several genomes with extremely low N50 contig lengths were observed, followed by sudden increase – around 50,000 bp – and a plateau in values (mode = 131,567 bp). All assemblies under 50,000 bp N50 contig length were judged to be low quality genomes and were removed from the dataset. In addition to N50 contig lengths, 5 other quality criteria were used, and 214 genomes were removed as detailed in Table 5. A total of 314 genomes passed quality control and were used for all further analyses.

Table 5 The number of genomes eliminated from the study from not meeting the quality criteria

Quality Criteria	Number of failed genomes
Duplicate genomes	67
Transformed genomes	46
N50 contig length < 50,000	95
Size between 1620000 bp and 2070000 bp	60
Invalid ST	73
No serotype	145
Total genomes removed	214

3.4 Typing Schemes

3.4.1 In Silico Serotyping

For all genomes, *in silico* serotyping was done using two methods: 1) by sequence alignment of the cps region against reference serotypes, or 2) transitively assigned by MLST analysis against the PubMLST database. The performance of the *in silico* serotyping methods was tested using the 82 NML strains which have been traditionally serotyped in the laboratory, using both PCR and SAST. Using the alignment approach, 61 strains (74.4%) were correctly serotyped, while the PubMLST method correctly serotyped 76 strains (92.7%). Using both serotyping methods 61 (74.4%) strains were correctly serotyped. Of the 555 genomes tested, 410 strains had congruent serotype assignments, while 145 had inconclusive serotypes (Table 6).

Figure 2 Scatterplot of N50 contig lengths for 555 *H. influenzae* genomes. The dataset comprised of in-house sequenced genomes, reference genomes downloaded from PATRIC and raw reads downloaded from SRA. The scatterplot was used to define a minimum N50 contig length threshold of 50,000 bps**.**

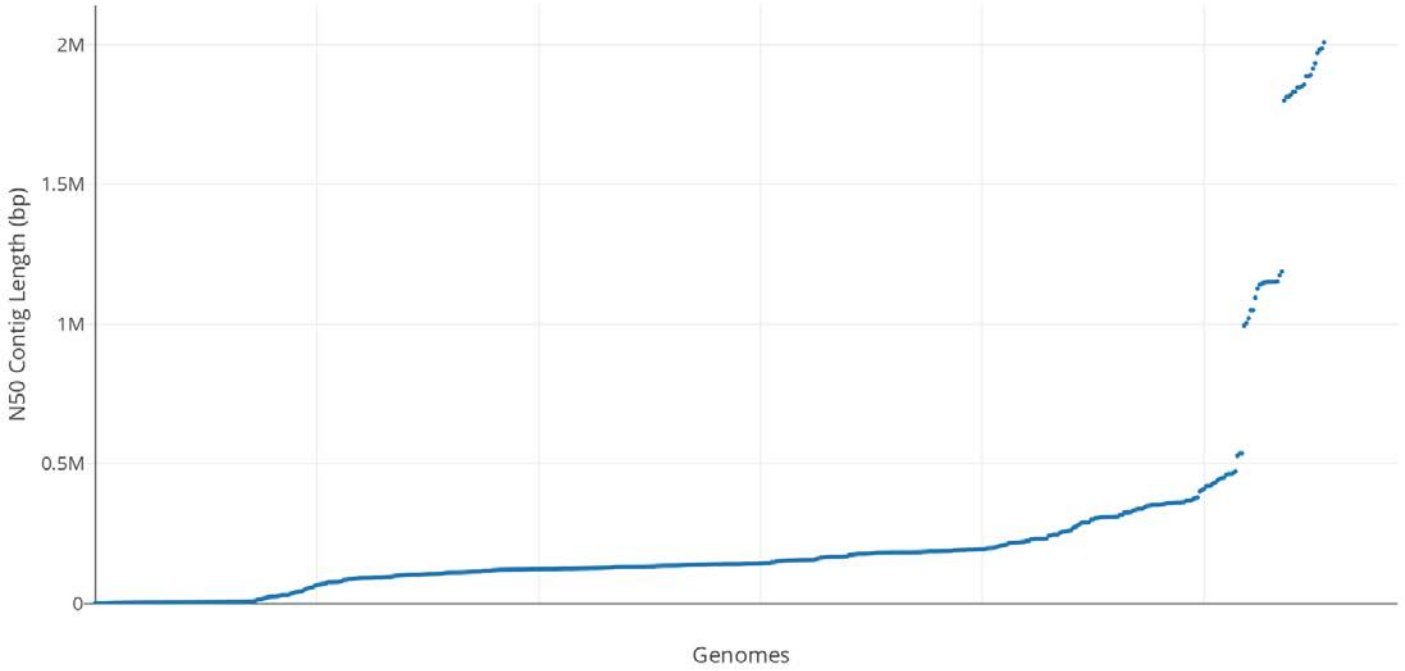

2M
1.5M
1M
0.5M
0
N50 Contig Length (bp)
Genomes

Table 6 The number of strains in each serotype group. All strains were serotyped using the cps region and PubMLST data. Strains that had no metadata or inconsistent serotypes were excluded from the study (ND).

Serotype	Number of strains
a	39
b	42
c	13
d	7
e	14
f	19
NT*	276
ND**	145

* Non-typeable
** Not-determined

3.4.2 Seven-gene MLST Schema

All genomes were typed using the *H. influenzae* 7-gene MLST schema. Of the 555 genomes, 73 were either missing one or more of the seven alleles, or a partial match was found, and 38 were assigned a novel sequence type (ST). A total of 80 novel alleles and 33 novel STs were found. The 304 genomes that were included in the final cgMLST schema were partitioned into 144 different STs.

3.4.3 Core genome MLST Schema

A pgMLST schema was defined by identifying all CDSs in all 314 genomes, and clustering orthologous genes. The schema had 3824 loci, 876 of which had only one allele (Figure 3). Allele mode size was calculated for each locus, which ranged from 165 bp to 9249 bp (Figure 4), with most loci falling between 165 bp and 1500 bp.

To filter out low quality genomes, the number of missing loci in each genome was examined. An exclusion threshold of 20, a parameter to determine the maximum number of loci allowed to be missing from any given genome, was chosen by examining the effects of different threshold

values on cgMLST schemas (Figure 5). Ten genomes were filtered out by this method and were removed from the final cgMLST. Additionally, 135 paralogous genes were identified during allele calling, and the loci were also removed from the final schema.

To define a robust cgMLST schema, low quality loci should be removed. The minimum percentage of genomes missing any given locus (*p*) is used to remove accessory loci. To determine the optimal parameter to use, schemas were generated using 8 different *p* values (65, 70, 75, 80, 85, 90, and 95%). Minimum spanning trees for each percentage were generated (Figure 6). Tree topographies remain relatively stable between 90% and 100% *p* values, and start to diverge for values lower than 85%. The final cgMLST scheme was defined using a *p* value of 100%, and had 980 loci.

A cluster stability analysis was conducted to partition the 980 loci into cgMLST types (CTs). Clusters are generated using every possible locus variant level (i.e. from SLV up to nLV). At the SLV level, the 304 genomes were clustered into 257 groups. The adjusted Wallace coefficient and the Shannon diversity index were calculated for every pair of neighbouring clusters. The plotted coefficients (Figure 7) reveal that the clusters stabilize around a threshold of 25, which was used for the assignment of CTs. The cgMLST schema was partitioned into 204 different CTs (Table 7).

Table 7 The number of loci and the number of distinct clusters in the 7-gene MLST, rMLST and cgMLST schemas.

Schema	**7-gene MLST**	**rMLST**	**cgMLST**
Number of loci	7	49	980
Number of clusters	204	148	144

Figure 3 Frequencies of loci sizes, based on the number of alleles in each locus.

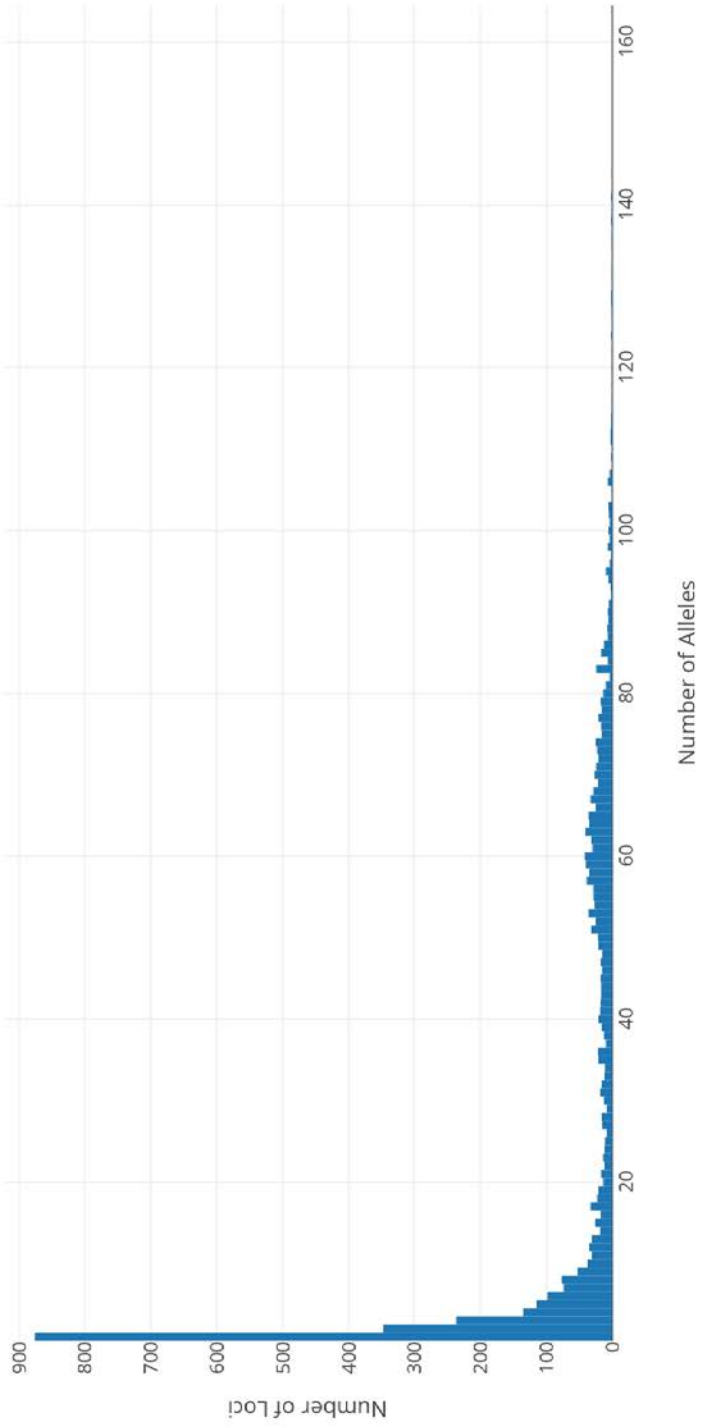

Number of Alleles
0
20
40
60
80
100
120
140
160
Number of Loci
0
100
200
300
400
500
600
700
800
900

Figure 4 Frequencies of allele mode sizes

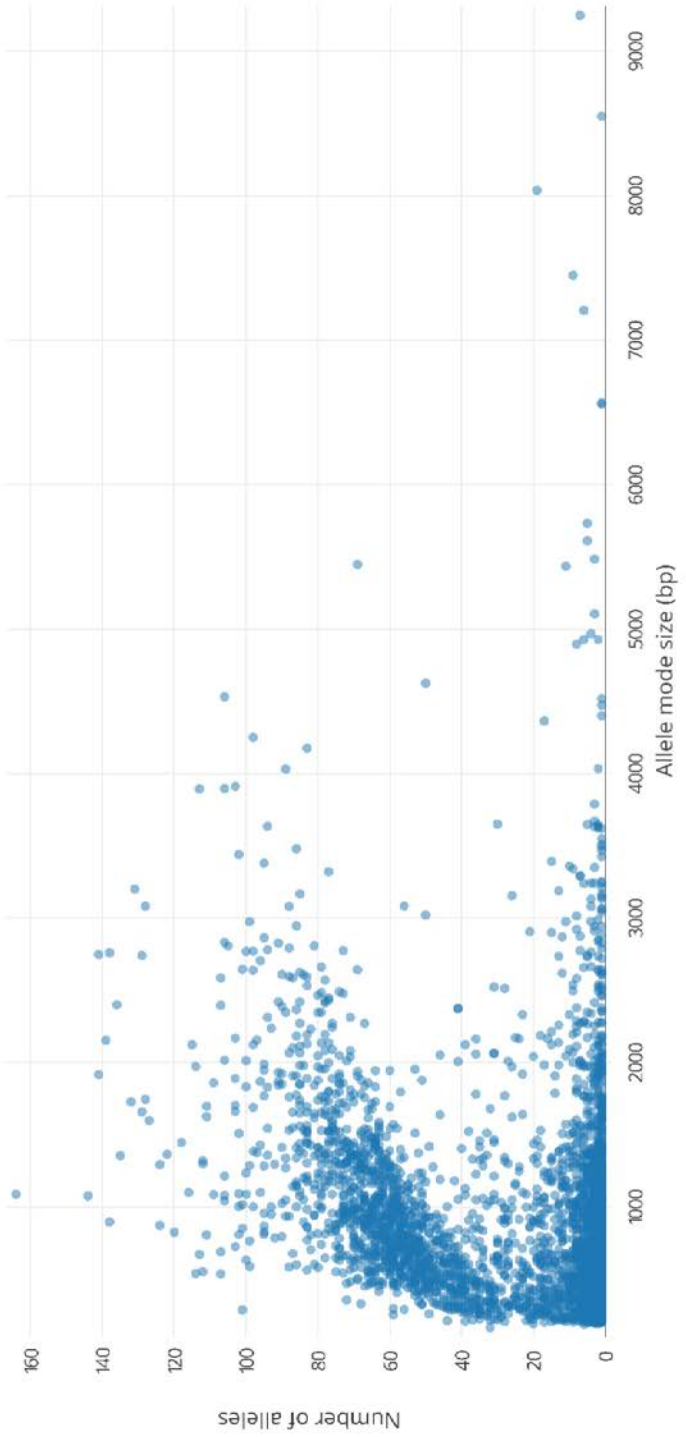

Number of alleles
0
20
40
60
80
100
120
140
160
1000
2000
3000
4000
5000
6000
7000
8000
9000
Allele mode size (bp)

Figure 5 Number of loci and number of genomes in every exclusion threshold level. The number of alleles found in 95%, 99%. 99.5% and 100% of genomes are shown.

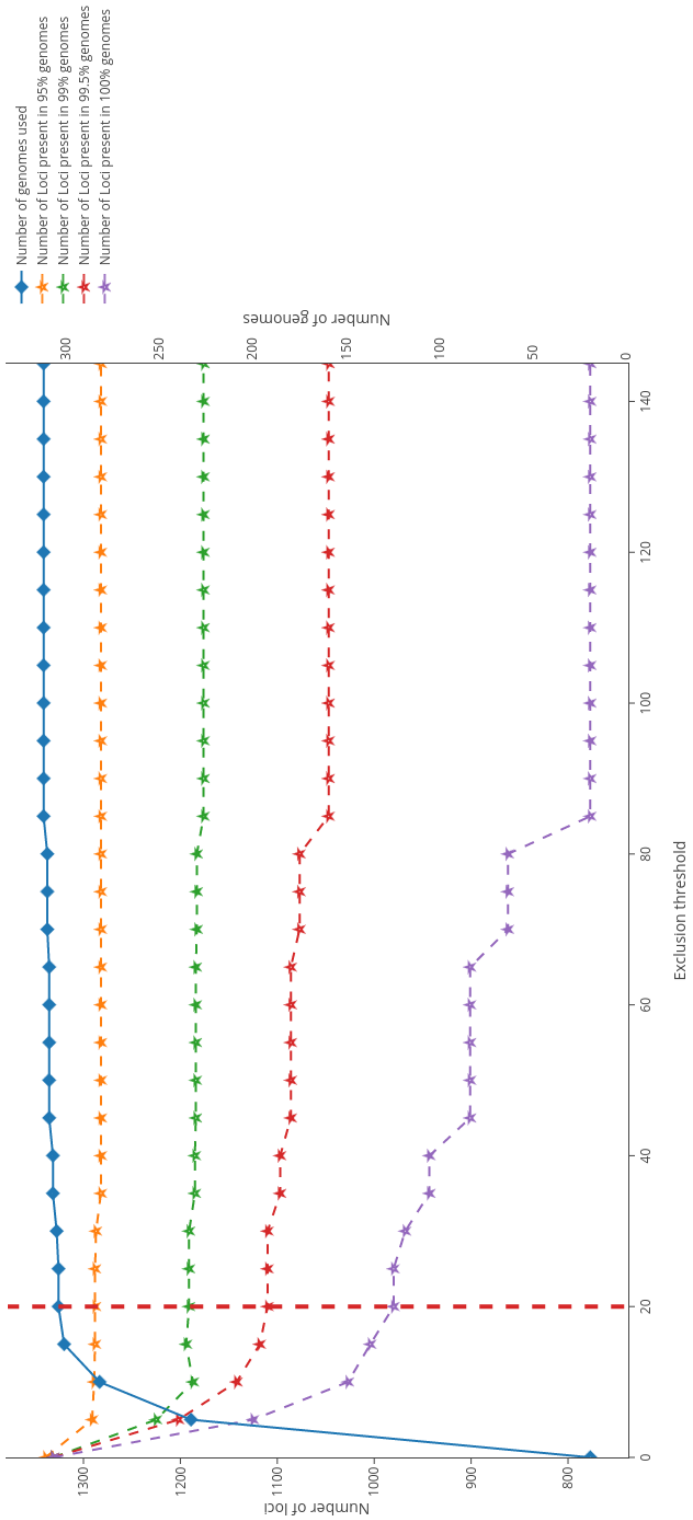
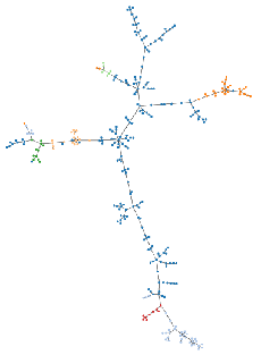
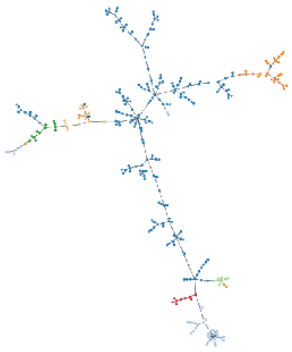

Number of genomes used
Number of Loci present in 95% genomes
Number of Loci present in 99% genomes
Number of Loci present in 99.5% genomes
Number of Loci present in 100% genomes
Number of genomes
300
250
200
150
100
50
0
Number of loci
1300
1200
1100
1000
900
800
Exclusion threshold
0
20
40
60
80
100
120
140

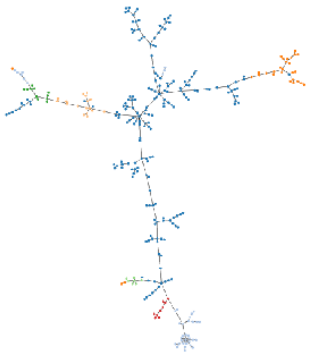
Figure 6 Core genome MLST minimum spanning trees generated using the goeBURST algorithm using 8 exclusion threshold levels, ranging from 0.65 to 1.0. A total of 304 genomes were used to generate the cgMLST schema.



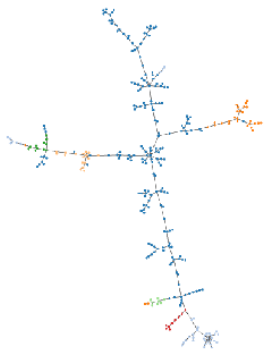
1.0 – 961 loci



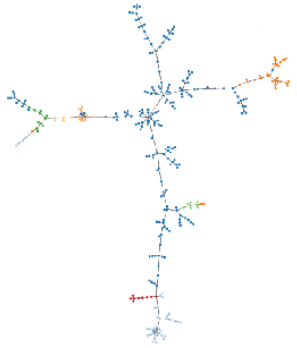
0.95 – 1285 loci



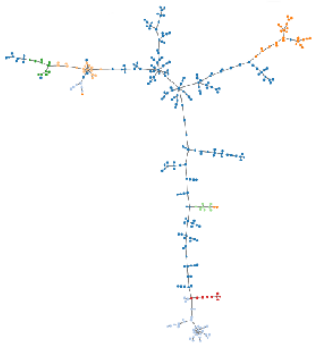
0.90 – 1321 loci



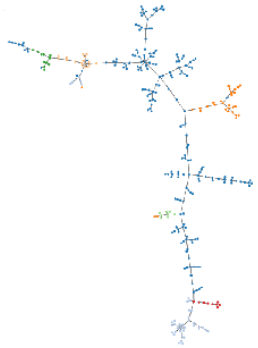
0.85 - 1348 loci



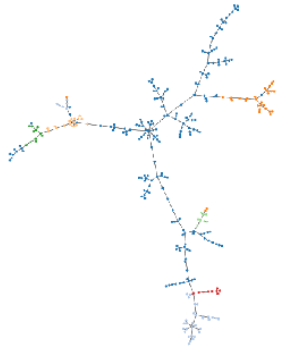
0.80 – 1370 loci



0.75 – 1384 loci

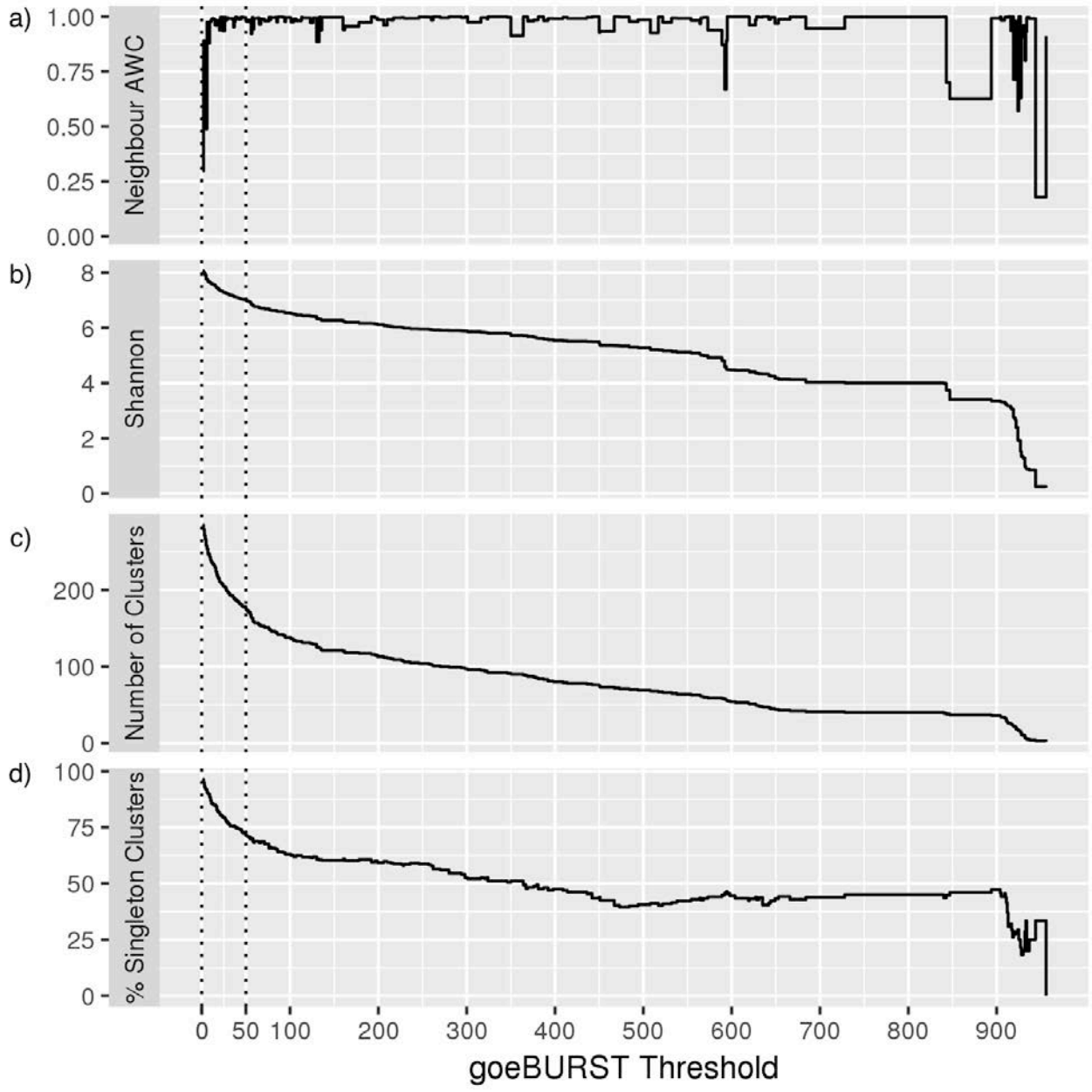


0.70 – 1394 loci



0.65 – 1410 loci

Figure 7 Cluster stability analysis of the cgMLST goeBURST clusters at 257 different locus variant levels. The adjusted Wallace coefficient and Shannon index was calculated for each pair of neighbouring clusters. The coefficients, number of clusters and percentage of singletons were plotted against all possible locus variant levels.


a)
1.00
0.75
0.50
0.25
0.00
Neighbour AWC
b)
8
6
4
2
0
Shannon
c)
200
100
0
Number of Clusters
d)
100
75
50
25
0
% Singleton Clusters
0
50
100
200
300
400
500
600
700
800
900
goeBURST Threshold

3.4.4 Ribosomal MLST Schema

A ribosomal protein MLST schema was defined based on a 53-locus schema downloaded from PubMLST. The chewBBACA workflow was used to call alleles for all 304 genomes that were in the final cgMLST schema. No paralogous genes were detected, and no genomes were removed from the dataset. Alleles were required to be present in 100% of genomes. Four alleles did not meet this requirement, and were removed from the schema. The final rMLST schema had 49 loci, and partitions 304 genomes into 148 ribosomal sequence types (rSTs).

3.4.5 Minimum Spanning Trees

Minimum spanning trees were built using the goeBURST algorithm for each of the MLST schemas (Figures 8, 9, and 10). The cgMLST schema was additionally clustered at nLV level 25 to reflect the different CTs. As expected, NT strains show the largest amount of diversity compared to the serotypeable strains in all three trees. While serotypes c, d, e and f were all clustered together in the cgMLST minimum spanning tree, serotype a strains were grouped into two distinct clades; the majority of strains clustered closer to serotypes c and d, while a smaller number of strains clustered more closely with serotypes e and f. Most serotype b strains were clustered among NT strains, except for 3 strains that were either clustered with Hia or Hic.

Despite a few similarities in clustering patterns, the minimum spanning trees produced by the rMLST schema and 7-gene MLST schema differ in topology, compared to the cgMLST tree. In the rMLST tree, Hia and Hib are clustered near each other; however, this isn’t supported by either the 7-gene MLST or the cgMLST trees. In the 7-gene MLST tree, Hib and Hie are most closely related, while the rest of the serotypes form distinct clusters among the NTHi strains.

3.5 Typing Methods Congruence Analysis

To evaluate the congruence among typing schemes used in this study, the Simpson's index of diversity (*D)*, the adjusted Rand coefficient (*AR*), and the adjusted Wallace coefficients (*AW*) were calculated, using the jackknife resampling approach to calculate confidence intervals.

Simpson's index of diversity calculates the probability that two randomly selected strains will be typed in two different partitions. The 7-gene MLST, rMLST and cgMLST all have high values for the Simpson index, with the cgMLST being slightly higher than the 7-gene MLST and rMLST schemas (Table 8). The confidence interval for all three MLST methods overlap, meaning that the null hypothesis cannot be rejected, and all three methods have the same discriminatory power. Interestingly, p values produced by the jackknife approach indicates that the differences in values between the cgMLST and all other typing methods, to be statistically significant ($p < 0.001$, using $\alpha=0.05$). Therefore, the null hypothesis that the cgMLST has the same discriminatory power as serotyping, the 7-gene MLST, and the rMLST schema, can be rejected. On the other hand, differences between the rMLST and the 7-gene MLST schemas were found to be to be statistically insignificant ($p = 0.081$; using $\alpha=0.05$), indicating that the discriminatory power between the two methods is likely equal.

Table 8 Simpson's diversity indices for serotyping, 7-gene MLST, rMLST and cgMLST schemas. The confidence intervals were calculated using the jackknife resampling approach

Typing scheme	Number of partitions (*S*)	Simpson's ID (*D*)	Confidence Interval
Serotype	7	0.576	(0.515 – 0.636)
7-gene MLST	144	0.989	(0.986 – 0.992)*
rMLST	148	0.986	(0.982 – 0.991)*
cgMLST	204	0.993	(0.990 – 0.996)*

* Overlapping confidence intervals at 95%

Figure 8 Minimum spanning tree generated using the cgMLST schema of 304 *H. influenzae* isolates. Tree nodes were clustered at locus variant level 25, and coloured by serotype.

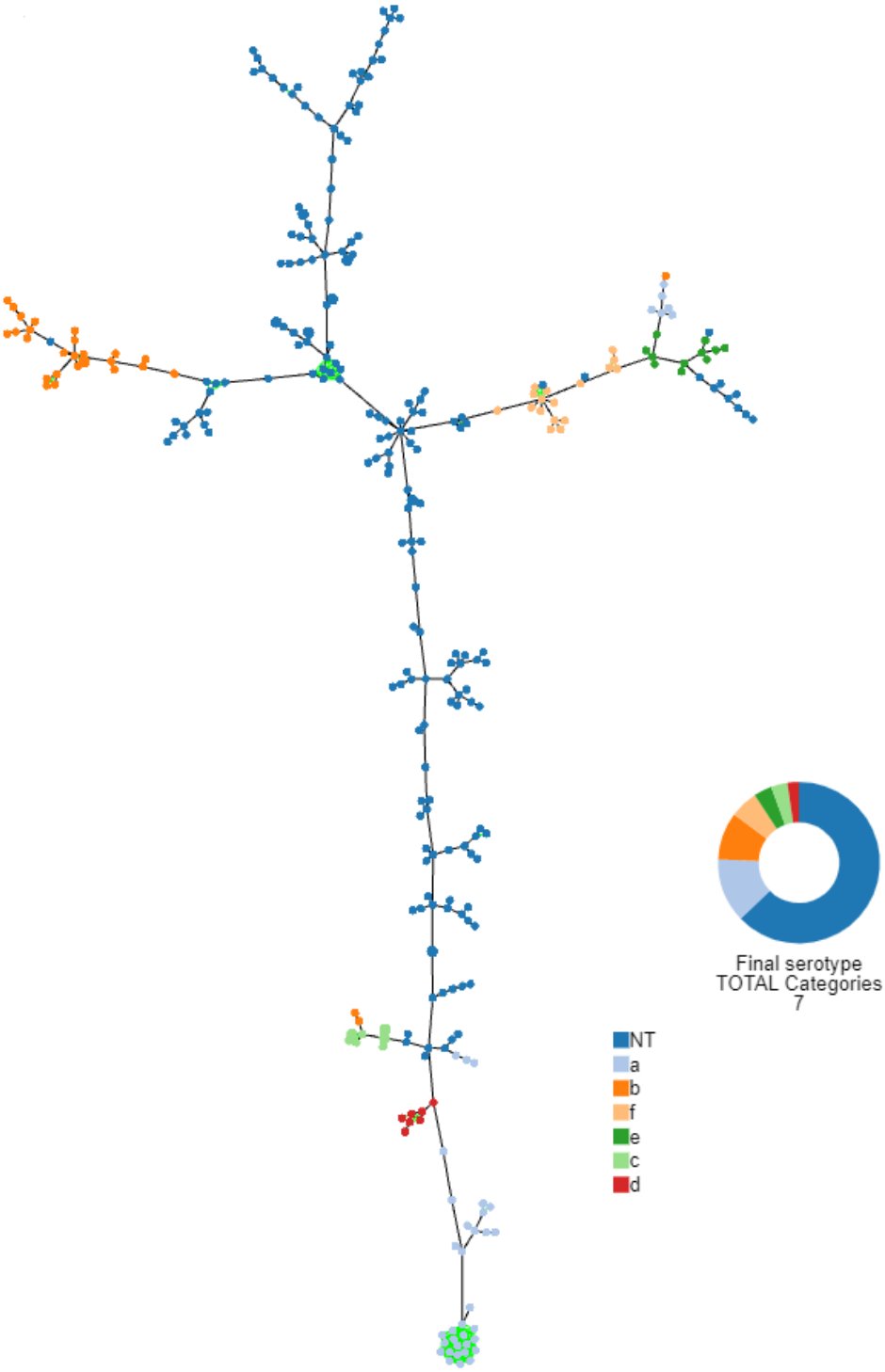

Final serotype
TOTAL Categories
7
NT
a
b
f
e
c
d

Figure 9 Minimum spanning tree generated using the rMLST schema of 304 *H. influenzae* isolates. Tree nodes were coloured by serotype.

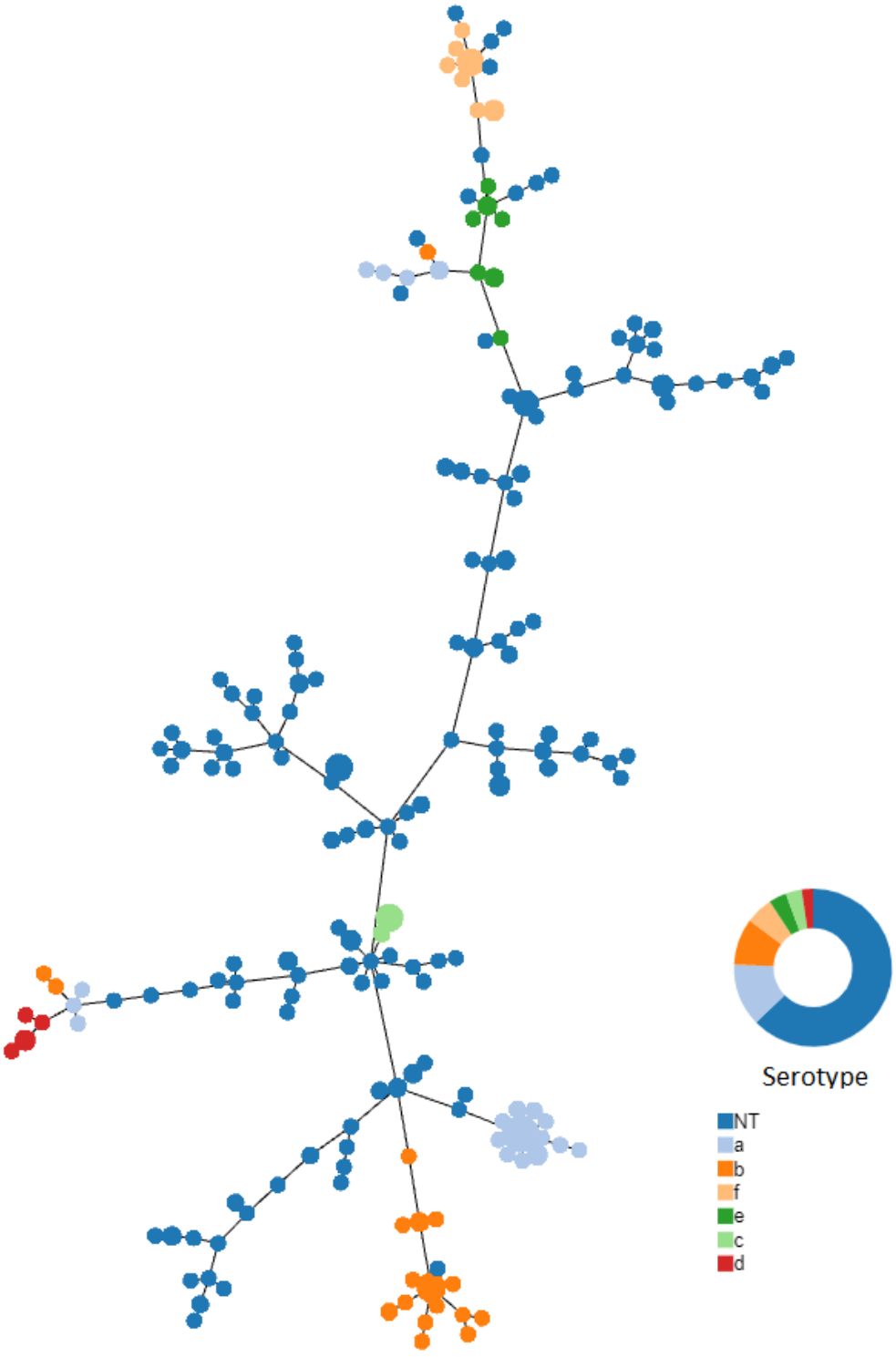
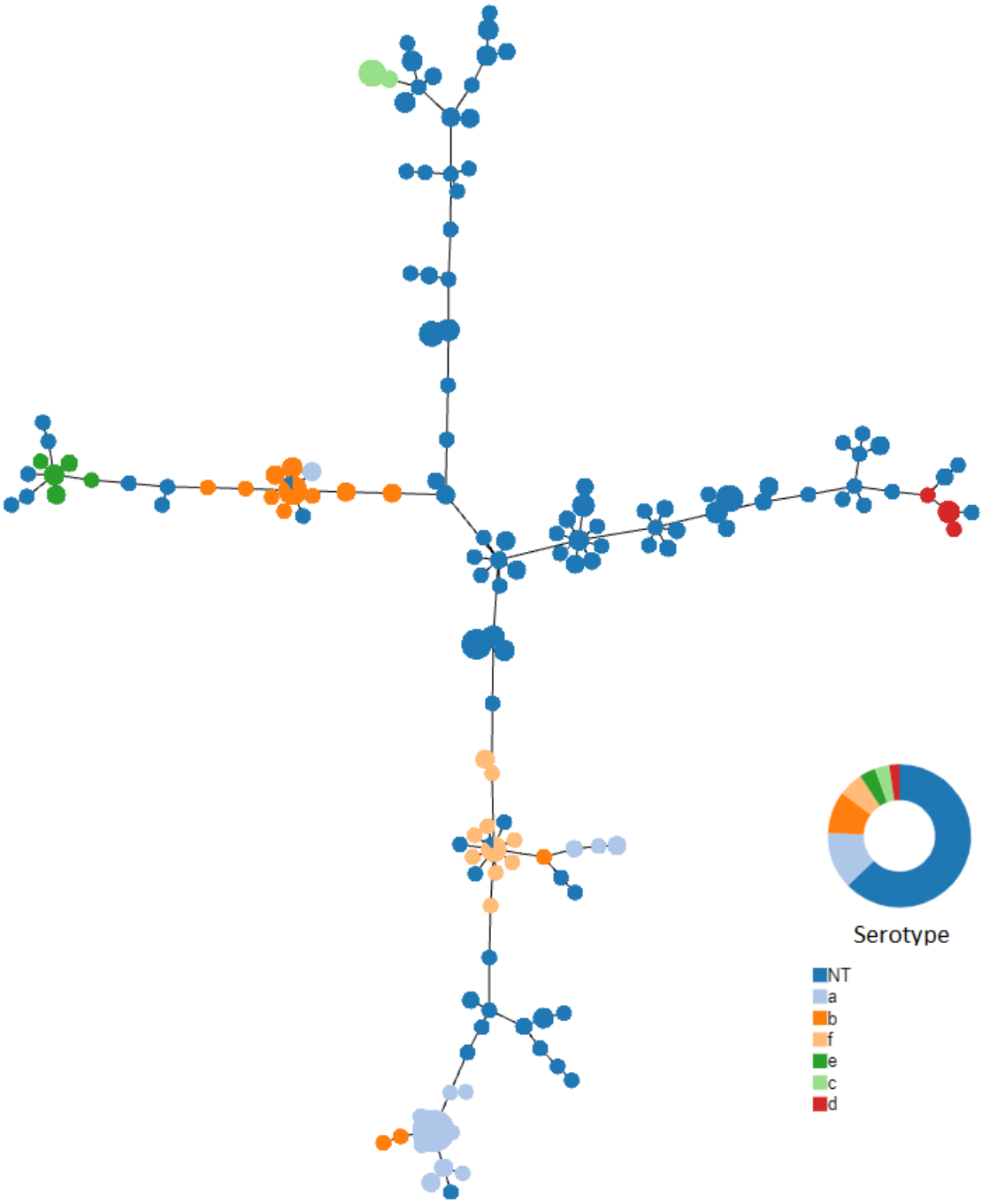

Serotype
NT
a
b
f
e
c
d

Figure 10 Minimum spanning tree generated using the 7-gene MLST schema of 304 *H. influenzae* isolates. Tree nodes were coloured by serotype.


Serotype
NT
a
b
f
e
c
d

The adjusted Rand coefficient calculates the concordance between two typing methods. The highest level of concordance was found between the 7-gene MLST scheme and the rMLST scheme (Table 9). Since at 95% the confidence intervals overlap, we can't reject the hypothesis that the 7-gene MLST scheme is as congruent with the cgMLST scheme as the rMLST schema ($p = 0.997$).

Table 9 Adjusted Rand coefficient and 95% confidence intervals for serotyping, 7-gene MLST, rMLST and cgMLST schemas

Typing Scheme	Serotype	7-gene MLST	rMLST	cgMLST
Serotype				
7-gene MLST	0.028 (0.015 – 0.040)¶			
rMLST	0.035 (0.016 – 0.051)¶	0.729 (0.632 – 0.827)¥		
cgMLST	0.017 (0.007 – 0.028)¶	0.649 (0.527 – 0.781)¥	0.649 (0.526 – 0.782)¥	

¥ ¶ Overlapping confidence intervals at 95%

Finally, the adjusted Wallace coefficient was calculated for all typing pairs (Table 10). If two strains are clustered in the same cluster in the cgMLST schema, there is 86.7% chance of the two strains also clustering in the same ST, however in the reverse situation, the probability is only 51.9%. In addition, the 95% confidence intervals do not overlap, meaning that the null hypothesis is rejected ($p < 0.001$). Similar comparisons between other typing methods revealed that all MLST schemes are more discriminatory than serotyping ($p < 0.001$) and cgMLST is more discriminatory than rMLST ($p < 0.001$). However, the 95% confidence interval of the 7-gene MLST and rMLST schemas overlap, meaning that both methods have equal discriminatory power ($p = 0.044$).

Table 10 Adjusted Wallace Coefficients and 95% confidence intervals using serotyping, 7-gene MLST, rMLST and cgMLST schemas

Typing Scheme	Serotype	7-gene MLST	rMLST	cgMLST
Serotype	-	0.014* (0.008-0.021)	0.018* (0.010-0.026)	0.009* (0.003-0.015)
7-gene MLST	0.952* (0.902-1.000)	-	0.816 (0.741-0.891)	0.519* (0.433-0.604)
rMLST	0.944* (0.882-1.000)	0.659 (0.549-0.769)	-	0.481* (0.373-0.589)
cgMLST	0.977* (0.950-1.000)	0.867* (0.740-0.994)	0.997* (0.993-1.000)	-

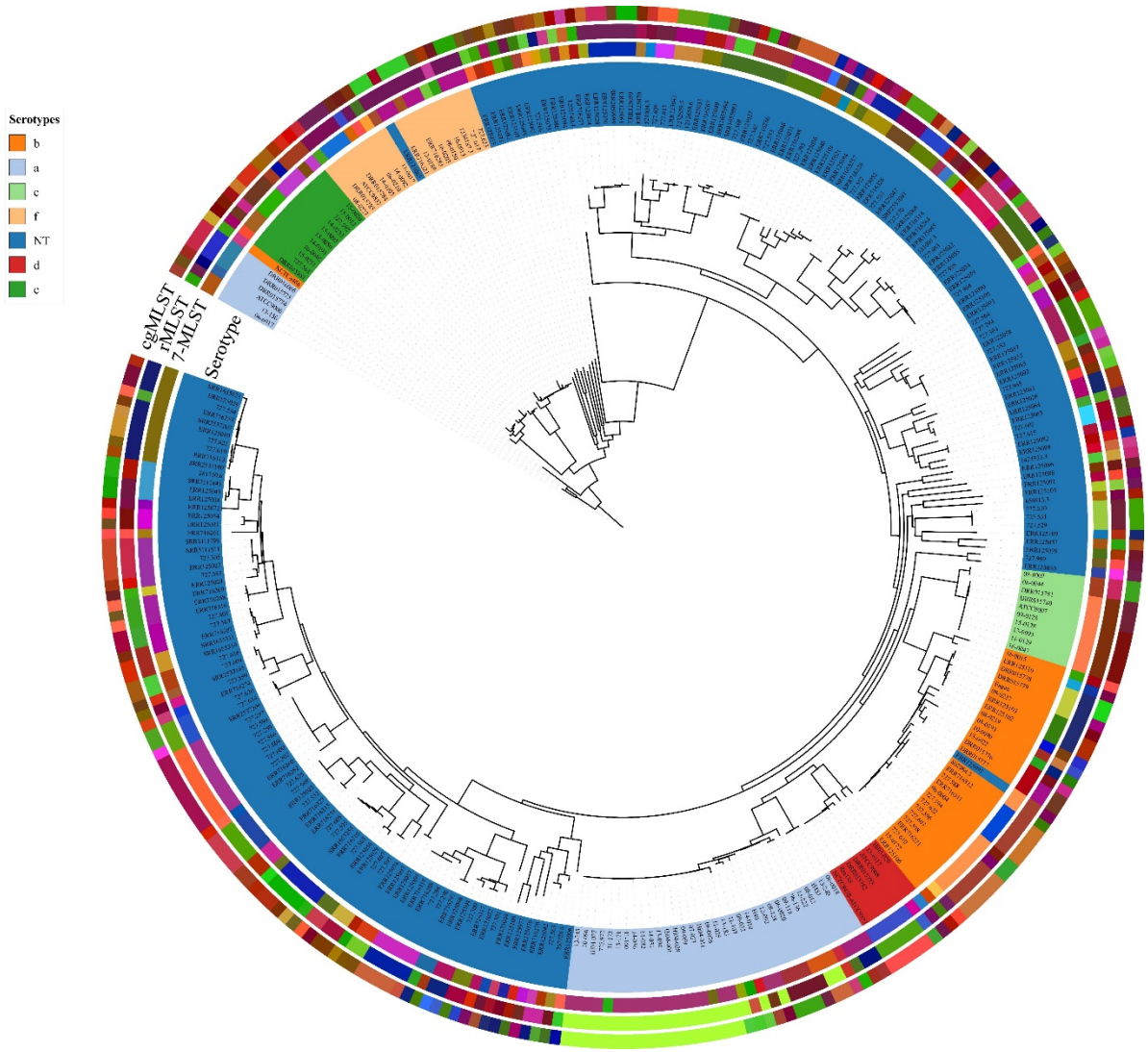
* Significance at $p < 0.001$ with reciprocal comparisons

3.6 Single Nucleotide Variant Analysis

3.6.1 *H. influenzae* Population Structure

Single nucleotide variant analysis was performed with SNVPhyl, using all 304 genomes in the final cgMLST schema. A total of 275,433 SNVs were detected among all *H. influenzae* genomes, however, after filtering, the final phylogenetic tree was generated using only 26,888 SNVs. A circular phylogenetic tree was generated, and annotated by using MLST-based schemes clusters, and serotype information (Figure 11). There are two distinct Hia strains, which are consistent with the two clusters found in the cgMLST schema minimum spanning tree. However, unlike the cgMLST minimum spanning tree, Hia and Hib are clustered together with serotypes c and d. Similarly, serotype f and e are clustered along with the smaller Hia group, and one Hib strains. Finally, NTHi strains ERR125065 and ERR125097 are clustered with serotypes f and b, respectively.

Figure 11 Phylogenetic tree based on SNV-data. Each isolate label was colourized using serotype data. Tracks for cgMLST, rMLST and 7-gene MLST clusters were plotted around the tree.


Serotypes
b
a
c
f
NT
d
e
cgMLST
rMLST
7-MLST
Serotype

Distance matrices using both the SNV data and MLST profile data were calculated, and neighbour joining trees were produced. A neighbour joining tree (NJT) for each method was generated, and each node tip coloured using serotype data (Supplementary Figure 1). The NJT produced from the SNV data was compared to NJ trees generated from all allele profile data produced from each of the MLST schemas. The 7-gene MLST tanglegram (Figure 12) shows the least amount of concordance, compared to the rMLST (Figure 13) and the cgMLST tanglegram (Figure 14). NTHi strains are more organized in the cgMLST tanglegram, compared to the 7-gene MLST tanglegram, and more clusters were kept intact, even if they moved around the tree.

3.6.2 Recombination Detection in the *cps* Region

Single nucleotide variant analysis was performed using the whole genome of Hia, as well as the *cps* region to uncover the amount of recombination in the *cps* genes. The whole genome SNV analysis detected 111058 total SNVs, and the phylogenetic tree was build using a total of 77318 SNVs. On the other hand, 95 SNVs were detected in the *cps* region analysis, and 73 SNVs were used to build the phylogenetic tree. Topology differences of the two trees were visualized by producing a tanglegram (Figure 15). The tanglegram indicates that some recombination occurs in the *cps* region.

Figure 12 A tanglegram generated using neighbour joining trees of SNV data (left), and 7-gene MLST profiles (right). Tanglegram connections were colorized based on serotype data.

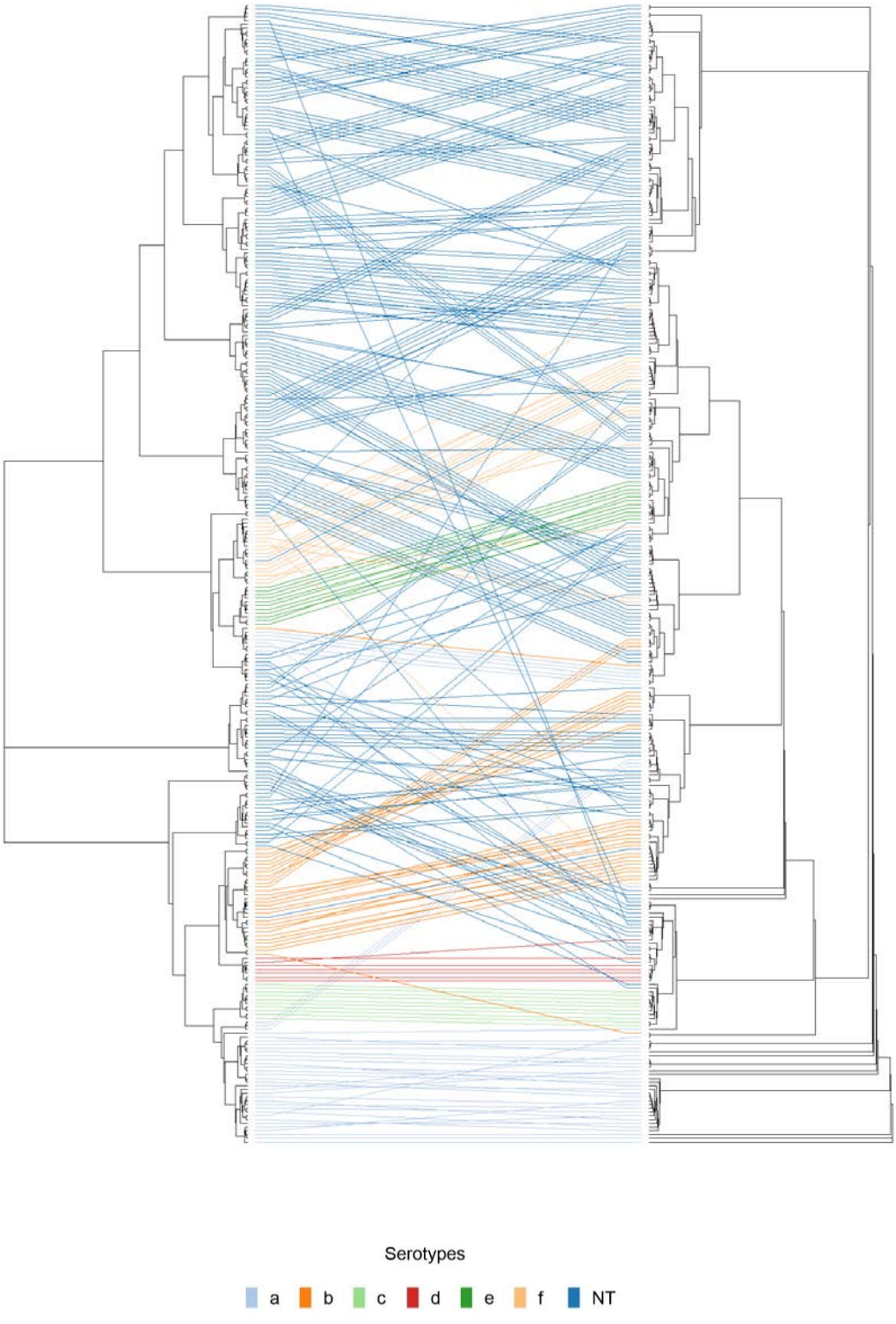

Serotypes
a
b
c
d
e
f
NT

Figure 13 A tanglegram generated using neighbour joining trees of SNV data (left), and rMLST profiles (right). Tanglegram connections were coloured by serotype.

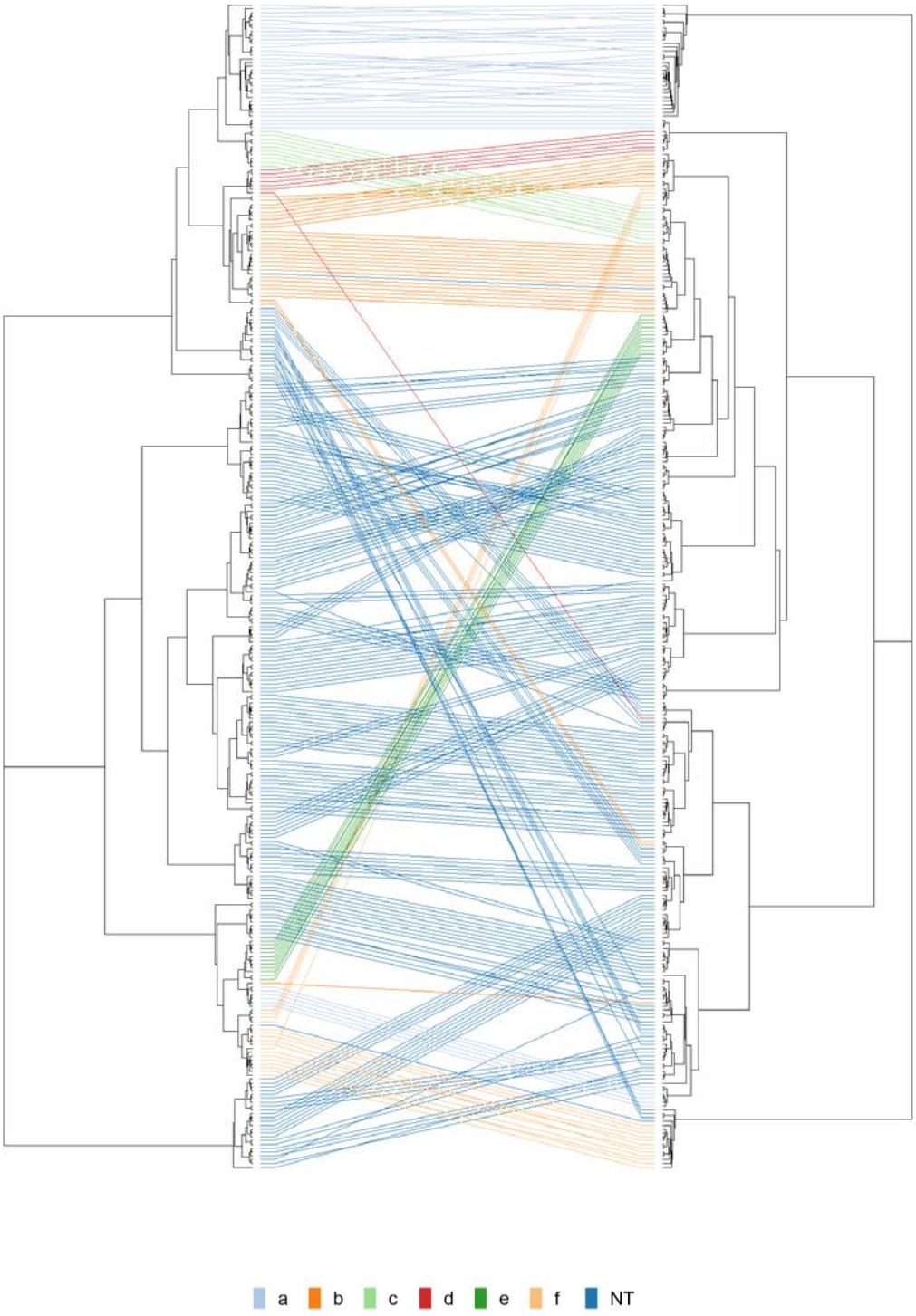

a
b
c
d
e
f
NT

Figure 14 A tanglegram generated using neighbour joining trees of SNV data (left), and cgMLST profiles (right). Tanglegram connections were coloured by serotype.

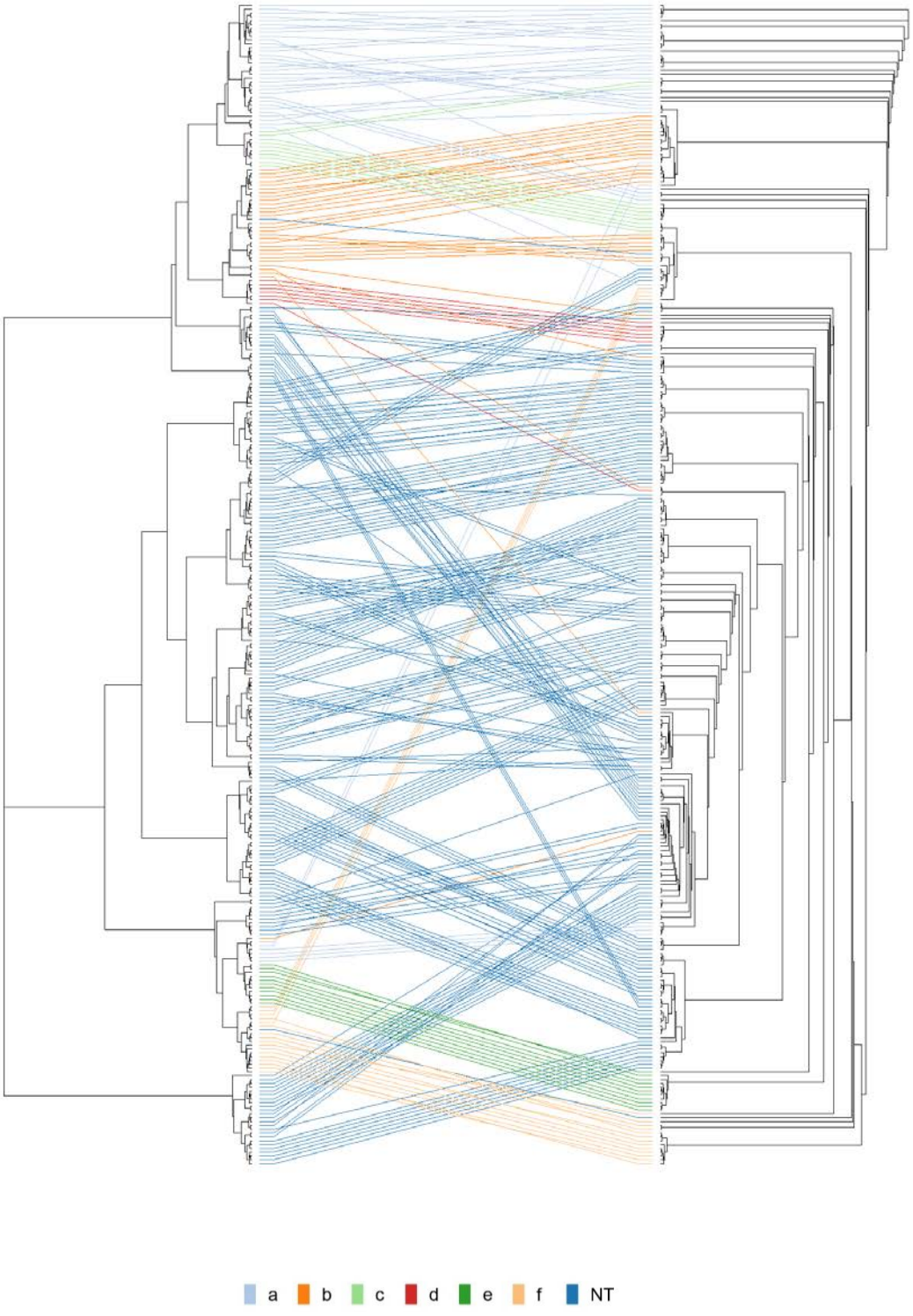
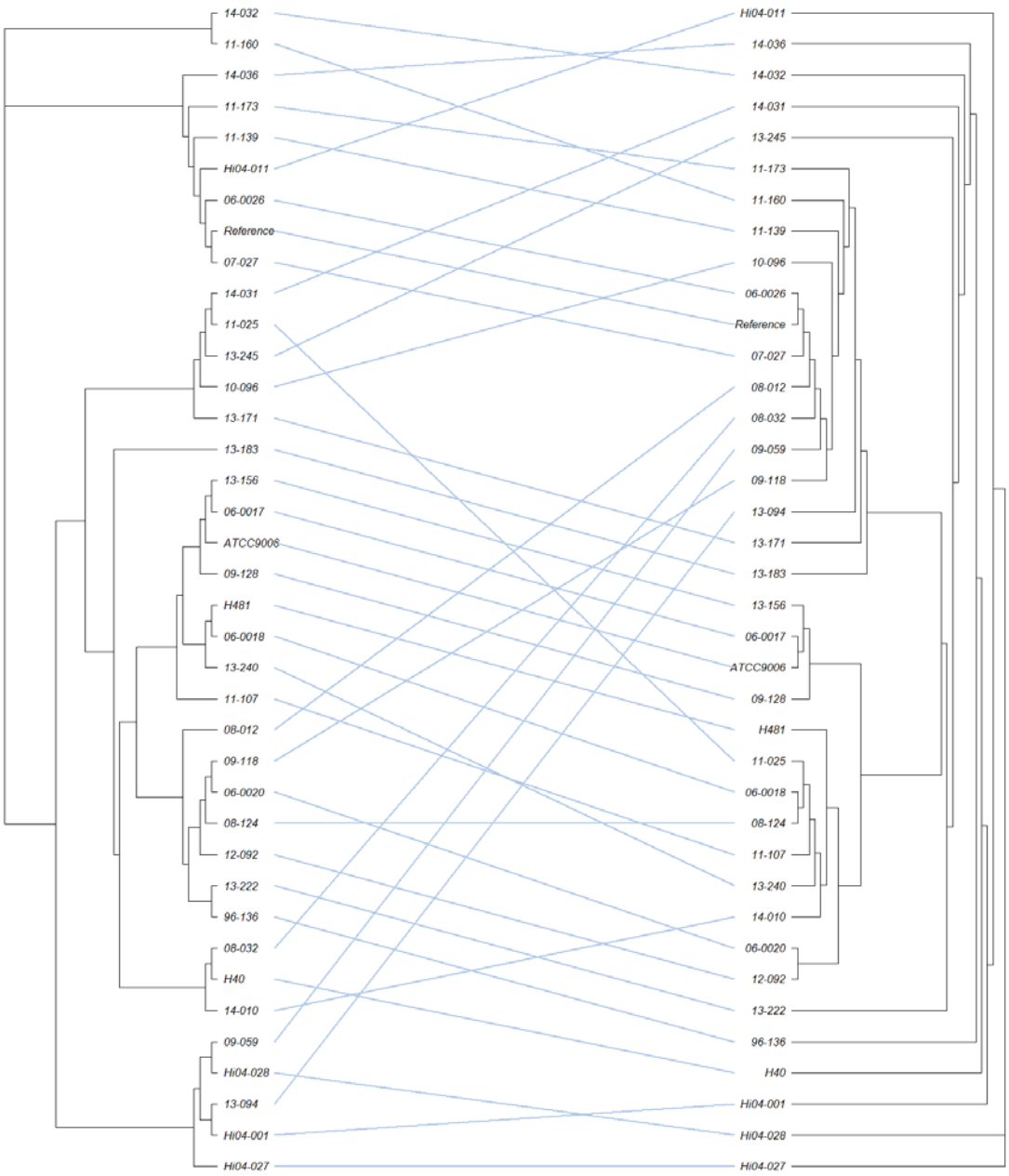

a
b
c
d
e
f
NT

Figure 15 A tanglegram generated using neighbour joining trees of Hia whole genome SNV data (left), and SNVs only based on the *cps* region (right).


14-032
11-160
14-036
11-173
11-139
HI04-011
06-0026
Reference
07-027
14-031
11-025
13-245
10-096
13-171
13-183
13-156
06-0017
ATCC9006
09-128
H481
06-0018
13-240
11-107
08-012
09-118
06-0020
08-124
12-092
13-222
96-136
08-032
H40
14-010
09-059
HI04-028
13-094
HI04-001
HI04-027
HI04-011
14-036
14-032
14-031
13-245
11-173
11-160
11-139
10-096
06-0026
Reference
07-027
08-012
08-032
09-059
09-118
13-094
13-171
13-183
13-156
06-0017
ATCC9006
09-128
H481
11-025
06-0018
08-124
11-107
13-240
14-010
06-0020
12-092
13-222
96-136
H40
HI04-001
HI04-028
HI04-027

Chapter 4 Discussion

4.1 The Molecular Epidemiology of *H. influenzae*

Haemophilus influenzae has been the cause of invasive and non-invasive disease globally. While this pathogen can be a commensal resident of the upper respiratory system, it can also be the cause of disease, mostly infecting young children and those with a weakened immune system. Historically, *H. influenzae* serotype b had been the cause of global outbreaks. A conjugate vaccine was introduced in the 1990's that drastically reduced the incidence and prevalence of Hib disease. However, non-Hib strains, including NTHi, have since replaced Hib as the most prevalent cause of disease. While NTHi and Hif are most predominant in the general population, Hia strains have been thriving in the Indigenous populations of North America and Australia. The population structure and diversity of *H. influenzae* is not fully known.

To assist in the epidemiological investigations of *H. influenzae*, and to research potential vaccines for non-Hib strains, it is essential to discern the population structure, and to develop a scheme for typing this pathogen. Whole genome sequencing has emerged as a powerful tool for surveillance and molecular typing of bacterial pathogens. In this study, I hypothesize that whole genome sequencing can be used to develop a cgMLST schema for the typing of *H. influenzae.* I also hypothesize that the cgMLST schema is concordant with existing molecular typing techniques, and has higher discriminatory power than existing methods.

4.1.1 Data Selection and Quality Control

To develop a typing scheme, a large dataset is optimal to ensure that the entire species is represented. In this study, only a set of 83 in-house sequenced genomes were available, and only serotypeable strains were represented, mostly serotypes a and b. To increase the size of the dataset, all publicly available genomes were downloaded either from SRA or PATRIC. While

genomes sequenced at the NML were high quality genomes, most of the reads downloaded from SRA or PATRIC were low quality and did not meet our minimum quality criteria. In contrast to the 7.2% of NML-sequenced genomes that were removed from the dataset, 40.6% of PATRIC genomes and 47% of SRA genomes were eliminated, either due to low quality or missing serotype or sequence type information. Unfortunately, this is a significant limitation for publicly available data.

4.1.2 The *in silico* Serotyping of *H. influenzae*

Another factor to consider is the evaluation of a new typing scheme. Typically, this evaluation is done by comparing the method's partitions to available epidemiological and outbreak data. In addition to the low quality, virtually no metadata was available for the publicly-available strains. For this reason, *in silico* serotyping was done to provide a starting reference for schema evaluation. Serotyping is traditionally done in a laboratory using either SAST or PCR, and there are currently no validated methods for *in silico* serotyping of *H. influenzae*. Two methods were used to perform *in silico* serotyping: 1) the *cps* region alignment method, and 2) the MLST method.

Using a set of 82 strains that were traditionally serotyped at the NML, the MLST method outperformed the *cps* alignment method. Using the MLST method, 92.7% of strains were correctly serotyped, but only 74.4% of strains were correctly typed using the *cps* alignment method. However, since neither method has been validated outside of this work, and the sample size is relatively small, consensus from both methods were used to serotype the full dataset, and genomes with conflicting serotype assignments (n = 145) were removed.

4.2 Development and Evaluation of the cgMLST Schema

In the present study, a core genome MLST schema was developed and validated as a potential typing method for *H. influenzae*. To define a high-quality schema, stringent quality criteria were set to eliminate low quality genomes, or genomes with missing information (Table 5). Since the 7-gene MLST ST designations were essential for evaluating the cgMLST schema, all 555 genomes in the dataset were typed *in silico*, and any genomes missing a ST or having partial matches to existing alleles were removed from the dataset (n = 73). In addition to the MLST criteria, other criteria were also used to filter low quality genomes. For instance, genomes with a N50 contig length of 50,000 bp or less were removed (n = 95), genomes under minimum size of 1.62 Mb or over a maximum size of 2.07 Mb were removed (n = 60), and laboratory transformed genomes were also removed (n = 46). Finally, due to obtaining sequences from multiple sources, some duplicated genomes were found in the dataset (n = 67), and were removed by cross-referencing the biosample accession numbers.

The pgMLST schema was defined using 314 genomes that passed quality control using chewBBACA. The schema is built by first identifying CDSs in all genomes, and using a recommended BSR value of 0.6 to cluster alleles into loci (n = 3824). To define a high quality cgMLST, only genomes missing less than 20 loci—which are present in at least 95% of all genomes—were used for schema definition. Paralogous genes were identified (n = 135), and removed from the pool of loci. Finally, a minimum cut-off was set to ensure that loci that are missing from a large proportion of the genomes are not included in the dataset. Minimum spanning trees were built using eight cut-offs, ranging from 65% minimum percentage of genomes having a given locus to 100% (Figure 6). While the tree topology was observed to

stabilize around a value of 85%, ultimately, a cut-off of 100% was chosen to define a true cgMLST schema. The final cgMLST schema is comprised of 980 loci, out of an average of 1,800 genes typically found in the *H. influenzae* genome (Hogg et al. 2007; Kappler et al. 2017). The schema was developed using 304 genomes, which were partitioned into 204 CTs.

To evaluate the cgMLST schema, the 7-gene MLST and rMLST schemas were used to compare the concordance and discriminatory power between all three methods. A 53-locus rMLST schema was downloaded from PubMLST, and alleles were called in the 304 genomes in the final cgMLST schema. However, some loci were missing from 1 or more genomes, and the final rMLST schema only had 49 loci. The 7-gene MLST schema partitioned the 304 genomes into 144 STs, while the rMLST schema had 148 partitions. The concordance and discriminatory power of all three schemas, as well as serotyping, were compared by calculating the Simpson's index, the adjusted Rand coefficient and the adjusted Wallace coefficient (Table 8, 9, and 10).

The validation of the cgMLST schema was complicated by the lack of epidemiological data associated with the publicly available strains. Epidemiological data could be used to compare the concordance between the schema clustering and outbreak data, such as strains grouped by geographical location or a specific period of time. This data can also be used to determine clustering thresholds when defining the schema. In the absence of this data, the schema was instead evaluated by using statistical methods to determine the discriminatory power of the typing schema, as well as comparisons for concordance with the 7-gene MLST schema, rMLST schema, and SNV-based phylogeny.

The Simpson diversity index (Table 8), the adjusted Rand coefficient (Table 9) and the adjusted Wallace coefficient (Table 10) were calculated for all three schemas (Severiano, Pinto,

et al. 2011; Hunter and Gaston 1988; Rand 1971). Both the Simpson index and the adjusted Rand coefficient found all three MLST schemas to be equally discriminatory. The Simpson's index measures the probability that two strains picked at random will belong to the same cluster. While this method is often used to measure the discriminatory power of typing methods, Simpson's index is heavily weighted by the most abundant type, and rare types have very little effect on the index (R. H. Whittaker 1972). For instance, in the cgMLST schema, CT-1 is the most abundant type, containing only serotype a strains, followed by CT-2 containing only NTHi strains. The Simpson index for the cgMLST will, therefore, be heavily weighted by those two groups, while other CTs may not be properly represented by the index. This bias is accentuated by the uneven serotype representation in the dataset. Curiously, p-values calculated using the jackknife resampling approach indicated that the discriminatory power of the cgMLST schema is significantly higher ($p < 0.001$) than the 7-gene MLST and rMLST schemas, despite the slight overlapping between the 95% confidence intervals among the groups.

The adjusted Rand score measures the overall concordance between two methods, but provides no directionality, i.e. the concordance of schema *A* to schema *B* is the same as the concordance of schema *B* to schema *A*. The 7-gene MLST and rMLST schemas have the highest adjusted Rand score, indicating high congruence between the two methods. The cgMLST schema, on the other hand, had a lower adjusted Rand score compared to the 7-gene MLST and rMLST schemes, however, this score does not indicate which of the methods has the higher discriminatory power, since the, the coefficient is not directional. In contrast, the adjusted Wallace coefficient is more informative, since it is directional, and can determine which of the two methods being compared is more discriminatory. Hence, the adjusted Wallace coefficient was calculated to show directional congruence. The results indicate that the cgMLST is

statistically a more discriminatory typing method compared to serotyping, the 7-gene MLST and the rMLST schemas. Taken together, these results reveal that while the cgMLST schema is not significantly congruent with the 7-gene MLST or the rMLST schemas, it is the most discriminatory typing method.

To confirm the cgMLST schema's discriminatory ability, a single nucleotide variants (SNV) analysis was performed using the SNVPhyl pipeline (Petkau et al. 2017). Since SNVPhyl is a highly validated method used for analysing bacterial population structure, the pipeline was used to estimate a phylogeny that most resembles the population structure of *H. influenzae*. Due to the diverse nature of *H. influenzae*, performing the SNV analysis was challenging. The SNVPhyl pipeline detects sequence regions with high SNV density, and masks those regions as it is an indication of possible recombination. Using a value between 2 bp and 20 bp for SNV density in a 400 bp window resulted in phylogenetic trees with poor support since they were constructed using only a handful of SNVs. On the other hand, using values over 30 bp resulted in an analysis using a large number of SNVs—over 150,000—and phylogenetic trees were not built due to the excessive computational requirement requirement. A value of 25 bp over a 400 bp window was chosen, and a phylogenetic tree was generated using 26,888 SNVs.

The SNV phylogeny (Figure 11) highlights the discriminatory ability of the cgMLST schema. Most of the 7-genome MLST schema STs were further subdivided by the cgMLST schema, while the reverse was not true, with some minor exceptions. Additionally, tanglegrams between each of the MLST-based schemas and the SNV-based neighbour joining trees were plotted. Although the tree topography of the cgMLST neighbour joining tree was different than that of the SNV-based tree, the same group of strains are seen clustered in both trees (Figure 14). In contrast, the 7-gene MLST tanglegram (Figure 12) reveals more scattering of strains,

particularly in non-typeable *H. influenzae*. The rMLST tanglegram (Figure 13) shows higher concordance to the SNV-based neighbour joining tree than the 7-gene MLST schema, but lower concordance compared to the cgMLST schema. Of all three tanglegrams, the cgMLST maintains the highest congruence against the SNV data, followed by the rMLST schema, and the 7-gene MLST schema.

4.3 *H. influenzae* Population Structure

The population structure of *H. influenzae* was investigated using minimum spanning trees. Trees were generated for the 7-gene MLST, rMLST and cgMLST schemas using the goeBURST algorithm (Figures 8, 9, and 10). Both the 7-gene MLST and rMLST were built using SLV, while the cgMLST schema was generated using a locus variant level 25. All three trees exhibited the same subclustering patterns for each of the serotypes, however, cluster placements on the tree differed among the three methods. The NTHi strains were the largest group in all three trees. Serotypes c, d e, and f each formed distinct clusters, while serotypes a and b strains each formed one large group, with a few strains scattered around the trees.

The cgMLST tree had four distinct branches; one branch contained only NTHi strains, one branch contained Hib strains along with NTHi strains, one branch mostly had serotypes e and f, while the last branch held serotypes a, c and d. While most Hia strains clustered on the same branch as Hic and Hid, a small set of strains were clustered with the Hie strains. Similarly, Hib stains were clustered with Hic and Hie. This is a possible indication of serotype switching occurring amongst serotyepable strains. If true, this exposes a severe limitation of traditional serotyping in that it may not represent the true population structure of *H. influenzae*.

Additionally, all branches had NTHi strains, which is an indication that serotypeable strains may not have descended from a common ancestor, a finding supported by St. Geme et al. (1994a).

The 7-gene MLST tree was also observed to have four distinct branches, however, the relationships between different clades did not correspond to the cgMLST tree. Serotype c and NTHi strains are found on one branch, Hid and NTHi found on another branch, Hib and Hie cluster together on the third branch, and Hia and Hif are grouped on the last branch. As in the cgMLST tree, some Hia strains are found clustered with Hif and Hib, while some Hib strains are found clustered with Hia and Hif. In contrast to the cgMLST and 7-gene MLST schemes, the rMLST tree had more ambiguous branches. Serotypes a and b were found to be grouped in close proximity, which contradicts the cgMLST and 7-gene MLST trees. However, there were some consistencies between the rMLST and the cgMLST trees. For example, Hif and Hie are grouped together, along with some Hia and Hib strains, and Hic and Hid were grouped on the same branch, also with some Hia and Hib strains.

Ribosomal proteins are typically used in bacterial typing since there is little to no expected recombination in those proteins, which could be an advantage over other schemas. While recombinogenic loci are likely included in the cgMLST schema, the effects of these loci should be minimal since a large number of loci are used. Only 49 loci were used in the rMLST schema, which could account for the differences observed between the rMLST and the cgMLST schemas. In addition, the rMLST schema partitioned the 304 genomes into 148 partitions, only four more partitions than the 7-gene MLST schema. The discriminatory power and concordance of the 7-gene MLST and the rMLST were found to be approximately similar with the Simpson index and the adjusted Rand coefficient, however, the adjust Wallace coefficient indicates that the discriminatory power of the 7-gene MLST schema is higher than that of the rMLST

schema,which is counterintuitive. These results are an indication that the rMLST schema is an questionable method for typing *H. influenzae.*

Previous studies using MLEE found that Hia and Hib typically have different lineages, whereas serotypes c, d, e and f formed a monophyletic group (Ulanova and Tsang 2014). While serotypes a and b have different lineages according to the cgMLST minimum spanning trees, the remaining serotypes do not form a monophyletic group. In fact, even 7-gene MLST data contradicts the MLEE findings in the literature (Ulanova and Tsang 2014; Hardy, Tudor, and Geme 2003; Musser et al. 1988; Meats et al. 2003). These contradicting results could be an indication of the low discriminatory power, and corresponding inadequacy of MLEE as a typing scheme for *H. influenzae,* compared to the MLST schemas.

Non-typeable *H. influenzae* forms a more diverse and recombinogenic phylogeny (Cody et al. 2003). The diversity of NTHi is further emphasised by the difficulties encountered when generating a SNV-based phylogeny. The SNV-based phylogeny indicates that the NTHi population is sub-divided into two distinct groups, consistent with what is described in the literature (Meats et al. 2003). In the SNV-based phylogeny, serotypes e and f were found to be clustered, along with a subset of Hia and Hib strains, which supports the findings from the cgMLST scheme minimum spanning tree. Serotypes a, b, c and d are grouped together. While the clustering of Hia, Hic and Hid is consistent with the cgMLST schema minimum spanning tree, the grouping of Hia and Hib contradicts the cgMLST tree topology. Since lenient parameters for masking recombination regions were used in SNVPhyl, the SNV-based phylogeny is likely affected by recombination. In contrast, using the entire genes as loci in the cgMLST schema likely masks recombination regions, which can account for the differences in tree topology between the cgMLST tree and the SNV-based tree.

4.4 Recombination in the Hia *cps* Region

Serotype a disease has been on the rise, particularly in the Indigenous populations of North America. Due to the similarities in virulence factors between Hia and Hib, the concern over the outbreak-causing potential of Hia grows. Since the Hib conjugate vaccine does not protect against Hia, there is a need to research and develop a vaccine for this serotype to stem future outbreaks. Since a vaccine for Hia will likely target the *cps* region (Desai et al. 2014), a knowledge of the structure and diversity of the *cps* region is important. To determine whether there is recombination in the *cps* region of Hia, SNV-based analysis was performed using the Hia whole genome and the Hia *cps* region. Phylogenetic trees from both analyses were compared, and the two trees were found to have differing topologies (Figure 15), which may be an indication of recombination in the Hia *cps* region. This capsular diversity in Hia can present challenges in developing a vaccine against this serotype. If recombination is common in the *cps* region of Hia, then future vaccines based on the *cps* may not protect against all Hia strains, and selection pressures on the population may cause the vaccine to be eventually ineffective.

4.5 Limitations of the Study

A major limitation in this study is the limited number of isolates used to define the cgMLST schema. Since many isolates were downloaded from public databases, the quality of many of the reads downloaded from SRA was low. Despite having an initial dataset of 555 genomes, only 304 passed all the quality criteria. Core genome MLST schemas defined for other organisms typically include between 1000 and 5000 isolates (Moura et al. 2016; Yoshida et al. 2016; Santona et al. 2016; Kluytmans-Van Den Bergh et al. 2016; Zhou et al. 2017). Additionally, not all subgroups were proportionately represented in the dataset; non-typeable strains made up 62.8% of all genomes in the schema, and serotypes c, d and e made up 3.3%, 2.3% and 3.6%,

respectively. Due to this imbalance, the cgMLST schema may not be representative of the true diversity of *H. influenzae.*

Another limitation is the lack of metadata or epidemiological data for most of the genomes used to define the cgMLST schema. Without this data, evaluation of the schema was done using only statistical methods and comparisons of tree topologies. Furthermore, clustering of the cgMLST was done using alternative methods, such as the cluster stability method. Serotyping of genomes missing data was done using two *in silico* methods. While the *cps* method could potentially be the most accurate, since it relies on identifying the serotype-specific genes in the *cps* region, the method has limitations. For instance, the presence of serotype-specific genes does not necessarily indicate that the strain is capsulated, since spontaneous capsule loss has been reported in *H. influenzae*. In addition, attempting to serotype draft genomes could be problematic if the *cps* region is not covered by sequencing – a false NTHi could be assigned to a capsulated strain.

4.6 Conclusions and Future Direction

In this thesis, we have shown that cgMLST schema has tremendous potential to become a typing scheme for *H. influenzae*. The 980-locus cgMLST schema has a much higher discriminatory power in *H. influenzae* typing compared to the 7-gene MLST schema. If validated, this schema can replace multiple traditional typing methods that can be tedious and time consuming, and enhance the workflow for studying outbreaks.

To define a more robust schema, more high-quality *H. influenzae* isolates need to be sequenced, and more epidemiological data should be available to guide with validation and interpretation. Additionally, a larger number of serotypeable strains should be used to ensure that

the diversity of *H. influenzae* is properly represented. In addition to the increased discrimination, the cgMLST schema can be used for *in silico* serotyping of *H. influenzae*. The cgMLST schema had a 97.7% chance of correctly assigning a serotype, as opposed to the 95.2% of the 7-gene MLST schema. The cgMLST schema has the potential for serotyping *H. influenzae* strains, similar to the *in silico* serotyping method being used for *Salmonella* (Yoshida et al. 2016), however, more studies and validation are required. Future studies should also evaluate the schema's ability to distinguish between Hib, Hib$^-$ and NTHi strains. It has been shown that *H. haemolyticus* is often mistaken with NTHi, due to the low discriminatory power of current methods (Nørskov-Lauritsen 2009). The cgMLST schema's potential to distinguish between *H. influenzae* and *H. haemolyticus* should be studied and evaluated.

References

Adderson, E. E., C. L. Byington, L. Spencer, A. Kimball, M. Hindiyeh, K. Carroll, S. Mottice, E. K. Korgenski, J. C. Christenson, and A. T. Pavia. 2001. "Invasive Serotype a *Haemophilus Influenzae* Infections With a Virulence Genotype Resembling *Haemophilus Influenzae* Type B: Emerging Pathogen in the Vaccine Era?" *Pediatrics* 108 (1). American Academy of Pediatrics: e18–e18. doi:10.1542/peds.108.1.e18.

Allali, Slimane, Martin Chalumeau, Odile Launay, Samir K. Ballas, and Mariane de Montalembert. 2016. "Conjugate *Haemophilus Influenzae* Type B Vaccines for Sickle Cell Disease." *Cochrane Database of Systematic Reviews* 2016 (2). doi:10.1002/14651858.CD011199.pub2.

Anyanwu, Juliana N., Carina A. Rodriguez, Katherine E. Fleming, and Elisabeth E. Adderson. 2003. "Pgi Genotyping Is a Surrogate for Serotyping of Encapsulated Haemophilus Influenzae." *Journal of Clinical Microbiology* 41 (5): 2080–83. doi:10.1128/JCM.41.5.2080-2083.2003.

Aubrey, Ruth, and Christoph Tang. 2003. "The Pathogenesis of Disease Due to Type B Haemophilus Influenzae." In *Haemophilus Influenzae Protocols*, edited by Mark A Herbert, Derek W Hood, and E. Richard Moxon, 71:29–50. Totowa, NJ, NJ: Humana P. http://link.springer.com/content/pdf/10.1385/1-59259-321-6:29.pdf%5Cnhttp://link.springer.com/protocol/10.1385/1-59259-321-6:29.

Barrett, Douglas J., John W. Sleasman, Desmond A. Schatz, and Michael Steinitz. 1992. "Human Anti-Pneumococcal Polysaccharide Antibodies Are Secreted by the CD5- B Cell Lineage." *Cellular Immunology* 143 (1): 66–79. doi:10.1016/0008-8749(92)90006-B.

Bayliss, Christopher D., Dawn Field, and E. Richard Moxon. 2001. "The Simple Sequence Contingency Loci of Haemophilus Influenzae and Neisseria Meningitidis." *Journal of Clinical Investigation* 107 (6): 657–62. doi:10.1172/JCI12557.

Beck, James M., Vincent B. Young, and Gary B. Huffnagle. 2012. "The Microbiome of the Lung." *Translational Research* 160 (4). Mosby, Inc: 258–66. doi:10.1016/j.trsl.2012.02.005.

Brown, Veronica M, Sharen Madden, Len Kelly, Frances B Jamieson, Raymond S W Tsang, and Marina Ulanova. 2009. "Invasive Haemophilus Influenzae Disease Caused by Non-Type B Strains in Northwestern Ontario, Canada, 2002-2008." *Clinical Infectious Diseases : An Official Publication of the Infectious Diseases Society of America* 49 (8): 1240–43. doi:10.1086/605671.

Bruce, Michael G., Tammy Zulz, Carolynn DeByle, Ros Singleton, Debby Hurlburt, Dana Bruden, Karen Rudolph, Thomas Hennessy, Joseph Klejka, and Jay D. Wenger. 2013. "Haemophilus Influenzae Serotype a Invasive Disease, Alaska, USA, 1983-2011." *Emerging Infectious Diseases* 19 (6). Centers for Disease Control and Prevention: 932–37. doi:10.3201/eid1906.121805.

Camacho, Christiam, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer, and Thomas L Madden. 2009. “BLAST+: Architecture and Applications.” *BMC Bioinformatics* 10 (1): 421. doi:10.1186/1471-2105-10-421.

Capeding, Maria Rosario, Josefina Cadorna-Carlos, May Book-Montellano, and Esteban Ortiz. 2008. “Immunogenicity and Safety of a DTaP–IPV//PRP~T Combination Vaccine given with Hepatitis B Vaccine: A Randomized Open- Label Trial.” *Bulletin of the World Health Organization* 86 (6). doi:10.2471/BLT.07.042143.

Cerquetti, Marina, Rita Cardines, Maria Giufrè, Annalisa Castella, Monica Rebora, Paola Mastrantonio, and Marta Luisa Ciofi Degli Atti. 2006. “Detection of Six Copies of the Capsulation B Locus in a *Haemophilus Influenzae* Type B Strain Isolated from a Splenectomized Patient with Fulminant Septic Shock.” *Journal of Clinical Microbiology* 44 (2): 640–42. doi:10.1128/JCM.44.2.640-642.2006.

Cerquetti, Marina, Rita Cardines, Maria Giufre, Tonino Sofia, Fabio D Ambrosio, Paola Mastrantonio, and Marta Luisa. 2006. “Genetic Diversity of Invasive Strains of Haemophilus Influenzae Type B before and after Introduction of the Conjugate Vaccine in Italy” 43 (October): 317–19. doi:10.1086/505499.

Cerquetti, Marina, Rita Cardines, Marta Luisa, Maria Giufre, Antonino Bella, Tonino Sofia, Paola Mastrantonio, and Mary Slack. 2005. “Presence of Multiple Copies of the Capsulation B Locus in Invasive Haemophilus Influenzae Type B (Hib) Strains Isolated from Children with Hib Conjugate Vaccine Failure.” *J Infect Dis* 192 (5): 819–23. doi:JID34086 [pii] 10.1086/432548.

Chin, Chen-Shan, David H Alexander, Patrick Marks, Aaron A Klammer, James Drake, Cheryl Heiner, Alicia Clum, et al. 2013. “Nonhybrid, Finished Microbial Genome Assemblies from Long-Read SMRT Sequencing Data.” *Nature Methods* 10 (6). Nature Research: 563–69. doi:10.1038/nmeth.2474.

Cintra, Felipe De Oliveira, and Mickie Takagi. 2015. “Study of the Chemical Stability of the Capsular Polysaccharide Produced by *Haemophilus Influenzae* Type B.” *Carbohydrate Polymers* 116 (February): 167–72. doi:10.1016/j.carbpol.2014.04.004.

Cleland, Gavin, Clare Leung, Jenny Wan Sai Cheong, Joshua Francis, Claire Heney, and Clare Nourse. 2017. “Paediatric Invasive *Haemophilus Influenzae* in Queensland, Australia, 2002-2011: Young Indigenous Children Remain at Highest Risk.” *Journal of Paediatrics and Child Health*, no. June. doi:10.1111/jpc.13662.

Cody, Alison J., Dawn Field, Edward J. Feil, Suzanna Stringer, Mary E. Deadman, Anthony G. Tsolaki, Brett Gratz, et al. 2003. “High Rates of Recombination in Otitis Media Isolates of Non-Typeable *Haemophilus Influenzae*.” *Infection, Genetics and Evolution* 3 (1): 57–66. doi:10.1016/S1567-1348(02)00152-1.

Corn, P G, J Anders, A K Takala, H Kayhty, and S K Hoiseth. 1993. “Genes Involved in *Haemophilus Influenzae* Type B Capsule Expression Are.” *Journal Of Infectious Diseases* 167 (2): 356–64.

Cress, Brady F., Jacob A. Englaender, Wenqin He, Dennis Kasper, Robert J. Linhardt, and Mattheos A.G. Koffas. 2014. “Masquerading Microbial Pathogens: Capsular Polysaccharides Mimic Host-Tissue Molecules.” *FEMS Microbiology Reviews* 38 (4): 660–97. doi:10.1111/1574-6976.12056.

Crisel, R. M., R. S. Baker, E Dorman, and D. E. Dorman. 1975. “Capsular Polymer of *Haemophilus Influenzae*, Tybe B.” *Journal of Biological Chemistry* 250 (13): 4926–30. http://www.jbc.org.uml.idm.oclc.org/content/250/13/4926.short.

Desai, S, R Tsang, M St. Laurent, and A Cox. 2014. “Collaboration on a Public Health Driven Vaccine Initiative.” *Canadian Communicable Disease Report* 40 (17): 365–68. https://www.phac-aspc.gc.ca/publicat/ccdr-rmtc/14vol40/dr-rm40-17/dr-rm40-17-vaccin-eng.php.

Dworkin, M. S., L. Park, and S. M. Borchardt. 2007. “The Changing Epidemiology of Invasive *Haemophilus Influenzae* Disease, Especially in Persons >=65 Years Old.” *Clinical Infectious Diseases* 44 (6): 810–16. doi:10.1086/511861.

Eldere, Johan Van, Lisa Brophy, Barbara Loynds, Patrick Celts, Ian Hancock, Stephen Carman, J. Simon Kroll, and E. Richard Moxon. 1995. “Region II of the *Haemophilus Influenzae* Type B Capsulation Locus as Involved in Serotype-specific Polysaccharide Synthesis.” *Molecular Microbiology* 15 (1): 107–18. doi:10.1111/j.1365-2958.1995.tb02225.x.

Eldere, Johan Van Van, Mary P. E. Slack, S. Ladhani, and Allan W. Cripps. 2014. “Non-Typeable Haemophilus Influenzae, an under-Recognised Pathogen.” *The Lancet Infectious Diseases* 3099 (14): 0–29.

Evans, N. M., and D. D. Smith. 1972. “The Effect of the Medium and Source of Growth Factors on the Satellitism Test for Haemophilus Species.” *Journal of Medical Microbiology* 5 (4). Microbiology Society: 509–14. doi:10.1099/00222615-5-4-509.

Evans, N. M., D. D. Smith, and A. J. Wicken. 1974. “Haemin and Nicotinamide Adenine Dinucleotide Requirements of *Haemophilus Influenzae* and *Haemophilus Parainfluenzae*.” *Journal of Medical Microbiology* 7 (3). Microbiology Society: 359–65. doi:10.1099/00222615-7-3-359.

Falla, T. J., D. W M Crook, L. N. Brophy, D. Maskell, J. S. Kroll, and E. R. Moxon. 1994. “PCR for Capsular Typing of *Haemophilus Influenzae*.” *Journal of Clinical Microbiology* 32 (10): 2382–86. http://jcm.asm.org.uml.idm.oclc.org/content/32/10/2382.short.

Feil, E J, E C Holmes, D E Bessen, M S Chan, N P Day, M C Enright, R Goldstein, et al. 2001. “Recombination within Natural Populations of Pathogenic Bacteria: Short-Term Empirical Estimates and Long-Term Phylogenetic Consequences.” *Proceedings of the National Academy of Sciences of the United States of America* 98 (1). National Academy of Sciences: 182–87. doi:10.1073/pnas.98.1.182.

Feil, Edward J., Bao C. Li, David M. Aanensen, William P. Hanage, and Brian G. Spratt. 2004. “eBURST: Inferring Patterns of Evolutionary Descent among Clusters of Related Bacterial

Genotypes from Multilocus Sequence Typing Data." *Journal of Bacteriology* 186 (5). American Society for Microbiology (ASM): 1518–30. doi:10.1128/JB.186.5.1518-1530.2004.

Follens, Anja, Maria Veiga-da-cunha, Rita Merckx, Emile Van, Johan Van Eldere, and Emile V A N Schaftingen. 2001. "acs1 of Haemophilus Influenzae Type a Capsulation Locus Region II Encodes a Bifunctional Ribulose 5-Phosphate Reductase– CDP-Ribitol Pyrophosphorylase." *J. Bacteriol.* 181 (7): 2001–7. http://jb.asm.org.uml.idm.oclc.org/content/181/7/2001.short.

Francisco, Alexandre P, Miguel Bugalho, Mário Ramirez, and João A Carriço. 2009. "Global Optimal eBURST Analysis of Multilocus Typing Data Using a Graphic Matroid Approach." *BMC Bioinformatics* 10 (1): 152. doi:10.1186/1471-2105-10-152.

Frosch, M., U. Edwards, K. Bousset, B. Krauße, and C. Weisgerber. 1991. "Evidence for a Common Molecular Origin of the Capsule Gene Loci in Gram-negative Bacteria Expressing Group II Capsular Polysaccharides." *Molecular Microbiology* 5 (5): 1251–63. doi:10.1111/j.1365-2958.1991.tb01899.x.

Fry, A M, P Lurie, M Gidley, S Schmink, J Lingappa, M Fischer, and N E Rosenstein. 2001. "*Haemophilus Influenzae* Type B Disease among Amish Children in Pennsylvania: Reasons for Persistent Disease." *Pediatrics* 108 (4): E60. doi:10.1542/peds.108.4.e60.

Geme, J. W. St., A. Takala, E. Esko, and S. Falkow. 1994. "Evidence for Capsule Gene Sequences among Pharyngeal Isolates of Nontypeable Haemophilus Influenzae." *Journal of Infectious Diseases* 169 (2). Oxford University Press: 337–42. doi:10.1093/infdis/169.2.337.

Grundmann, H., S. Hori, and G. Tanner. 2001. "Determining Confidence Intervals When Measuring Genetic Diversity and the Discriminatory Abilities of Typing Methods for Microorganisms." *Journal of Clinical Microbiology* 39 (11). American Society for Microbiology: 4190–92. doi:10.1128/JCM.39.11.4190-4192.2001.

Habermehl, Pirmin, Geert Leroux-Roels, Roland Sänger, Gudrun Mächler, and Dominique Boutriau. 2010. "Combined Haemophilus Influenzae Type B and Neisseria Meningitidis Serogroup C (HibMenC) or Serogroup C and Y-Tetanus Toxoid Conjugate (and HibMenCY) Vaccines Are Well-Tolerated and Immunogenic When Administered according to the 2,3,4 Months Schedule with ." *Human Vaccines* 6 (8): 640–51. doi:10.4161/hv.6.8.12154.

Hallström, Teresia, and Kristian Riesbeck. 2010. "*Haemophilus Influenzae* and the Complement System." *Trends in Microbiology* 18 (6): 258–65. doi:10.1016/j.tim.2010.03.007.

Hamborsky, J, A Kroger, and S Wolfe, eds. 2015. "Haemophilus Influenzae Type B." In *Center for Disease Control and Prevention: Epidemiology and Prevention of Vaccine-Preventable Diseases 13th Edition*, 119–34. Washington, DC, DC. https://www.cdc.gov/vaccines/pubs/pinkbook/hib.html.

Hardy, Gail G, Simone M Tudor, and Joseph W St Geme. 2003. "The Pathogenesis of Disease due to Nontypeable Haemophilus Influenzae." In *Haemophilus Influenzae Protocols*, edited by Mark A Herbert, Derek W Hood, and E. Richard Moxon, 71:1–28. Totowa, NJ, NJ: Humana P.

Hargreaves, R M, M P Slack, A J Howard, E Anderson, and M E Ramsay. 1996. “Changing Patterns of Invasive Haemophilus Influenzae Disease in England and Wales after Introduction of the Hib Vaccination Programme.” *BMJ (Clinical Research Ed.)* 312 (7024): 160–61. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2349799&tool=pmcentrez&rendertype=abstract.

High, Nicola J., Feinan Fan, and Joseph D. Schwartzman. 2015. *Haemophilus Influenzae*. *Molecular Medical Microbiology*. Elsevier Ltd. doi:10.1016/B978-0-12-397169-2.00097-4.

Hirschmann, J V, and E D Everett. 1979. “*Haemophilus Influenzae* Infections in Adults: Report of Nine Cases and a Review of the Literature.” *Medicine*. http://journals.lww.com.uml.idm.oclc.org/md-journal/Citation/1979/01000/Haemophilus_Influenzae_Infections_in_Adults_.5.aspx.

Hogg, Justin S, Fen Z Hu, Benjamin Janto, Robert Boissy, Jay Hayes, Randy Keefe, J Christopher Post, and Garth D Ehrlich. 2007. “Characterization and Modeling of the *Haemophilus Influenzae* Core and Supragenomes Based on the Complete Genomic Sequences of Rd and 12 Clinical Nontypeable Strains.” *Genome Biology* 8 (6): R103. doi:10.1186/gb-2007-8-6-r103.

Hoiseth, S. K., C. J. Connelly, and E. R. Moxon. 1985. “Genetics of Spontaneous, High-Frequency Loss of B Capsule Expression in *Haemophilus Influenzae*.” *Infection and Immunity* 49 (2): 389–95.

Hoiseth, Susan K, E R Moxon, and Richard P Silver. 1986. “Genes Involved in *Haemophilus Influenzae* Type B Capsule Expression Are Part of an 18-Kilobase Tandem Duplication.” *Proceedings of the National Academy of Sciences of the United States of America* 83 (4): 1106–10. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=323020&tool=pmcentrez&rendertype=abstract.

Howard, A J, K T Dunkin, and Millar G W. 1988. “Nasopharyngeal Carriage and Antibiotic Resistance of Haemophilus Influenzae in Healthy Children.” *Epidemiology and Infection* 100: 193–203. http://www.bioline.org.br/pdf?mb09029.

Hubert, Lawrence, and Phipps Arabie. 1985. “Comparing Partitions.” *Journal of Classification* 2 (1). Springer-Verlag: 193–218. doi:10.1007/BF01908075.

Hunter, P. R., and M. A. Gaston. 1988. “Numerical Index of the Discriminatory Ability of Typing Systems: An Application of Simpson’s Index of Diversity.” *Journal of Clinical Microbiology* 26 (11): 2465–66. doi:0095-1137/88/112465-02$02.00/0.

Huson, Daniel H, and Celine Scornavacca. 2012. “Software for Systematics and Evolution Dendroscope 3: An Interactive Tool for Rooted Phylogenetic Trees and Networks.” *Syst. Biol* 61 (6): 1061–67. doi:10.1093/sysbio/sys062.

Hyatt, Doug, Gwo-Liang Chen, Philip F Locascio, Miriam L Land, Frank W Larimer, and Loren J Hauser. 2010. “Prodigal: Prokaryotic Gene Recognition and Translation Initiation Site

Identification." *BMC Bioinformatics* 11 (March). BioMed Central: 119. doi:10.1186/1471-2105-11-119.

Jolley, Keith A, and Martin CJ Maiden. 2010. "BIGSdb: Scalable Analysis of Bacterial Genome Variation at the Population Level." *BMC Bioinformatics* 11 (1): 595. doi:10.1186/1471-2105-11-595.

Kapogiannis, Bill G, Sarah Satola, Harry L Keyserling, and Monica M Farley. 2005. "Invasive Infections with *Haemophilus Influenzae* Serotype a Containing an IS*1016-bexA* Partial Deletion: Possible Association with Virulence." *Clinical Infectious Diseases* 41 (11): E97–103. doi:10.1086/498028.

Kappler, Ulrike, Rabeb Dhouib, Remya Purushothaman Nair, and Alastair G. McEwan. 2017. "Draft Genome Sequences of Three Nontypeable Strains of Haemophilus Influenzae , C188, R535, and 1200, Isolated from Different Types of Disease." *Genome Announcements* 5 (12): e00035-17. doi:10.1128/genomeA.00035-17.

Kelly, L., R. S. W. Tsang, A. Morgan, F. B. Jamieson, and M. Ulanova. 2011. "Invasive Disease Caused by Haemophilus Influenzae Type a in Northern Ontario First Nations Communities." *Journal of Medical Microbiology* 60 (3). Microbiology Society: 384–90. doi:10.1099/jmm.0.026914-0.

Keylock, C. J. 2005. "Simpson Diversity and the Shannon-Wiener Index as Special Cases of a Generalized Entropy." *Oikos* 109 (1). Munksgaard International Publishers: 203–7. doi:10.1111/j.0030-1299.2005.13735.x.

Kingry, Luke C., Lori A. Rowe, Laurel B. Respicio-Kingry, Charles B. Beard, Martin E. Schriefer, and Jeannine M. Petersen. 2016. "Whole Genome Multilocus Sequence Typing as an Epidemiologic Tool for Yersinia Pestis." *Diagnostic Microbiology and Infectious Disease* 84 (4). Elsevier B.V.: 275–80. doi:10.1016/j.diagmicrobio.2015.12.003.

Kluytmans-Van Den Bergh, Marjolein F.Q., John W.A. Rossen, Patricia C.J. Bruijning-Verhagen, Marc J.M. Bonten, Alexander W. Friedrich, Christina M.J.E. Vandenbroucke-Grauls, Rob J.L. Willems, and Jan A.J.W. Kluytmans. 2016. "Whole-Genome Multilocus Sequence Typing of Extended-Spectrum-Beta-Lactamase-Producing Enterobacteriaceae." *Journal of Clinical Microbiology* 54 (12). American Society for Microbiology: 2919–27. doi:10.1128/JCM.01648-16.

Kristensen, Martin. 1922. "Investigations into the Occurrence and Classification of the Haemoglobinophilic Bacteria." Levin & Munksgaard. http://agris.fao.org/agris-search/search.do?recordID=US201300284003.

Kroll, J. S., B. Loynds, L. N. Brophy, and E. R. Moxon. 1990. "The Bex Locus in Encapsulated *Haemophilus Influenzae*: A Chromosomal Region Involved in Capsule Polysaccharide Export." *Molecular Microbiology* 4 (11): 1853–62. doi:10.1111/j.1365-2958.1990.tb02034.x.

Kroll, J. S., B. M. Loynds, and E. R. Moxon. 1991. “The *Haemophilus Influenzae* Capsulation Gene Cluster: A Compound Transposon.” *Molecular Microbiology* 5 (6): 1549–60. doi:10.1111/j.1365-2958.1991.tb00802.x.

Kroll, J. Simon, Isobel Hopkins, and E. Richard Moxon. 1988. “Capsule Loss in H. Influenzae Type B Occurs by Recombination-Mediated Disruption of a Gene Essential for Polysaccharide Export.” *Cell* 53 (3): 347–56. doi:10.1016/0092-8674(88)90155-9.

Kroll, J. Simon, E. Richard Moxon, and Barbara M. Loynds. 1993. “An Ancestral Mutation Enhancing the Fitness and Increasing the Virulence of *Haemophilus Influenzae* Type B.” *Journal of Infectious Diseases* 168 (1): 172–76. doi:10.1093/infdis/168.1.172.

Kuo, J.S-C., V W Doelling, J F Graveline, and D W McCoy. 1985. “Evidence for Covalent Attachment of Phospholipid to the Capsular Polysaccharide of *Haemophilus Influenzae* Type B.” *J Bacteriol* 163 (2): 769–73.

LaClaire, Leslye L., Maria Lucia C Tondella, David S. Beall, Corie A. Noble, Pratima L. Raghunathan, Nancy E. Rosenstein, Tanja Popovic, et al. 2003. “Identification of Haemophilus Influenzae Serotypes by Standard Slide Agglutination Serotyping and PCR-Based Capsule Typing.” *Journal of Clinical Microbiology* 41 (1): 393–96. doi:10.1128/JCM.41.1.393-396.2003.

Ladhani, Shamez, Mary P.E. Slack, Paul T. Heath, Anne Von Gottberg, Manosree Chandra, Mary E. Ramsay, Peter McIntyre, et al. 2010. “Invasive *Haemophilus Influenzae* Disease, Europe, 1996-2006.” *Emerging Infectious Diseases* 16 (3): 455–63. doi:10.3201/eid1603.090290.

Lâm, T. T., H. Claus, M. Frosch, and U. Vogel. 2016. “Analysis of Non-Typeable *Haemophilus Influenzae* in Invasive Disease Reveals Lack of the Capsule Locus.” *Clinical Microbiology and Infection* 22 (1). Elsevier Ltd: 63.e7-63.e8. doi:10.1016/j.cmi.2015.09.027.

Lâm, Thiên Trí, Heike Claus, Matthias Frosch, and Ulrich Vogel. 2011. “Sequence Analysis of Serotype-Specific Synthesis Regions II of *Haemophilus Influenzae* Serotypes c and D: Evidence for Common Ancestry of Capsule Synthesis in Pasteurellaceae and *Neisseria Meningitidis*.” *Research in Microbiology* 162 (5): 483–87. doi:10.1016/j.resmic.2011.04.002.

Langereis, Jeroen D., and Marien I. De Jonge. 2015. “Invasive Disease Caused by Nontypeable *Haemophilus Influenzae*.” *Emerging Infectious Diseases* 21 (10): 1711–18. doi:10.3201/eid2110.150004.

Langmead, Ben, and Steven L Salzberg. 2012. “Fast Gapped-Read Alignment with Bowtie 2.” *Nature Methods* 9 (4). Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.: 357–59. doi:10.1038/nmeth.1923.

Leinonen, Rasko, Hideaki Sugawara, and Martin Shumway. 2011. “The Sequence Read Archive.” *Nucleic Acids Research* 39 (SUPPL. 1). Oxford University Press: D19-21. doi:10.1093/nar/gkq1019.

Letunic, Ivica, and Peer Bork. 2016. “Interactive Tree of Life (iTOL) v3: An Online Tool for the Display and Annotation of Phylogenetic and Other Trees.” *Nucleic Acids Research*. doi:10.1093/nar/gkw290.

Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. 2009. “The Sequence Alignment/Map Format and SAMtools.” *Bioinformatics* 25 (16): 2078–79. doi:10.1093/bioinformatics/btp352.

Lima, Josilene B. T., Guilherme S. Ribeiro, Soraia M. Cordeiro, Edilane L. Gouveia, Kátia Salgado, Brian G. Spratt, Daniel Godoy, Mitermayer G. Reis, Albert I. Ko, and Joice N. Reis. 2010. “Poor Clinical Outcome for Meningitis Caused by *Haemophilus Influenzae* Serotype A Strains Containing the IS 1016-bexA Deletion.” *The Journal of Infectious Diseases* 202 (10). Oxford University Press: 1577–84. doi:10.1086/656778.

Lin, Hsin Hung, and Yu Chieh Liao. 2015. “Evaluation and Validation of Assembling Corrected Pacbio Long Reads for Microbial Genome Completion via Hybrid Approaches.” *PLoS ONE* 10 (12). Public Library of Science: e0144305. doi:10.1371/journal.pone.0144305.

Livorsi, Daniel J., Jessica R. MacNeil, Amanda C. Cohn, Joseph Bareta, Shelly Zansky, Susan Petit, Ken Gershman, et al. 2012. “Invasive *Haemophilus Influenzae* in the United States, 1999-2008: Epidemiology and Outcomes.” *Journal of Infection* 65 (6): 496–504. doi:10.1016/j.jinf.2012.08.005.

Maiden, Martin C. J., Jane A. Bygraves, Edward Feil, Giovanna Morelli, Joanne E. Russell, Rachel Urwin, Qing Zhang, et al. 1998. “Multilocus Sequence Typing: A Portable Approach to the Identification of Clones within Populations of Pathogenic Microorganisms.” *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences. doi:10.2307/44811.

Mäkelä, P Helena, and Helena Käyhty. 2002. “Evolution of Conjugate Vaccines.” *Expert Review of Vaccines* 1 (3): 399–410. doi:10.1586/14760584.1.3.399.

Meats, Emma, Edward J. Feil, Suzanna Stringer, Alison J. Cody, Richard Goldstein, J. Simon Kroll, Tanja Popovic, and Brian G. Spratt. 2003. “Characterization of Encapsulated and Noncapsulated *Haemophilus Influenzae* and Determination of Phylogenetic Relationships by Multilocus Sequence Typing.” *Journal of Clinical Microbiology* 41 (4). American Society for Microbiology: 1623–36. doi:10.1128/JCM.41.4.1623-1636.2003.

Millar, E. V., K. L. O’Brien, J. P. Watt, J. Lingappa, R. Pallipamu, N. Rosenstein, D. Hu, R. Reid, and M. Santosham. 2005. “Epidemiology of Invasive *Haemophilus Influenzae* Type A Disease among Navajo and White Mountain Apache Children, 1988-2003.” *Clinical Infectious Diseases* 40 (6). Oxford University Press: 823–30. doi:10.1086/428047.

Moran, Anthony P., Martina M. Prendergast, and Ben J. Appelmelk. 1996. “Molecular Mimicry of Host Structures by Bacterial Lipopolysaccharides and Its Contribution to Disease.” *FEMS Immunology and Medical Microbiology* 16 (2): 105–15. doi:10.1016/S0928-8244(96)00072-7.

Morris, Denise E., David W. Cleary, and Stuart C. Clarke. 2017. “Secondary Bacterial Infections Associated with Influenza Pandemics.” *Frontiers in Microbiology* 8 (JUN): 1–17. doi:10.3389/fmicb.2017.01041.

Moura, Alexandra, Alexis Criscuolo, Hannes Pouseele, Mylène M. Maury, Alexandre Leclercq, Cheryl Tarr, Jonas T. Björkman, et al. 2016. “Whole Genome-Based Population Biology and Epidemiological Surveillance of Listeria Monocytogenes.” *Nature Microbiology* 2 (October). Nature Publishing Group: 16185. doi:10.1038/nmicrobiol.2016.185.

Murphy, Timothy F., Howard Faden, Lauren O. Bakaletz, Jennelle M. Kyd, Arne Forsgren, Jose Campos, Mumtaz Virji, and Stephen I. Pelton. 2009. “Nontypeable Haemophilus Influenzae as a Pathogen in Children.” *The Pediatric Infectious Disease Journal* 28 (1): 43–48. doi:10.1097/INF.0b013e318184dba2.

Murphy, Timothy F, Aimee L Brauer, Sanjay Sethi, Mogens Kilian, Xueya Cai, Alan J Lesse, Source The, et al. 2007. “Haemophilus Haemolyticus : A Human Respiratory Tract Commensal to Be Distinguished from Haemophilus Influenzae Linked References Are Available on JSTOR for This Article : Haemophilus Haemolyticus : A Human Respiratory Tract Commensal to Be Distinguished F” 195 (1): 81–89.

Musser, J. M., J. S. Kroll, E. R. Moxon, and R. K. Selander. 1988. “Clonal Population Structure of Encapsulated Haemophilus Influenzae.” *Infection and Immunity* 56 (8). American Society for Microbiology: 1837–45. http://www.ncbi.nlm.nih.gov/pubmed/2899551.

Myers, Angela L., Mary Anne Jackson, Lixin Zhang, Douglas S. Swanson, and Janet R. Gilsdorf. 2017. “*Haemophilus Influenzae* Type B Invasive Disease in Amish Children, Missouri, USA, 2014.” *Emerging Infectious Diseases* 23 (1): 112–14. doi:10.3201/eid2301.160593.

Nascimento, Marta, Adriano Sousa, Mário Ramirez, Alexandre P. Francisco, João A. Carriço, and Cátia Vaz. 2017. “PHYLOViZ 2.0: Providing Scalable Data Integration and Visualization for Multiple Phylogenetic Inference Methods.” *Bioinformatics* 33 (1). Oxford University Press: 128–29. doi:10.1093/bioinformatics/btw582.

Nørskov-Lauritsen, Niels. 2009. “Detection of Cryptic Genospecies Misidentified as Haemophilus Influenzae in Routine Clinical Samples by Assessment of Marker Genes fucK, Hap, and sodC.” *Journal of Clinical Microbiology* 47 (8): 2590–92. doi:10.1128/JCM.00013-09.

Page, Andrew J., Carla A. Cummins, Martin Hunt, Vanessa K. Wong, Sandra Reuter, Matthew T.G. Holden, Maria Fookes, Daniel Falush, Jacqueline A. Keane, and Julian Parkhill. 2015. “Roary: Rapid Large-Scale Prokaryote Pan Genome Analysis.” *Bioinformatics* 31 (22): 3691–93. doi:10.1093/bioinformatics/btv421.

Paradis, E., J. Claude, and K. Strimmer. 2004. “APE: Analyses of Phylogenetics and Evolution in R Language.” *Bioinformatics* 20 (2). Oxford University Press: 289–90. doi:10.1093/bioinformatics/btg412.

Parisi, Dana N, and Luis R Martinez. 2014. “Intracellular Haemophilus Influenzae Invades the Brain: Is Zyxin a Critical Blood Brain Barrier Component Regulated by TNF-A?” *Virulence* 5 (6): 645–47. doi:10.4161/viru.36086.

Park, S-J, M Saito-Adachi, Y Komiyama, and K Nakai. 2016. “Advances, Practice, and Clinical Perspectives in High-Throughput Sequencing.” *Oral Diseases* 22 (5): 353–64. doi:10.1111/odi.12403.

Peltola, H, H Kayhty, A Sivonen, and H Makela. 1977. “Haemophilus Influenzae Type B Capsular Polysaccharide Vaccine in Children: A Double-Blind Field Study of 100,000 Vaccinees 3 Months to 5 Years of Age in Finland.” *Pediatrics* 60 (5): 730–37.

Perdue, D G, L R Bulkow, B G Gellin, M Davidson, K M Petersen, R J Singleton, and A J Parkinson. 2000. “Invasive Haemophilus Influenzae Disease in Alaskan Residents Aged 10 Years and Older before and after Infant Vaccination Programs.” *JAMA, Journal of the American Medical Association* 283 (23): 3089–94. doi:10.1001/jama.283.23.3089.

Petkau, Aaron, Philip Mabon, Cameron Sieffert, Natalie C. Knox, Jennifer Cabral, Mariam Iskander, Mark Iskander, et al. 2017. “SNVPhyl: A Single Nucleotide Variant Phylogenomics Pipeline for Microbial Genomic Epidemiology.” *Microbial Genomics* 3 (6). doi:10.1099/mgen.0.000116.

Pfeiffer, R. 1892. “I.-Preliminary Communication on the Exciting Causes of Influenza.” *British Medical Journal* 1 (1620): 128. doi:10.1136/bmj.1.1620.128.

Pinto, Francisco R., José Melo-Cristino, and Mário Ramirez. 2008. “A Confidence Interval for the Wallace Coefficient of Concordance and Its Application to Microbial Typing Methods.” Edited by Enrico Scalas. *PLoS ONE* 3 (11). Public Library of Science: e3696. doi:10.1371/journal.pone.0003696.

Pittman, M. 1931. “Variation and Type Specificity in the Bacterial Species Hemophilus Influenzae.” *Journal of Experimental Medicine* 53 (4). Rockefeller University Press: 471–92. doi:10.1084/jem.53.4.471.

Price, Erin P, Derek S Sarovich, Elizabeth Nosworthy, Jemima Beissbarth, Robyn L Marsh, Janessa Pickering, Lea-Ann S Kirkham, Anthony D Keil, Anne B Chang, and Heidi C Smith-Vaughan. 2015. “Haemophilus Influenzae: Using Comparative Genomics to Accurately Identify a Highly Recombinogenic Human Pathogen.” *BMC Genomics* 16. BMC Genomics: 641. doi:10.1186/s12864-015-1857-x.

Rand, William M. 1971. “Objective Criteria for the Evaluation of Clustering Methods.” *Journal of the American Statistical Association* 66 (336): 846–50. doi:10.1080/01621459.1971.10482356.

Resman, F., M. Ristovski, J. Ahl, A. Forsgren, J. R. Gilsdorf, A. Jasir, B. Kaijser, G. Kronvall, and K. Riesbeck. 2011. “Invasive Disease Caused by *Haemophilus Influenzae* in Sweden 1997-2009; Evidence of Increasing Incidence and Clinical Burden of Non-Type B Strains.” *Clinical Microbiology and Infection* 17 (11): 1638–45. doi:10.1111/j.1469-0691.2010.03417.x.

Rhoads, Anthony, and Kin Fai Au. 2015. “PacBio Sequencing and Its Applications.” *Genomics, Proteomics and Bioinformatics* 13 (5): 278–89. doi:10.1016/j.gpb.2015.08.002.

Ribeiro-Gonçalves, Bruno, Alexandre P Francisco, Cátia Vaz, Mário Ramirez, and João André Carriço. 2016. “PHYLOViZ Online: Web-Based Tool for Visualization, Phylogenetic Inference, Analysis and Sharing of Minimum Spanning Trees.” *Nucleic Acids Research* 44 (W1). Oxford University Press: W246-51. doi:10.1093/nar/gkw359.

Roberson, Emily B., and Mary K. Firestone. 1992. “Relationship between Desiccation and Exopolysaccharide Production in a Soil *Pseudomonas* Sp.” *Applied and Environmental Microbiology* 58 (4). American Society for Microbiology: 1284–91. http://www.ncbi.nlm.nih.gov/pubmed/16348695.

Roberts, Ian S. 1996. “The Biochemistry and Genetics of Capsular Polysaccharide Production in Bacteria.” *Annual Review of Microbiology* 50 (1). Annual Reviews 4139 El Camino Way, P.O. Box 10139, Palo Alto, CA 94303-0139, USA: 285–315. doi:10.1146/annurev.micro.50.1.285.

Roux, David M. le, and Heather J. Zar. 2017. “Community-Acquired Pneumonia in Children — a Changing Spectrum of Disease.” *Pediatric Radiology* 47 (11). Pediatric Radiology: 1392–98. doi:10.1007/s00247-017-3827-8.

Santona, Antonella, Elisa Taviani, Massimo Deligios, Ahmed A. Al-Qahtani, Mohammed N. Al-Ahdal, Salvatore Rubino, and Bianca Paglietti. 2016. “Vancomycin-Resistant Enterococcus Faecium High-Resolution Typing by Core Genome Multilocus Sequence Typing.” *Journal of Infection in Developing Countries* 10 (10): 1159–61. doi:10.3855/jidc.9223.

Satola, Sarah W., Julie T. Collins, Ruth Napier, and Monica M. Farley. 2007. “Capsule Gene Analysis of Invasive *Haemophilus Influenzae*: Accuracy of Serotyping and Prevalence of IS1016 among Nontypeable Isolates.” *Journal of Clinical Microbiology* 45 (10): 3230–38. doi:10.1128/JCM.00794-07.

Schumacher, S K, C D Marchant, A M Loughlin, V Bouchet, A Stevenson, and S I Pelton. 2012. “Prevalence and Genetic Diversity of Nontypeable Haemophilus Influenzae in the Respiratory Tract of Infants and Primary Caregivers.” *Pediatric Infectious Disease* 31 (2): 145–49. doi:10.1002/ana.22528.Toll-like.

Severiano, Ana, João A. Carriço, D. Ashley Robinson, Mário Ramirez, and Francisco R. Pinto. 2011. “Evaluation of Jackknife and Bootstrap for Defining Confidence Intervals for Pairwise Agreement Measures.” Edited by Fabio Rapallo. *PLoS ONE* 6 (5). Public Library of Science: e19539. doi:10.1371/journal.pone.0019539.

Severiano, Ana, Francisco R. Pinto, Mário Ramirez, and João A. Carriço. 2011. “Adjusted Wallace Coefficient as a Measure of Congruence between Typing Methods.” *Journal of Clinical Microbiology* 49 (11). American Society for Microbiology: 3997–4000. doi:10.1128/JCM.00624-11.

Shively, R. G., J. T. Shigei, E. M. Peterson, and L. M. De La Maza. 1981. “Typing of Haemophilus Influenzae by Coagglutination and Conventional Slide Agglutination.” *Journal of Clinical Microbiology* 14 (6): 706–8.

Shuel, Michelle, Linda Hoang, Dennis K.S. Law, and Raymond Tsang. 2011. “Invasive *Haemophilus Influenzae* in British Columbia: Non-Hib and Non-Typeable Strains Causing Disease in Children and Adults.” *International Journal of Infectious Diseases* 15 (3). International Society for Infectious Diseases: e167–73. doi:10.1016/j.ijid.2010.10.005.

Sill, Michelle L., Dennis K.S. Law, Jianwei Zhou, Stuart Skinner, John Wylie, and Raymond S.W. Tsang. 2007. “Population Genetics and Antibiotic Susceptibility of Invasive *Haemophilus Influenzae* in Manitoba, Canada, from 2000 to 2006.” *FEMS Immunology and Medical Microbiology* 51 (2): 270–76. doi:10.1111/j.1574-695X.2007.00299.x.

Silva, Mickael, Miguel P. Machado, Mirko Rossi, Jacob Moran-Gilad, Sergio Santos, Mario Ramirez, and Joao Andre Carrico. 2017. “chewBBACA: A Complete Suite for Gene-by-Gene Schema Creation and Strain Identification.” *bioRxiv*, August. Cold Spring Harbor Laboratory, 173146. doi:10.1101/173146.

Spinola, S M, J Peacock, F W Denny, D L Smith, and J G Cannon. 1986. “Epidemiology of Colonization by Nontypable Haemophilus Influenzae in Children: A Longitudinal Study.” *The Journal of Infectious Diseases* 154 (1): 100–109. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=3486923.

Sriram, Krishna Bajee, Amanda J. Cox, Robert L. Clancy, Mary P. E. Slack, and Allan W. Cripps. 2017. “Nontypeable *Haemophilus Influenzae* and Chronic Obstructive Pulmonary Disease: A Review for Clinicians.” *Critical Reviews in Microbiology* 7828 (October): 1–18. doi:10.1080/1040841X.2017.1329274.

Staples, M., R. M. A. Graham, and A. V. Jennison. 2017. “Characterisation of Invasive Clinical *Haemophilus Influenzae* Isolates in Queensland, Australia Using Whole-Genome Sequencing.” *Epidemiology and Infection* 145 (8): 1727–36. doi:10.1017/S0950268817000450.

Sukupolvi-Petty, Soila, Susan Grass, and Joseph W. St. Geme. 2006. “The *Haemophilus Influenzae* Type B hcsA and hcsB Gene Products Facilitate Transport of Capsular Polysaccharide across the Outer Membrane and Are Essential for Virulence.” *Journal of Bacteriology* 188 (11): 3870–77. doi:10.1128/JB.01968-05.

Sullivan, Christopher B., Matthew A. Diggle, and Stuart C. Clarke. 2005. “Multilocus Sequence Typing: Data Analysis in Clinical Microbiology and Public Health.” *Molecular Biotechnology* 29 (3): 245–54. doi:10.1385/MB:29:3:245.

Swords, W Edward, Paul a Jones, and Michael a Apicella. 2003. “The Lipo-Oligosaccharides of Haemophilus Influenzae: An Interesting Array of Characters.” *Journal of Endotoxin Research* 9 (3): 131–44. doi:10.1179/096805103125001531.

Timens, W I M, Adriana Boes, Thea Rozeboom-uiterwijk, and Sibrand Poppema. 1989. “Immaturity of the Human Splenic Marginal Zone in Infancy. Possible Contribution to the Deficient Infant Immune Response.” *The Journal of Immunology* 143 (10): 3200–3206. http://www.ncbi.nlm.nih.gov/pubmed/12901578.

Tsang, Raymond S.W., Y. Anita Li, Angie Mullen, Maureen Baikie, Kathleen Whyte, Michelle Shuel, Gregory Tyrrell, Jenny A.L. Rotondo, Shalini Desai, and John Spika. 2016. “Laboratory Characterization of Invasive *Haemophilus Influenzae* Isolates from Nunavut, Canada, 2000-2012.” *International Journal of Circumpolar Health* 75 (0): 29798. doi:10.3402/ijch.v75.29798.

Tsang, Raymond S.W., Jean Francois Proulx, Kristy Hayden, Michelle Shuel, Brigitte Lefebvre, Andree Anne Boisvert, and Dorothy Moore. 2017. “Characteristics of Invasive Haemophilus Influenzae Serotype a (Hia) from Nunavik, Canada and Comparison with Hia Strains in Other North American Arctic Regions.” *International Journal of Infectious Diseases* 57. International Society for Infectious Diseases: 104–7. doi:10.1016/j.ijid.2017.02.003.

Ulanova, Marina, and Tsang, Raymond S.W. 2009. “Invasive*Haemophilus Influenzae* Disease: Changing Epidemiology and Host-Parasite Interactions in the 21st Century.” *Infection, Genetics and Evolution* 9 (4): 594–605. doi:10.1016/j.meegid.2009.03.001.

Ulanova, Marina, and Tsang, Raymond S.W. 2014. “Haemophilus Influenzae Serotype a as a Cause of Serious Invasive Infections.” *The Lancet Infectious Diseases* 14 (1): 70–82. doi:10.1016/S1473-3099(13)70170-1.

Wallace, David L. 1983. “A Method for Comparing Two Hierarchical Clusterings: Comment.” *Journal of the American Statistical Association*. doi:10.1080/01621459.1983.10478009.

Wang, Huabin, Jonathan J. Wilksch, Richard A. Strugnell, and Michelle L. Gee. 2015. “Role of Capsular Polysaccharides in Biofilm Formation: An AFM Nanomechanics Study.” *ACS Applied Materials and Interfaces* 7 (23): 13007–13. doi:10.1021/acsami.5b03041.

Wattam, Alice R., James J. Davis, Rida Assaf, Sébastien Boisvert, Thomas Brettin, Christopher Bun, Neal Conrad, et al. 2017. “Improvements to PATRIC, the All-Bacterial Bioinformatics Database and Analysis Resource Center.” *Nucleic Acids Research* 45 (D1): D535–42. doi:10.1093/nar/gkw1017.

Whittaker, R H. 1972. “Evolution and Measurement of Species Diversity.” *Taxon* 21 (2/3). IAPT: 213. doi:10.2307/1218190.

Whittaker, Robert, Assimoula Economopoulou, Joana Gomes Dias, Elizabeth Bancroft, Miriam Ramliden, and Lucia Pastore Celentano. 2017. “Epidemiology of Invasive *Haemophilus Influenzae* Disease, Europe, 2007–2014.” *Emerging Infectious Diseases* 23 (3): 396–404. doi:10.3201/eid2303.161552.

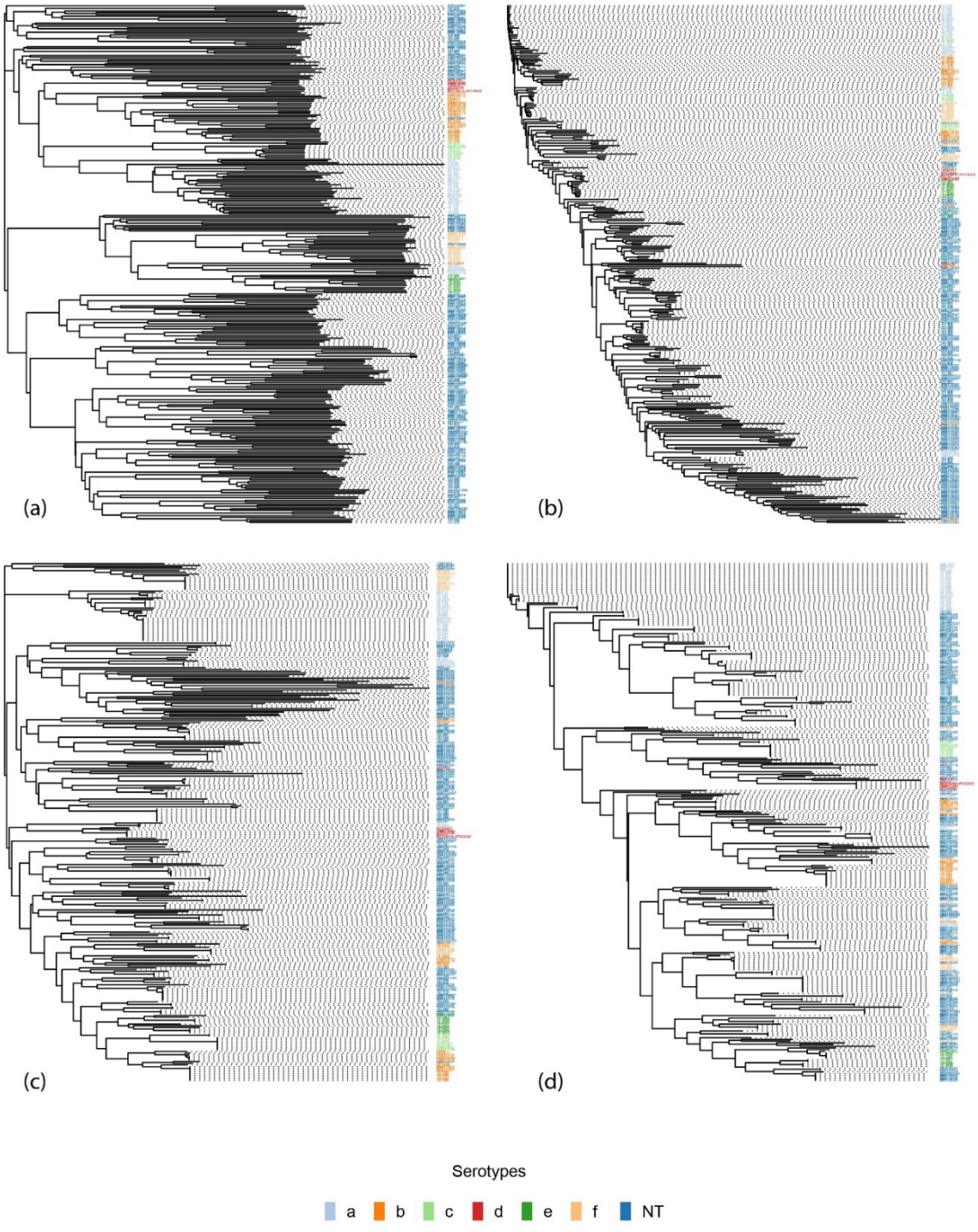
Wilson, R. 1991. “The Role of *Haemophilus Influenzae* in the Pathogenesis of Pneumonia.” *Reviews of Infectious Diseases* 13 (Supplement 6): S518–27. doi:10.1093/clinids/13.Supplement_6.S518.

Yoshida, Catherine E., Peter Kruczkiewicz, Chad R. Laing, Erika J. Lingohr, Victor P.J. Gannon, John H.E. Nash, and Eduardo N. Taboada. 2016. “The Salmonella in Silico Typing Resource (SISTR): An Open Web-Accessible Tool for Rapidly Typing and Subtyping Draft Salmonella Genome Assemblies.” Edited by Michael Hensel. *PLoS ONE* 11 (1). Public Library of Science: e0147101. doi:10.1371/journal.pone.0147101.

Zhou, Haijian, Wenbing Liu, Tian Qin, Chen Liu, and Hongyu Ren. 2017. “Defining and Evaluating a Core Genome Multilocus Sequence Typing Scheme for Whole-Genome Sequence-Based Typing of *Klebsiella Pneumoniae*.” *Frontiers in Microbiology* 8 (March). Frontiers: 371. doi:10.3389/fmicb.2017.00371.

Zwahlen, André, J. Simon Kroll, Lorry G. Rubin, and E. Richard Moxon. 1989. “The Molecular Basis of Pathogenicity in *Haemophilus Influenzae*: Comparative Virulence of Genetically-Related Capsular Transformants and Correlation with Changes at the Capsulation Locus Cap.” *Microbial Pathogenesis* 7 (3): 225–35. doi:10.1016/0882-4010(89)90058-2.

Appendix



Supplementary Figure 1 Neighbour joining trees based on the a) SNV data, b) cgMLST schema, c) rMLST schema, and d) 7-gene MLST schema. The branch tips were colorized using serotyping information