# Nontypeable *Haemophilus influenzae* High Molecular Weight Adhesins: Molecular Epidemiology, Evolution, and Within-Host Population Dynamics

by

Gregory S. Davis

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Epidemiological Science)
in the University of Michigan
2013

Doctoral Committee:

Professor Janet Gilsdorf, Co-Chair
Professor Denise Kirschner, Co-Chair
Assistant Professor Suzanne Dawid
Associate Professor Carl F. Marrs
Professor Harry Mobley
Research Assistant Professor Lixin Zhang

**Acknowledgements**

I would like to thank my co-mentors, Janet Gilsdorf and Denise Kirschner, and dissertation committee members Suzanne Dawid, Harry Mobley, Carl Marrs, and Lixin Zhang for their support, guidance, insight, and lively discussions.

I was fortunate enough to call three labs home, the Gilsdorf lab, the Kirschner lab, and the Mobley lab. I would like to thank everyone in these labs, past and present, for their support. I would especially like to thank May for keeping things running smoothly and peacefully in the Gilsdorf laboratory, and always being there to lend a helping hand. Joe Waliga and Simeone Marino kept me updated, connected and sane, and for that I am very appreciative. I would also like to thank Stef Himpsl, Erika Flannery, Ariel Brumbaugh, and Chris Alteri for thoughtful discussions, technical advice, and necessary distractions.

I would further like to thank the Department of Epidemiology and my fellow students in the School of Public Health who made my time in Michigan most enjoyable. I was fortunate to have a great group of friends to share my PhD experience with, including Justin Henderson, Brian Davis, Ashley Hazel, Meghan Milbrath, and Mariana Rosenthal. KJ Hoggatt, Felice Le, Sarah Lyon-Callo, and Sarah Reeves were instrumental in my understanding of epidemiologic theory. And, none of this would have been possible without help from Dawn Reed, Sally Bauzin, Jody Gray, Jess Whipple, and Nancy Francis; thanks for making things happen.

Finally, I would like to thank my wife, Ines, and my family, Bill, Barbara, and Heather, for their support, encouragement, understanding, and inspiration.

## Table of Contents

## List of Figures

## List of Tables

## Abstract


**Nontypeable *Haemophilus influenzae* High Molecular Weight Adhesins: Molecular Epidemiology, Evolution, and Within-Host Population Dynamics**

by

Gregory S. Davis

Nontypeable *Haemophilus influenzae* (NTHi), a Gram-negative bacterium that commonly resides within the human pharynx as a commensal, is also capable of causing localized infections of the respiratory tract and invasive disease. Among children, NTHi is a leading cause of acute otitis media (AOM) which is a significant cause of childhood morbidity and the most common reason for prescribing antibiotics in the US. In adults, NTHi is often associated with acute exacerbations of chronic obstructive disease (COPD), the fourth leading cause of death worldwide. Because NTHi is human restricted its long term survival is dependent upon its ability to successfully colonize new hosts. Adherence to host epithelium, mediated by bacterial adhesins, is proposed to be one of the first steps in bacterial colonization and since disease-causing NTHi strains originate from strains that colonize the pharynges, adherence also marks one of the first steps in NTHi pathogenesis. NTHi encode several adhesins, including the high molecular weight (HMW) adhesins that mediate attachment to the respiratory epithelium where they interact with the host immune system, eliciting a strong humoral response. *hmwA*, which encodes the HMW adhesin, displays marked amino acid diversity and also demonstrated phase variation mediated by 7-base pair tandem repeats located within the *hmwA* promoter region. Thus, to gain further insight into the role of HMW adhesins during colonization and disease I: (1) described a *bexB*-based molecular typing method that differentiates typeable from

true nontypeable *H. influenzae*, (2) sequenced the *hmwA* core binding domains from a collection of 170 geographically distributed commensal and AOM isolates, demonstrating that they form four distinct phylogenetic clusters, (3) found evidence of positive selection operating on the *hmwA* region that encodes the core binding domain, and (4) demonstrated, using a mathematical model, that the occurrence of large, yet rare, phase variable deletion events allows for the stable maintenance of a small population of adherent NTHi cells during colonization in spite of HMW-specific antibody mediated immunity. This work has direct implications for ongoing efforts aimed at developing an effective NTHi vaccine and, at a more basic level, advances our understanding of host-pathogen, and importantly, host-commensal interactions.

# Chapter 1
# Introduction

**Public health significance of acute otitis media and chronic obstructive pulmonary diseases.** Acute otitis media (AOM), which is an infection of the middle ear, is the most common childhood bacterial disease for which antibiotics are prescribed (30). During 2005 there were an estimated 709 million cases of AOM globally, reflecting an annual incidence rate of approximately 11%, with the majority of those cases occurring in children under five years of age (61). In the United States, approximately 83% of children have had at least one episode of AOM by the age of three years and 45% have suffered at least three AOM episodes (96). Recurrent episodes of AOM are associated with long term sequelae including hearing loss, which in turn can lead to impaired speech development and impaired cognitive ability (3, 66). In addition, otitis media with effusion resulting from AOM is the leading reason for performing tympanostomy-tube insertions, the most common childhood operation (1, 59). AOM imposes a significant burden on the health care system. In 2009, the total annual cost of AOM in the US was estimated to be $4.2 billion (69). Thus, reducing the incidence of AOM, ideally through preventive measures, will reduce a significant burden on our healthcare system.

Both viral and bacterial pathogens are associated with AOM and often an episode of bacterial AOM is preceded by a viral upper respiratory tract infection (73). Three predominant bacterial species isolated from the middle ear spaces of children with AOM are *Haemophilus influenzae* (Hi), *Streptococcus pneumoniae*, and *Moraxella catarrhalis*. The most recent estimates suggest that the percentage of AOM cases attributable to *S. pneumoniae* ranges from 42 – 48%, nontypeable *H. influenzae* 36 – 45%, and *M. catarrhalis* 13 – 16% (17).

Interestingly, these same three bacterial pathogens are associated with another human disease, namely chronic obstructive pulmonary disease (COPD) (46, 82). COPD is a chronic, progressive lung disease resulting from interactions between environmental factors (*e.g.*, smoking), host immunity, and microbial factors associated with the microbial communities that

colonize the respiratory tract. COPD prevalence increases with age, reaching approximately 12% among US adults aged 65 – 74 (77). In the US alone, the total economic cost of COPD/asthma for 2008 was estimated to be $68 billion (43) and chronic lower respiratory tract infections, which include COPD, were the third leading cause of death (77). On a global scale, the WHO estimates that during 2008 COPD was responsible for 3.28 million deaths worldwide, making it the fourth leading cause of death (70, accessed September 19, 2012). Bacteria commonly isolated from the sputum of COPD patients include *Streptococcus pneumoniae*, *Moraxella catarrhalis*, and *Haemophilus influenzae*, the same three bacterial species most often associated the AOM in children (46, 82). Importantly, nontypeable *H. influenzae* are commonly associated with acute exacerbations of COPD and these events are often treated with antibiotics (46, 56).

Given the significant contribution of nontypeable *H. influenzae* (NTHi) to both AOM and COPD, decreasing the incidence of NTHi-associated diseases can reduce a significant burden on our healthcare system and has the potential to decrease antibiotic usage and the associated concerns regarding emerging antibiotic resistance. To these ends, current efforts to lower NTHi disease incidence, especially AOM, are primarily focused on developing preventive measures such as vaccines.

***Haemophilus influenzae.*** *Haemophilus influenzae* are small, nonmotile, Gram-negative coccobacilli that reside exclusively within humans, typically within the nasopharynges. *H. influenzae* strains can be designated as either typeable or nontypeable based upon their reactivity with typing antisera developed against each of six immunologically distinct polysaccharide capsules. Because of the capsule, typeable and nontypeable *H. influenzae* are fundamentally different, but differentiating between typeable and nontypeable strains is challenging. Serum agglutination, using type specific serum directed against each of the six immunologically distinct *H. influenzae* capsules, has classically been used to test for the presence of a capsule and to determine capsular type. By these criteria, strains that fail to react with typing sera are considered “nontypeable”, but strains can fail to react with typing sera for several different reasons. These reasons include false negative typing reactions, failure to produce functional capsule due to mutations in the *bexA* gene that encodes a protein required for capsule export, and the complete lack of the capsule genetic locus due to long past evolutionary events. Only in the latter case, that is, the strain in question lacks the capsule locus as a consequence of its

evolutionary past, would the strain be considered a true nontypeable *H. influenzae*. Thus, "nontypeable" strains defined only by lack of reactivity with typing sera are a heterogeneous collection of strains and this can confound epidemiologic studies.

More recently, partially based on evidence for poor reliability of serum based methods, typing schemes have been transitioning from phenotypic based typing to the use of genetic typing. Genetic typing methods rely on polymerase chain reaction (PCR) based detection of *cap*-locus genes that encode the proteins necessary for capsule synthesis and expression. The *cap*-locus consists of three distinct regions (I to III). Region I and III genes flank the *cap* locus, are highly conserved across all capsular types, and encode proteins responsible for transporting the capsule across the outer membrane (49-51, 79, 93). Region II genes encode proteins specific to each capsular type, a to f, and, thus vary by capsular type. The most commonly employed *H. influenzae* genetic typing scheme involves PCR detection of the conserved region I gene *bexA* and one of the six capsule-specific region II genes (28, 58). This approach is cumbersome, requiring seven PCR reactions to identify a true NTHi strain (*i.e.*, a strain that lacks the entire capsule locus). Thus, re-examination of the current *H. influenzae* typing scheme is warranted.

Typeable *H. influenzae* tend to be relatively clonal, with lineages demarcated by capsular type, whereas extensive genetic diversity is a hallmark of nontypeable *H. influenzae* (NTHi) (14, 29, 53, 64, 98). Encapsulated strains are commonly associated with invasive human infections such as meningitis and septicemia among non-immune children. In contrast, NTHi are generally associated with localized infections of the respiratory tract such as pneumonia, sinusitis, and acute otitis media (AOM).

NTHi is a common inhabitant of the human pharynx with colonization prevalence among healthy children between 25 – 84% (15, 27, 42). NTHi colonization occurs early during childhood and is a dynamic process marked by simultaneous colonization with multiple strains and presumably high rates of strain turnover (27, 39, 62, 91, 97, 101). Recent studies suggest that two-year old children are at a significantly greater risk of being colonized with *H. influenzae* if their mother is also colonized, but interestingly, the strains that colonize mother-child pairs are genetically distinct (54). This suggests that host genetic factors may play a role in susceptibility to *H. influenzae* colonization even though NTHi does not appear to be vertically transmitted; instead, strains may be acquired from close contacts, for example, siblings or playmates.

AOM-associated pathogenic NTHi arise from the community of colonizing strains that reside within the pharynx (Figure 1.1). While not fully elaborated, the pathogenic pathway for an AOM strain involves person to person transmission, colonization of the pharynx, ascension of the Eustachian tube, establishment in the middle ear space, and finally, induction of an inflammatory response. One of the first steps in NTHi AOM pathogenesis, therefore, is adherence of bacterial cells to cells of the host respiratory tract.

**NTHi adhesins.** NTHi adherence to the respiratory epithelium is mediated by a number of surface exposed pilin and non-pilin adhesins (85-88, 95). The pilin adhesins include hemagglutinating pili (33, 86, 92, 99), *H. influenzae* surface fibril, Hsf (85, 86), P5-fimbriae (6, 63, 78, 83), and type IV pili (4). There are also numerous non-pilin adhesins including the *Haemophilus* adherence and penetration protein, Hap, and lipooligosaccharide (LOS) both of which mediate attachment to and invasion of epithelial cells (87, 95). The cell envelope opacity-associated protein A, OpaA, is also involved in epithelial cell adherence (76). Finally, nearly all NTHi cells encode either a homolog of the Hib Hsf adhesin, Hia, or a pair of high molecular weight adhesins, HMW1 and HMW2 (11, 13, 24, 90).

**The HMW adhesins.** The HMW adhesins were originally identified by radioimmunoprecipitation assay and are the predominant NTHi non-pilin adhesins (8, 10, 13, 88). Across a wide diversity of NTHi strain types, including commensal, invasive, AOM-, and COPD-associated strains, the prevalence of either *hmwA* genes or functional HMW ranges from approximately 36 to 75% (11, 24-26, 90, 100). HMW is more prevalent among disease isolates than colonizing strains (11, 24, 90), suggesting that they may contribute to AOM pathogenesis.

HMW adhesins are encoded by the *hmw* loci, each of which contains three genes, *hmwA, hmwB,* and *hmwC* (Fig. 1.2) (12). Each *hmw*-positive strain possesses two complete copies of the *hmw* locus, designated *hmw1* and *hmw2*, located in conserved but unlinked locations within the NTHi chromosome that are likely the results of a gene duplication event that occurred early during the evolution of NTHi (11, 16). The functional HMWs are encoded by *hmw1A* and *hmw2A*, which in strain 12 are 71% identical (80% similarity) at the amino acid level (11). HMWB, encoded by *hmwB*, is an outer membrane protein responsible for transporting HMW to the bacterial cell surface (20, 38, 89). HMWC, encoded by *hmwC*, is a cytoplasmic glycosylase

required for HMW maturation (36, 37, 40). In contrast to HMWA, which displays marked amino acid diversity, NTHi strain 12 HMW1B/2B and HMW1C/2C are highly conserved at the amino acid level, displaying ≈ 99% and 97% identity, respectively (11).

NTHi strain 12 *hmw1A* and *hmw2A* have been extensively characterized. HMW proteins are translated as preproteins that can be divided into three distinct functional regions: (1) the signal sequence region, (2) the core-binding domain, and (3) a conserved carboxy-terminus (20). The first 441 amino acids of HMW encode a highly conserved signal sequence that is required for Sec-dependent transport across the cytoplasmic membrane and trafficking through the periplasm (11, 20). In its functional form, the glycosylated HMW adhesin is displayed on the bacterial cell surface and remains tethered to the bacterial cell via non-covalent interactions between HMWB and the carboxy terminus of HMWA (5, 38, 89).

**HMW adhesin diversity.** The HMW binding domain, located in the amino terminus of the mature protein, is responsible for binding to the host epithelium (20). Amino acid diversity within the binding domain is quite high both within and between strains; for example, strain 12 HMW1 and HMW2 binding domains share only ≈50% amino acid identity (11, 16, 20, 23, 34). Furthermore, the majority of the glycosylated asparagine residues within strain 12 HMW1 are also localized to the binding domain region (40). HMW adhesin amino acid diversity, and possibly glycosylation, affects both NTHi tissue tropism and antigenic diversity.

HMW1 and HMW2 adhesins bestow differing *in vitro* adherence properties to NTHi. Strain 12 HMW1 mediates high level adherence to Chang cells, Hep-2, HaCaT, and NCI-H292 cells, interacting specifically with α-2,3 N-linked sialic acids, whereas strain 12 HMW2 primarily mediates adherence to HaCaT and NCI-H292 cells and the exact nature of its interaction with these cells remains elusive (88, 89). Despite the relatively high amino acid diversity among HMW adhesin core binding domains most strains demonstrate an adherence profile similar to that of strain 12 with one HMWA conferring HMW1-like adherence and the other HMW2-like adherence (16). Phylogenetic relationships among the core binding domains reveal two major sequence clusters that correspond with *in vitro* adherence characteristics, that is, sequences that encode the HMW1-like adhesins form one cluster and those encoding the HMW2-like adhesins form the second cluster (16, 34). Thus, while some of the amino acid

diversity observed among the HMW-family of adhesins defines their tissue tropism, there is also selective pressure to conserve functionality.

Pharyngeal NTHi colonization stimulates the host's adaptive immune response resulting in serum IgG, IgM, and IgA (9, 10, 35, 47, 74). Many of these antibodies target outer membrane proteins, including HMW (10, 35, 41, 67); in fact, HMW proteins are the immunodominant NTHi outer membrane proteins (10). Numerous studies have demonstrated a role for immune driven positive selection, which can generate antigenic diversity in surface exposed proteins (31, 60, 68, 84). Thus, while yet to be formally tested, it is likely that antibody mediated immunity applies positive selective pressure upon the HMW adhesins favoring antigenic diversity via the process of diversifying selection. Empirical evidence for HMW antigenic diversity is demonstrated by the varying levels of cross reactivity of anti-HMW antibodies against the HMW adhesins from heterologous strains (100). Antigenic variation resulting from amino acid substitutions or recombination events is a relatively slow process that begets species level diversity and operates between hosts. With regard to the HMW adhesins, species level diversity can be maintained because allele-specific host immunity imposes a fitness cost on those strains expressing the most prevalent HMW alleles whereas rare HMW variants experience a fitness advantage.

**HMW phase variation.** Phase variation involves heritable, reversible, and stochastic phenotypic changes mediated through various different mechanisms such as invertible elements, gene conversion, or changes in the numbers of simple sequence repeats. Simple sequence repeats (SSR) consist of multiple tandem copies of short, for example, one to nine nucleotides (75), repetitive DNA sequences. One hallmark of phase variation is that it generates phenotypic changes at rates much greater than that of the average background mutation rate. For example, applying Drake's Rule (22, 94), the overall *H. influenzae* mutation rate, assuming a genome size of 1.8 Mb, is approximately $1.14 \times 10^{-9}$ mutations per nucleotide site per generation compared to mutation rates of $2.03 \times 10^{-4}$ mutations per SSR per generation for *H. influenzae* tetranucleotide SSRs (21). Thus, background mutation rates generate less than a single mutation, on average, per NTHi generation (assuming a genome size of $1.8 \times 10^{-6}$). Through the gain or loss of repeat units by slipped stranded mispairing during DNA synthesis (*e.g.*, during DNA replication), phase variation can affect gene transcription or translation, depending on the location of the repeat

tract. SSR mediated phase variation is a common theme among bacterial pathogens, such as *Campylobacter jejuni* (72), *Neisseria meningitidis* (80), and *Helicobacter pylori* (2) all of which contain numerous phase variable loci. The net result of phase variation is the generation of population level phenotypic diversity that accumulates in the absence of any external stimuli. This diversity increases the probability that at least some members of a population are able to survive rapid environmental changes, such as those incurred during transmission between hosts. In contrast to the relatively slow pace of antigenic variation as outlined above, phase variation occrus at relatively high rates and affects the bacterial population within an individual host.

*H. influenzae* possesses numerous SSRs. Tetranucleotide repeats are the most numerous, and best characterized, type of Hi SSRs (32, 44). A 2009 study of the *H. influenzae* pangenome, based on only 16 *Hi* strains, identified over 750 SSRs of which 199 were tetranucleotide SSRs, but, only a fraction of the SSRs were potentially phase variable (75). Analysis of four complete *H. influenzae* genomes determined that 14/18 tetranucleotide SSRs mediated on-off phase variable protein expression of Hi virulence associated genes (44, 75). In contrast, 3/5 heptanucleotide SSRs were phase variable and only one of the SSRs was located within an open reading frame (75).

The HMW adhesin genes are phase variable. Both *hmwA* alleles possess 7-base pair (bp) SSRs located within their promoter regions (Figure 1.2) (19). Repeats are gained and lost stochastically during DNA synthesis by Rec-independent slipped stranded mispairing (19). Because the SSRs are located within the promoter region, changes in repeat number affect *hmwA* transcription while the amino acid composition of the HMW adhesin remains unaltered. As the number of repeats increases, *hmwA* transcription and translation decrease in a graded fashion (19). Furthermore, studies of tetranucleotide SSRs observed that there was a linear relationship between the rate at which repeats are gained or lost and the overall length of the SSR tract (21). The mutation rate of the *hmwA* SSR has not been determined. Assuming that *hmwA* phase variation rates increase with SSR number, this means that the *hmwA* SSRs regulate both the level of protein expressed and the rate at which expression levels change. The net result of this process is that any given NTHi population contains cells displaying varying amounts of surface exposed HMW. Variable HMW production has two implications: it affects the cell's adherence capacity

and, given that HMW is highly immunogenic, it impacts a cell's ability to avoid clearance due to antibody mediated immunity.

Phase variable HMW adhesin expression has important implications for NTHi pathogenesis, during both AOM and COPD, and for vaccine development. Healthy children and adults are often simultaneously colonized with multiple different strains of NTHi, but during AOM the NTHi population is homogeneous and there is typically only a single strain present in both the middle ear and the nasopharynx (14, 48, 57, 65). Strains isolated from these different sites are often isogenic based on study-specific typing criteria, but examination of the SSR repeats within the *hmwA* promoter often reveal differences in repeat numbers. More specifically, middle ear isolates generally have a greater number of SSR repeats, and express less HMW adhesin, than do their isogenic counterparts isolated from the nasopharynx (19). *hmwA*-associated SSRs have also been studied longitudinally among individuals suffering from COPD by comparing isogenic NTHi strains collected from sputum samples during repeated physician visits. In general, the number of SSRs within the *hmwA* promoter increased over time and this increase was associated with decreased *in vitro* adherence to respiratory epithelial cells (18). Finally, phase variable expression of HMW adhesins has the potential to mediate escape from antibody mediated immunity as was highlighted in a vaccine study exploring the impact of immunization with HMW using a chinchilla model of experimental otitis media (7, 19). In this study, 5/10 immunized animals were protected when challenged with an isogenic NTHi isolate (7) and further examination of the NTHi isolates collected from the unprotected animals revealed an association between vaccine failure and increased numbers of *hmwA* SSRs (19). Each of these studies suggest that pathogenic NTHi strains benefit from down-regulation of HMW adhesins and highlight the complexity of the association, which varies in both space and time, between HMW adhesin phase variation and disease.

**Mathematical analyses of SSR phase variation.** Estimating mutation rates *in vivo* is, at best, a laborious process owing, in part, to the relative rarity of most mutational events. These challenges can be compounded by the absence of an easily detectable phenotype associated with the mutational event of interest. For example, changes in the *hmwA*-associated SSRs result in incremental changes in HMW-protein production which are difficult to monitor at the level of individual cells. While some experimental systems of mutational processes are amenable to

manipulation, for example, using recombinant DNA technology to insert reporters such as *lacZ*, this is not always possible. Mathematical modeling provides a way to generate predictions, hypotheses, and to extend our intuition on experimentally intractable systems. As a case in point, mathematical modeling has be used to elucidate the conditions under which phase variable gene regulation would evolve and to estimate optimal phase variation rates under varying environmental conditions (45, 52, 55, 71).

To date, mathematical models of SSR phase variation have primarily focused on two-state ON and OFF systems; that is, the phase variable character is either expressed or it is not. These models typically consist of two loci, a mutator locus and a two-state phenotypic locus, within the context of a fluctuating environment. The fitness of each phenotype in a given environment is defined with a selection coefficient, *s*, and the length of time spent in each environment is defined as *T*. Natural selection acts directly on the phenotypic locus but selection on the mutator locus is indirect and is a consequence of genetic linkage between the two loci. Within the context of on-off phase variation, the mutator locus would represent a SSR region and the phenotypic locus a gene responsible for encoding an adhesin; changes in SSR number shift the adhesin gene in or out of frame resulting in on-off phase variable adhesin expression.

When only two different environmental conditions exist and the rate of mutation is the same in both directions, that is, the on-off rate is equal to the off-on rate, then the reciprocal of the average duration to time spent in each environment (1/T), approximates the optimal mutation rate (45, 55). Extending the two-locus model to incorporate multiple phenotypes and multiple environments, Kussell and Leibler (52) demonstrated that the optimal mutation rates were solely determined by the duration of time spent in each environment and the probability of an environmental change. Their analysis assumed, however, sufficient time was spent in each environment for the population to reach equilibrium. Under these conditions, optimal mutation rates from phenotype *a* to *b* were inversely proportional to the time spent in the environment most suitable for phenotype *a* and proportional to the probability the environment changes from one favorable to phenotype *a* to one favorable to phenotype *b*.

Saunders *et al*, investigated the impact of phase variation on population structure using a dynamical model in which bacteria display one of two phenotypes, A or B, and each phenotype has a specified fitness. The rate of switching between phenotypes could vary; that is, the rate of

switching from A to B does not have to be the same as from B to A, as could the fitness associated with each phenotype (81). In the absence of fitness differences, mutation rates were the sole determinant of the population structure at equilibrium and, if switching rates were equal, the time to equilibrium was approximately equal to the reciprocal of the mutation rate. When the phenotypes differed in fitness, the population structure and the time it took for the population to reach equilibrium was determined by the fitness difference. Within the context of a colonizing bacterial population, this means that even when the mutation rates are low–so low that, in the absence of selection, equilibrium would not be achievable in a biologically plausible time frame–a rare phenotype with a fitness advantage of 10% could nearly replace the initial phenotype in a matter of days. Specific immune responses can presumably result in fitness differences much greater than 10% (81) thus, even when the mutation rate is low, a rare phenotype capable of evading specific host immune responses (*e.g.*, antibody mediated immunity), could nearly replace the less-fit phenotype within a matter of days. To frame this within the context of *hmwA* phase variation, in the presence of a threshold effect such that cells expressing very low levels of HMW adhesins were capable of evading antibody-mediated immunity, these cells would realize an increased fitness relative to cells with high levels of HMW adhesins and, thus, would predominate.

More recently, Palmer *et al.* (71), simulated changes in a population of individuals under varying selective pressure (*s*) and duration of time in each environment (*T*), where each individual encodes a single SSR locus with mutation rates that vary depending on the length of the SSR tract; as the tract length increases mutation rate increases. This still represents a two-locus model of phase variation with one locus consisting of a SSR and the other an on-off selectable phenotype, but it is a significant extension of previous models for the following reasons: (1) phase variation rates vary among individuals as a function of SSR tract length and (2) it parameterizes the mutation rates based on the *in vitro* behavior of *H. influenzae* tetranucleotide SSRs , with respect to mutation rates and types of mutational events as determined by De Bolle *et al.*(21). They found that a wide range of conditions favor the evolution of SSRs with adjustable mutation rate; that is, the mutation rate is a function of SSR tract length. In general, when the product of the selection coefficient (*s*) and the duration of time spent in each environment (*T*) is large enough for the favored phenotype to reach fixation before the environment switched, regardless of whether *s* and *T* were symmetrical, phase variation was

favored. Nonminimal phase variation rates were also favored when *s*T* was low but symmetrical, *i.e.*, the product of the selection coefficient and the duration of time is equivalent in both environments. In contrast, when *s*T* was small and differed between environments (*i.e.*, *s*T* was asymmetrical), only minimal phase variation rates evolved, meaning that the SSR tracts were relatively short and repeats were gained or lost at a slow rate.

HMW adhesin phase variation can be placed within a two-locus framework, but, there are some critical distinctions between the classic two-locus model and HMW adhesin phase variation. The HMW adhesins do not display on-off phase variation but, instead, changes in the number of SSRs within the *hmwA* promoter (the mutator locus) affect HMW adhesin production (the phenotypic locus) in a graded fashion. A second distinction involves the way in which environmental changes have been modeled, as classic two-locus models generally consider alternating transitions between two different environments. Alternating environments can occur both within and between hosts, *e.g.*, during a transmission events where each host would represent a distinct environment. One of the more significant challenges facing a NTHi population during colonization is likely to be antibody mediated immunity; this is especially true with regard to the HMW adhesins as they represent an immunodominant surface protein. In the scenario of a naïve host, antibody mediated immunity would apply a gradually increasing selective pressure, not an alternating selective pressure as most classic two-locus models assume. To date I am not aware of models that capture the behavior of systems analogous to HMW adhesin SSR mediated phase variation, but this system is amenable to mathematical modeling that could provide valuable insight into the role of HMW adhesin phase variation during colonization.

**Research objectives.** The research presented in this dissertation aims to: (1) describe a molecular typing method to differentiate typeable from true nontypeable *H. influenzae*, (2) characterize a geographically distributed collection of NTHi isolates with respect to *hmwA* prevalence and sequence diversity, (3) test for evidence of selective pressures operating on *hmwA*, and (4) develop a mathematical model that will provide a theoretical framework for exploring how *hmwA* phase variation and host mediated immunity interact to shape NTHi within-host population structure. I tackle these aims from an interdisciplinary perspective, synthesizing epidemiological data, molecular evolutionary analyses, and mathematical modeling.

Since the HMW adhesins are still being pursued as potential candidates for inclusion in a multi-valent NTHi vaccine, this work has both practical and theoretical implications for ongoing efforts aimed at developing an effective NTHi vaccine. At a more basic level, this work will advance our understanding of host-pathogen interactions, and importantly, host-commensal interactions with implications that reach well beyond *H. influenzae*, as phase and antigenic variation are common themes among human-restricted bacteria, for example, *Helicobacter pylori* and *Neisseria meningitidis*.

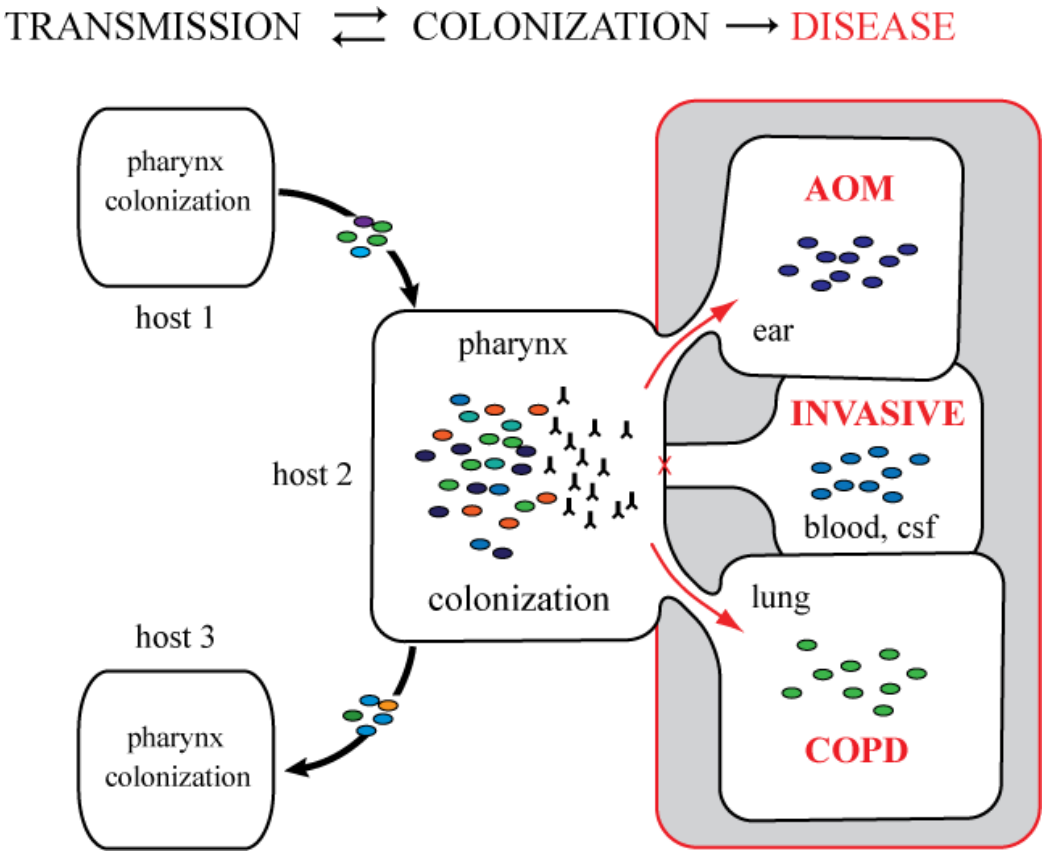



Figure 1.1. Conceptual model of *H. influenzae* transmission, colonization, and disease. *H. influenzae* are small cococobacilli that reside exclusively within humans, typically within the nasopharynges. *H. influenzae* are transmitted person to person via infected respiratory droplets. Encapsulated strains are divided into six capsular serotypes (a–f), with serotype b (Hib) being commonly associated with invasive human diseases such as meningitis and septicemia among non-immune children. In contrast, nonencapsulated strains, which are commonly referred to as nontypeable *H. influenzae* (NTHi), are generally associated with localized infections of the respiratory tract such as pneumonia, sinusitis, and acute otitis media (AOM). NTHi colonization occurs early during childhood and is a dynamic process marked by simultaneous colonization with multiple strains (indicated with different colors in the figure) and presumably high rates of strain turnover (27, 38, 61, 90, 96, 100). NTHi colonization stimulates an adaptive immune response resulting in serum IgG, IgM, and IgA, many of which target outer membrane proteins (9, 10, 34, 46, 73). Pathogenic NTHi, *e.g.*, those causing AOM, arise from the community of colonizing NTHi strains. While the community of *H. influenzae* colonizing the pharynx can contain multiple different strains, disease-causing isolates collected from normally sterile sites are typically clonal (as indicated with the homogeneous strains in each disease site). While the exact mechanisms leading to invasion of normally sterile sites by disease causing strains is not fully understood, the bacterial factors associated with pathogenicity, such as, virulence factors, may differ by site of infection (indicated by the different colored cells in associated with AOM, invasive disease, and COPD).

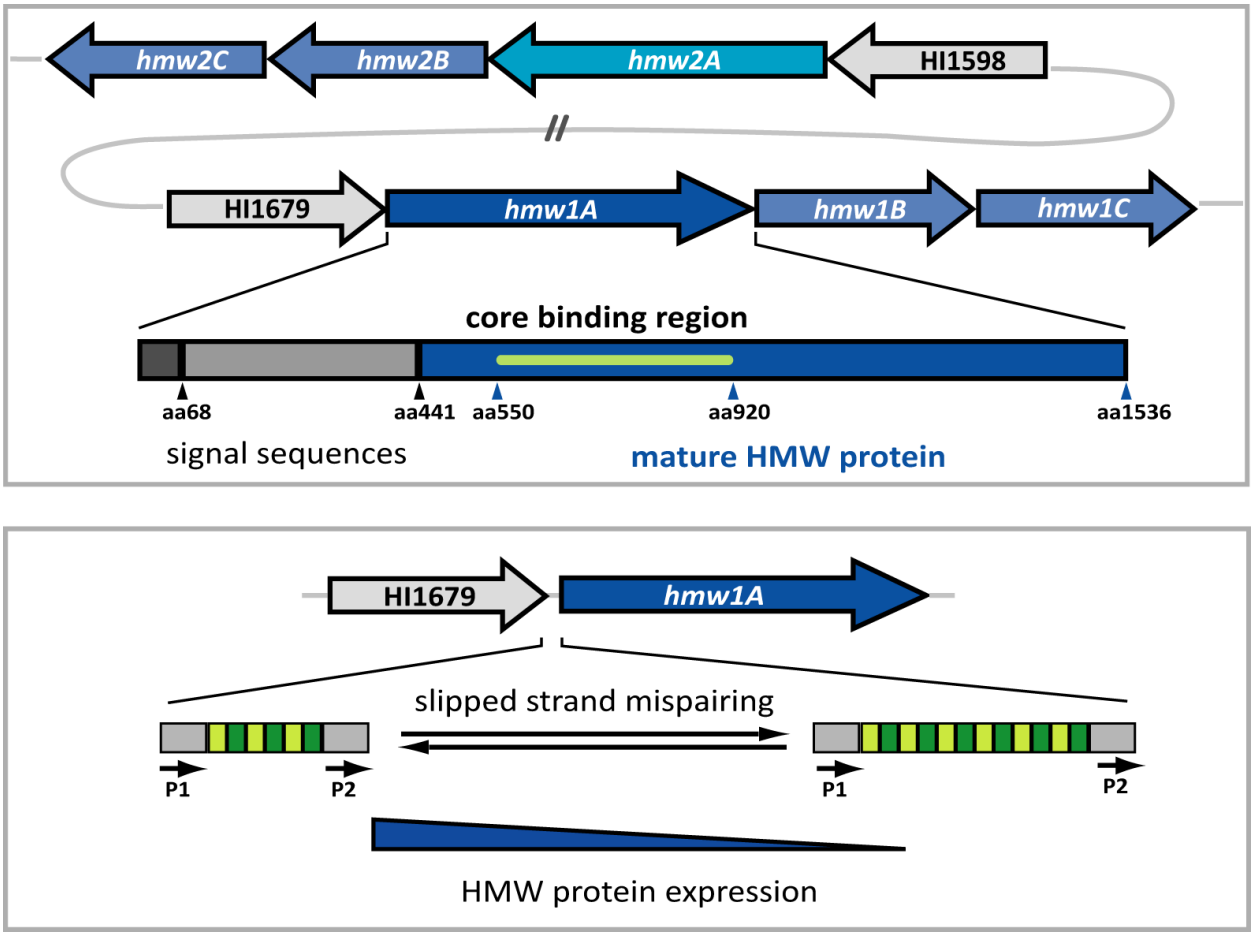



Figure 1.2. Chromosomal arrangement of the *hmw* loci (upper) and arrangement of the 7-bp repeat region upstream of *hmwA* (lower). HMW adhesins are encoded by the *hmw* locus which contains three genes, *hmwA*, *hmwB*, and *hmwC* (12). Each *hmw*-positive strain possesses two complete copies of the *hmw* locus, designated *hmw1* and *hmw2*, which are located in conserved but unlinked locations within the NTHi chromosome (16). The functional HMW adhesins are encoded by *hmw1A* and *hmw2A*, which in strain 12 share 71% amino acid identity (80% similarity) (11). The HMW binding domain is responsible for binding the host epithelium (20) (lower) Expression of both *hmwA* copies is phase variable and is mediated by 7-base pair (bp) tandem nucleotide repeats located within the *hmwA* promoter region (19). Repeats are gained and lost stochastically during DNA synthesis by slipped stranded mispairing and changes in the number of tandem repeats affects *hmwA* transcription and translation; as the repeat tract length increases, *hmwA* transcription and translation decreases (19).

## Literature Cited


1. Ah-Tye, C., J. L. Paradise, and D. K. Colborn. 2001. Otorrhea in young children after tympanostomy-tube placement for persistent middle-ear effusion: prevalence, incidence, and duration. Pediatrics 107:1251-8.
2. Alm, R. A., L. S. Ling, D. T. Moir, B. L. King, E. D. Brown, P. C. Doig, D. R. Smith, B. Noonan, B. C. Guild, B. L. deJonge, G. Carmel, P. J. Tummino, A. Caruso, M. Uria-Nickelsen, D. M. Mills, C. Ives, R. Gibson, D. Merberg, S. D. Mills, Q. Jiang, D. E. Taylor, G. F. Vovis, and T. J. Trust. 1999. Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen Helicobacter pylori. Nature 397:176-80.
3. American Speech, L. a. H. A. 2007, posting date. Causes of hearing loss in children 1997 - 2003. [Online.]
4. Bakaletz, L. O., B. D. Baker, J. A. Jurcisek, A. Harrison, L. A. Novotny, J. E. Bookwalter, R. Mungur, and R. S. Munson, Jr. 2005. Demonstration of Type IV pilus expression and a twitching phenotype by *Haemophilus influenzae*. Infect Immun 73:1635-43.
5. Bakaletz, L. O., and S. J. Barenkamp. 1994. Localization of high-molecular-weight adhesion proteins of nontypeable *Haemophilus influenzae* by immunoelectron microscopy. Infect Immun 62:4460-8.
6. Bakaletz, L. O., B. M. Tallan, T. Hoepf, T. F. DeMaria, H. G. Birck, and D. J. Lim. 1988. Frequency of fimbriation of nontypable *Haemophilus influenzae* and its ability to adhere to chinchilla and human respiratory epithelium. Infect Immun 56:331-5.
7. Barenkamp, S. J. 1996. Immunization with high-molecular-weight adhesion proteins of nontypeable *Haemophilus influenzae* modifies experimental otitis media in chinchillas. Infect Immun 64:1246-51.
8. Barenkamp, S. J. 1992. Outer membrane proteins and lipopolysaccharides of nontypeable *Haemophilus influenzae*. J Infect Dis 165 Suppl 1:S181-4.
9. Barenkamp, S. J. 1986. Protection by serum antibodies in experimental nontypable Haemophilus influenzae otitis media. Infect Immun 52:572-8.
10. Barenkamp, S. J., and F. F. Bodor. 1990. Development of serum bactericidal activity following nontypable *Haemophilus influenzae* acute otitis media. Pediatr Infect Dis J 9:333-9.
11. Barenkamp, S. J., and E. Leininger. 1992. Cloning, expression, and DNA sequence analysis of genes encoding nontypeable *Haemophilus influenzae* high-molecular-weight surface-exposed proteins related to filamentous hemagglutinin of *Bordetella pertussis*. Infect Immun 60:1302-13.
12. Barenkamp, S. J., and J. W. St Geme, 3rd. 1994. Genes encoding high-molecular-weight adhesion proteins of nontypeable *Haemophilus influenzae* are part of gene clusters. Infect Immun 62:3320-8.
13. Barenkamp, S. J., and J. W. St Geme, 3rd. 1996. Identification of a second family of high-molecular-weight adhesion proteins expressed by non-typable *Haemophilus influenzae*. Mol Microbiol 19:1215-23.
14. Berrens, Z. J., C. F. Marrs, M. M. Pettigrew, S. A. Sandstedt, M. Patel, and J. R. Gilsdorf. 2007. Genetic diversity of paired middle-ear and pharyngeal nontypeable *Haemophilus influenzae* isolates from children with acute otitis media. J Clin Microbiol 45:3764-7.

15. Bou, R., A. Dominguez, D. Fontanals, I. Sanfeliu, I. Pons, J. Renau, V. Pineda, E. Lobera, C. Latorre, M. Majo, and L. Salleras. 2000. Prevalence of Haemophilus influenzae pharyngeal carriers in the school population of Catalonia. Working Group on invasive disease caused by Haemophilus influenzae. Eur J Epidemiol 16:521-6.
16. Buscher, A. Z., K. Burmeister, S. J. Barenkamp, and J. W. St Geme, 3rd. 2004. Evolutionary and functional relationships among the nontypeable *Haemophilus influenzae* HMW family of adhesins. J Bacteriol 186:4209-17.
17. Casey, J. R., D. G. Adlowitz, and M. E. Pichichero. 2010. New patterns in the otopathogens causing acute otitis media six to eight years after introduction of pneumococcal conjugate vaccine. Pediatr Infect Dis J 29:304-9.
18. Cholon, D. M., D. Cutter, S. K. Richardson, S. Sethi, T. F. Murphy, D. C. Look, and J. W. St Geme, 3rd. 2008. Serial isolates of persistent *Haemophilus influenzae* in patients with chronic obstructive pulmonary disease express diminishing quantities of the HMW1 and HMW2 adhesins. Infect Immun 76:4463-8.
19. Dawid, S., S. J. Barenkamp, and J. W. St Geme, 3rd. 1999. Variation in expression of the *Haemophilus influenzae* HMW adhesins: a prokaryotic system reminiscent of eukaryotes. Proc Natl Acad Sci U S A 96:1077-82.
20. Dawid, S., S. Grass, and J. W. St Geme, 3rd. 2001. Mapping of binding domains of nontypeable *Haemophilus influenzae* HMW1 and HMW2 adhesins. Infect Immun 69:307-14.
21. De Bolle, X., C. D. Bayliss, D. Field, T. van de Ven, N. J. Saunders, D. W. Hood, and E. R. Moxon. 2000. The length of a tetranucleotide repeat tract in *Haemophilus influenzae* determines the phase variation rate of a gene with homology to type III DNA methyltransferases. Mol Microbiol 35:211-22.
22. Drake, J. W. 1991. A constant rate of spontaneous mutation in DNA-based microbes. Proc Natl Acad Sci U S A 88:7160-4.
23. Ecevit, I. Z., K. W. McCrea, C. F. Marrs, and J. R. Gilsdorf. 2005. Identification of new hmwA alleles from nontypeable *Haemophilus influenzae*. Infect Immun 73:1221-5.
24. Ecevit, I. Z., K. W. McCrea, M. M. Pettigrew, A. Sen, C. F. Marrs, and J. R. Gilsdorf. 2004. Prevalence of the *hifBC, hmw1A, hmw2A, hmwC*, and hia Genes in *Haemophilus influenzae* Isolates. J Clin Microbiol 42:3065-72.
25. Erwin, A. L., K. L. Nelson, T. Mhlanga-Mutangadura, P. J. Bonthuis, J. L. Geelhood, G. Morlin, W. C. Unrath, J. Campos, D. W. Crook, M. M. Farley, F. W. Henderson, R. F. Jacobs, K. Muhlemann, S. W. Satola, L. van Alphen, M. Golomb, and A. L. Smith. 2005. Characterization of genetic and phenotypic diversity of invasive nontypeable Haemophilus influenzae. Infect Immun 73:5853-63.
26. Erwin, A. L., S. A. Sandstedt, P. J. Bonthuis, J. L. Geelhood, K. L. Nelson, W. C. Unrath, M. A. Diggle, M. J. Theodore, C. R. Pleatman, E. A. Mothershed, C. T. Sacchi, L. W. Mayer, J. R. Gilsdorf, and A. L. Smith. 2008. Analysis of genetic relatedness of *Haemophilus influenzae* isolates by multilocus sequence typing. J Bacteriol 190:1473-83.
27. Faden, H., L. Duffy, A. Williams, D. A. Krystofik, and J. Wolf. 1995. Epidemiology of nasopharyngeal colonization with nontypeable *Haemophilus influenzae* in the first 2 years of life. J Infect Dis 172:132-5.
28. Falla, T. J., D. W. Crook, L. N. Brophy, D. Maskell, J. S. Kroll, and E. R. Moxon. 1994. PCR for capsular typing of *Haemophilus influenzae*. J Clin Microbiol 32:2382-6.

29. Farjo, R. S., B. Foxman, M. J. Patel, L. Zhang, M. M. Pettigrew, S. I. McCoy, C. F. Marrs, and J. R. Gilsdorf. 2004. Diversity and sharing of *Haemophilus influenzae* strains colonizing healthy children attending day-care centers. Pediatr Infect Dis J 23:41-6.
30. Finkelstein, J. A., J. P. Metlay, R. L. Davis, S. L. Rifas-Shiman, S. F. Dowell, and R. Platt. 2000. Antimicrobial use in defined populations of infants and young children. Arch Pediatr Adolesc Med 154:395-400.
31. Fitzpatrick, D. A., and J. O. McInerney. 2005. Evidence of positive Darwinian selection in Omp85, a highly conserved bacterial outer membrane protein essential for cell viability. J Mol Evol 60:268-73.
32. Fleischmann, R. D., M. D. Adams, O. White, R. A. Clayton, E. F. Kirkness, A. R. Kerlavage, C. J. Bult, J. F. Tomb, B. A. Dougherty, J. M. Merrick, and et al. 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science 269:496-512.
33. Gilsdorf, J. R., K. W. McCrea, and C. F. Marrs. 1997. Role of pili in *Haemophilus influenzae* adherence and colonization. Infect Immun 65:2997-3002.
34. Giufre, M., M. Muscillo, P. Spigaglia, R. Cardines, P. Mastrantonio, and M. Cerquetti. 2006. Conservation and diversity of HMW1 and HMW2 adhesin binding domains among invasive nontypeable *Haemophilus influenzae* isolates. Infect Immun 74:1161-70.
35. Gnehm, H. E., S. I. Pelton, S. Gulati, and P. A. Rice. 1985. Characterization of antigens from nontypable Haemophilus influenzae recognized by human bactericidal antibodies. Role of Haemophilus outer membrane proteins. J Clin Invest 75:1645-58.
36. Grass, S., A. Z. Buscher, W. E. Swords, M. A. Apicella, S. J. Barenkamp, N. Ozchlewski, and J. W. St Geme, 3rd. 2003. The *Haemophilus influenzae* HMW1 adhesin is glycosylated in a process that requires HMW1C and phosphoglucomutase, an enzyme involved in lipooligosaccharide biosynthesis. Mol Microbiol 48:737-51.
37. Grass, S., C. F. Lichti, R. R. Townsend, J. Gross, and J. W. St Geme, 3rd. 2010. The Haemophilus influenzae HMW1C protein is a glycosyltransferase that transfers hexose residues to asparagine sites in the HMW1 adhesin. PLoS Pathog 6:e1000919.
38. Grass, S., and J. W. St Geme, 3rd. 2000. Maturation and secretion of the non-typable *Haemophilus influenzae* HMW1 adhesin: roles of the N-terminal and C-terminal domains. Mol Microbiol 36:55-67.
39. Gratten, M., J. Montgomery, G. Gerega, H. Gratten, H. Siwi, A. Poli, and G. Koki. 1989. Multiple colonization of the upper respiratory tract of Papua New Guinea children with Haemophilus influenzae and Streptococcus pneumoniae. Southeast Asian J Trop Med Public Health 20:501-9.
40. Gross, J., S. Grass, A. E. Davis, P. Gilmore-Erdmann, R. R. Townsend, and J. W. St Geme, 3rd. 2008. The Haemophilus influenzae HMW1 adhesin is a glycoprotein with an unusual N-linked carbohydrate modification. J Biol Chem 283:26010-5.
41. Hansen, M. V., D. M. Musher, and R. E. Baughn. 1985. Outer membrane proteins of nontypable Haemophilus influenzae and reactivity of paired sera from infected patients with their homologous isolates. Infect Immun 47:843-6.
42. Harabuchi, Y., H. Faden, N. Yamanaka, L. Duffy, J. Wolf, and D. Krystofik. 1994. Nasopharyngeal colonization with nontypeable *Haemophilus influenzae* and recurrent otitis media. Tonawanda/Williamsville Pediatrics. J Infect Dis 170:862-6.
43. Health, N. I. o. 2012. Morbidity and mortality: 2012 chart book on cardiovascular, lung, and blood diseases. *In* N. H. L. a. B. Institute. (ed.). National Institutes of Health.

44. Hood, D. W., M. E. Deadman, M. P. Jennings, M. Bisercic, R. D. Fleischmann, J. C. Venter, and E. R. Moxon. 1996. DNA repeats identify novel virulence genes in Haemophilus influenzae. Proc Natl Acad Sci U S A 93:11121-5.
45. Ishii, K., H. Matsuda, Y. Iwasa, and A. Sasaki. 1989. Evolutionarily stable mutation rate in a periodically changing environment. Genetics 121:163-74.
46. Iyer Parameswaran, G., and T. F. Murphy. 2009. Chronic obstructive pulmonary disease: role of bacteria and updated guide to antibacterial selection in the older patient. Drugs Aging 26:985-95.
47. Karasic, R. B., C. E. Trumpp, H. E. Gnehm, P. A. Rice, and S. I. Pelton. 1985. Modification of otitis media in chinchillas rechallenged with nontypable Haemophilus influenzae and serological response to outer membrane antigens. J Infect Dis 151:273-9.
48. Krasan, G. P., D. Cutter, S. L. Block, and J. W. St Geme, 3rd. 1999. Adhesin expression in matched nasopharyngeal and middle ear isolates of nontypeable *Haemophilus influenzae* from children with acute otitis media. Infect Immun 67:449-54.
49. 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. Mol Microbiol 4:1853-62.
50. Kroll, J. S., B. M. Loynds, and E. R. Moxon. 1991. The *Haemophilus influenzae* capsulation gene cluster: a compound transposon. Mol Microbiol 5:1549-60.
51. Kroll, J. S., S. Zamze, B. Loynds, and E. R. Moxon. 1989. Common organization of chromosomal loci for production of different capsular polysaccharides in *Haemophilus influenzae*. J Bacteriol 171:3343-7.
52. Kussell, E., and S. Leibler. 2005. Phenotypic diversity, population growth, and information in fluctuating environments. Science 309:2075-8.
53. Lacross, N. C., C. F. Marrs, M. Patel, S. A. Sandstedt, and J. R. Gilsdorf. 2008. High genetic diversity of nontypeable *Haemophilus influenzae* isolates from two children attending a day care center. J Clin Microbiol 46:3817-21.
54. Lebon, A., H. A. Moll, M. Tavakol, W. J. van Wamel, V. W. Jaddoe, A. Hofman, H. A. Verbrugh, and A. van Belkum. 2010. Correlation of bacterial colonization status between mother and child: the Generation R Study. J Clin Microbiol 48:960-2.
55. Leigh, E. G., Jr. 1970. Natural Selection and Mutability. The American Naturalist 104:301-305.
56. Lindenauer, P. K., P. Pekow, S. Gao, A. S. Crawford, B. Gutierrez, and E. M. Benjamin. 2006. Quality of care for patients hospitalized for acute exacerbations of chronic obstructive pulmonary disease. Ann Intern Med 144:894-903.
57. Loos, B. G., J. M. Bernstein, D. M. Dryja, T. F. Murphy, and D. P. Dickinson. 1989. Determination of the epidemiology and transmission of nontypable Haemophilus influenzae in children with otitis media by comparison of total genomic DNA restriction fingerprints. Infect Immun 57:2751-7.
58. Maaroufi, Y., J. M. De Bruyne, C. Heymans, and F. Crokaert. 2007. Real-time PCR for determining capsular serotypes of *Haemophilus influenzae*. J Clin Microbiol 45:2305-8.
59. McIsaac, W. J., P. C. Coyte, R. Croxford, C. V. Asche, J. Friedberg, and W. Feldman. 2000. Otolaryngologists' perceptions of the indications for tympanostomy tube insertion in children. CMAJ 162:1285-8.

60. Mes, T. H., and J. P. van Putten. 2007. Positively selected codons in immune-exposed loops of the vaccine candidate OMP-P1 of Haemophilus influenzae. J Mol Evol 64:411-22.
61. Monasta, L., L. Ronfani, F. Marchetti, M. Montico, L. Vecchi Brumatti, A. Bavcar, D. Grasso, C. Barbiero, and G. Tamburlini. 2012. Burden of disease caused by otitis media: systematic review and global estimates. PLoS One 7:e36226.
62. Mukundan, D., Z. Ecevit, M. Patel, C. F. Marrs, and J. R. Gilsdorf. 2007. Pharyngeal colonization dynamics of *Haemophilus influenzae* and *Haemophilus haemolyticus* in healthy adult carriers. J Clin Microbiol 45:3207-17.
63. Munson, R. S., Jr., S. Grass, and R. West. 1993. Molecular cloning and sequence of the gene for outer membrane protein P5 of Haemophilus influenzae. Infect Immun 61:4017-20.
64. Munson, R. S., Jr., A. Harrison, A. Gillaspy, W. C. Ray, M. Carson, D. Armbruster, J. Gipson, M. Gipson, L. Johnson, L. Lewis, D. W. Dyer, and L. O. Bakaletz. 2004. Partial analysis of the genomes of two nontypeable Haemophilus influenzae otitis media isolates. Infect Immun 72:3002-10.
65. Murphy, T. F., J. M. Bernstein, D. M. Dryja, A. A. Campagnari, and M. A. Apicella. 1987. Outer membrane protein and lipooligosaccharide analysis of paired nasopharyngeal and middle ear isolates in otitis media due to nontypable Haemophilus influenzae: pathogenetic and epidemiological observations. J Infect Dis 156:723-31.
66. Murphy, T. F., H. Faden, L. O. Bakaletz, J. M. Kyd, A. Forsgren, J. Campos, M. Virji, and S. I. Pelton. 2009. Nontypeable *Haemophilus influenzae* as a Pathogen in Children. Pediatr Infect Dis J.
67. Musher, D. M., M. Hague-Park, R. E. Baughn, R. J. Wallace, Jr., and B. Cowley. 1983. Opsonizing and bactericidal effects of normal human serum on nontypable Haemophilus influenzae. Infect Immun 39:297-304.
68. Muzzi, A., M. Moschioni, A. Covacci, R. Rappuoli, and C. Donati. 2008. Pilus operon evolution in Streptococcus pneumoniae is driven by positive selection and recombination. PLoS One 3:e3660.
69. O'Brien, M. A., L. A. Prosser, J. L. Paradise, G. T. Ray, M. Kulldorff, M. Kurs-Lasky, V. L. Hinrichsen, J. Mehta, D. K. Colborn, and T. A. Lieu. 2009. New vaccines against otitis media: projected benefits and cost-effectiveness. Pediatrics 123:1452-63.
70. Organization, W. H. June 2011 2011, posting date. The 10 leading causes of death by broad income group (2008). World Health Organization. [Online.]
71. Palmer, M. E., M. Lipsitch, E. R. Moxon, and C. D. Bayliss. 2013. Broad conditions favor the evolution of phase-variable Loci. MBio 4.
72. Parkhill, J., M. Achtman, K. D. James, S. D. Bentley, C. Churcher, S. R. Klee, G. Morelli, D. Basham, D. Brown, T. Chillingworth, R. M. Davies, P. Davis, K. Devlin, T. Feltwell, N. Hamlin, S. Holroyd, K. Jagels, S. Leather, S. Moule, K. Mungall, M. A. Quail, M. A. Rajandream, K. M. Rutherford, M. Simmonds, J. Skelton, S. Whitehead, B. G. Spratt, and B. G. Barrell. 2000. Complete DNA sequence of a serogroup A strain of Neisseria meningitidis Z2491. Nature 404:502-6.
73. Pelton, S. I., and E. Leibovitz. 2009. Recent advances in otitis media. Pediatr Infect Dis J 28:S133-7.
74. Pichichero, M. E., R. Kaur, J. R. Casey, A. Sabirov, M. N. Khan, and A. Almudevar. 2010. Antibody response to Haemophilus influenzae outer membrane protein D, P6, and

OMP26 after nasopharyngeal colonization and acute otitis media in children. Vaccine 28:7184-92.
75. Power, P. M., W. A. Sweetman, N. J. Gallacher, M. R. Woodhall, G. A. Kumar, E. R. Moxon, and D. W. Hood. 2009. Simple sequence repeats in *Haemophilus influenzae*. Infect Genet Evol 9:216-28.
76. Prasadarao, N. V., E. Lysenko, C. A. Wass, K. S. Kim, and J. N. Weiser. 1999. Opacity-associated protein A contributes to the binding of Haemophilus influenzae to chang epithelial cells. Infect Immun 67:4153-60.
77. Prevention, C. f. D. C. a. 2011. Chronic obstructive pulmonary disease among adults--United States, 2011. MMWR Morb Mortal Wkly Rep 61:938-43.
78. Reddy, M. S., J. M. Bernstein, T. F. Murphy, and H. S. Faden. 1996. Binding between outer membrane proteins of nontypeable Haemophilus influenzae and human nasopharyngeal mucin. Infect Immun 64:1477-9.
79. Satola, S. W., P. L. Schirmer, and M. M. Farley. 2003. Complete sequence of the cap locus of *Haemophilus influenzae* serotype b and nonencapsulated b capsule-negative variants. Infect Immun 71:3639-44.
80. Saunders, N. J., A. C. Jeffries, J. F. Peden, D. W. Hood, H. Tettelin, R. Rappuoli, and E. R. Moxon. 2000. Repeat-associated phase variable genes in the complete genome sequence of Neisseria meningitidis strain MC58. Mol Microbiol 37:207-15.
81. Saunders, N. J., E. R. Moxon, and M. B. Gravenor. 2003. Mutation rates: estimating phase variation rates when fitness differences are present and their impact on population structure. Microbiology 149:485-95.
82. Sethi, S., and T. F. Murphy. 2001. Bacterial infection in chronic obstructive pulmonary disease in 2000: a state-of-the-art review. Clin Microbiol Rev 14:336-63.
83. Sirakova, T., P. E. Kolattukudy, D. Murwin, J. Billy, E. Leake, D. Lim, T. DeMaria, and L. Bakaletz. 1994. Role of fimbriae expressed by nontypeable Haemophilus influenzae in pathogenesis of and protection against otitis media and relatedness of the fimbrin subunit to outer membrane protein A. Infect Immun 62:2002-20.
84. Smith, N. H., J. Maynard Smith, and B. G. Spratt. 1995. Sequence evolution of the porB gene of Neisseria gonorrhoeae and Neisseria meningitidis: evidence of positive Darwinian selection. Mol Biol Evol 12:363-70.
85. St Geme, J. W., 3rd, and D. Cutter. 1995. Evidence that surface fibrils expressed by Haemophilus influenzae type b promote attachment to human epithelial cells. Mol Microbiol 15:77-85.
86. St Geme, J. W., 3rd, and D. Cutter. 1996. Influence of pili, fibrils, and capsule on in vitro adherence by Haemophilus influenzae type b. Mol Microbiol 21:21-31.
87. St Geme, J. W., 3rd, M. L. de la Morena, and S. Falkow. 1994. A *Haemophilus influenzae* IgA protease-like protein promotes intimate interaction with human epithelial cells. Mol Microbiol 14:217-33.
88. St Geme, J. W., 3rd, S. Falkow, and S. J. Barenkamp. 1993. High-molecular-weight proteins of nontypable *Haemophilus influenzae* mediate attachment to human epithelial cells. Proc Natl Acad Sci U S A 90:2875-9.
89. St Geme, J. W., 3rd, and S. Grass. 1998. Secretion of the Haemophilus influenzae HMW1 and HMW2 adhesins involves a periplasmic intermediate and requires the HMWB and HMWC proteins. Mol Microbiol 27:617-30.

90. St Geme, J. W., 3rd, V. V. Kumar, D. Cutter, and S. J. Barenkamp. 1998. Prevalence and distribution of the *hmw* and *hia* genes and the HMW and Hia adhesins among genetically diverse strains of nontypeable *Haemophilus influenzae*. Infect Immun 66:364-8.
91. St Sauver, J., C. F. Marrs, B. Foxman, P. Somsel, R. Madera, and J. R. Gilsdorf. 2000. Risk factors for otitis media and carriage of multiple strains of *Haemophilus influenzae* and *Streptococcus pneumoniae*. Emerg Infect Dis 6:622-30.
92. Stull, T. L., P. M. Mendelman, J. E. Haas, M. A. Schoenborn, K. D. Mack, and A. L. Smith. 1984. Characterization of Haemophilus influenzae type b fimbriae. Infect Immun 46:787-96.
93. Sukupolvi-Petty, S., S. Grass, and J. W. St Geme, 3rd. 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. J Bacteriol 188:3870-7.
94. Sung, W., M. S. Ackerman, S. F. Miller, T. G. Doak, and M. Lynch. 2012. Drift-barrier hypothesis and mutation-rate evolution. Proc Natl Acad Sci U S A 109:18488-92.
95. Swords, W. E., B. A. Buscher, K. Ver Steeg Ii, A. Preston, W. A. Nichols, J. N. Weiser, B. W. Gibson, and M. A. Apicella. 2000. Non-typeable Haemophilus influenzae adhere to and invade human bronchial epithelial cells via an interaction of lipooligosaccharide with the PAF receptor. Mol Microbiol 37:13-27.
96. Teele, D. W., J. O. Klein, and B. Rosner. 1989. Epidemiology of otitis media during the first seven years of life in children in greater Boston: a prospective, cohort study. J Infect Dis 160:83-94.
97. Trottier, S., K. Stenberg, and C. Svanborg-Eden. 1989. Turnover of nontypable *Haemophilus influenzae* in the nasopharynges of healthy children. J Clin Microbiol 27:2175-9.
98. van Alphen, L., D. A. Caugant, B. Duim, M. O'Rourke, and L. D. Bowler. 1997. Differences in genetic diversity of nonecapsulated *Haemophilus influenzae* from various diseases. Microbiology 143 ( Pt 4):1423-31.
99. van Alphen, L., L. Geelen-van den Broek, L. Blaas, M. van Ham, and J. Dankert. 1991. Blocking of fimbria-mediated adherence of Haemophilus influenzae by sialyl gangliosides. Infect Immun 59:4473-7.
100. van Schilfgaarde, M., P. van Ulsen, P. Eijk, M. Brand, M. Stam, J. Kouame, L. van Alphen, and J. Dankert. 2000. Characterization of adherence of nontypeable *Haemophilus influenzae* to human epithelial cells. Infect Immun 68:4658-65.
101. Vives, M., M. E. Garcia, P. Saenz, M. A. Mora, L. Mata, H. Sabharwal, and C. Svanborg. 1997. Nasopharyngeal colonization in Costa Rican children during the first year of life. Pediatr Infect Dis J 16:852-8.

## Chapter 2

## Use of *bexB* to detect the capsule locus in *Haemophilus influenzae*[1]


### Abstract

*Haemophilus influenzae* are classified as typeable or nontypeable (NTHi) based upon the presence or absence of capsule. In addition to serotyping, which is subject to false positive results, typeable strains can be identified through the detection of the capsular export gene *bexA* and one of six capsule-specific genes, which is resource intensive, especially when characterizing large numbers of strains. To address these challenges, we developed a *bexB*-based method to differentiate true NTHi from typeable strains. We validated a PCR-based method to detect *bexB* in ten strains whose capsule status was well-defined. Among 40 strains that were previously serotype positive in clinical microbiology laboratories, five lacked *bexA, bexB* and capsule type-specific genes by PCR analysis and, thus likely represent false positive serotyping results. Among 94 additional otitis media, commensal, and serotype b negative invasive strains, 85 were *bexA* and *bexB* negative and nine contained either a complete or partial capsule locus, *i.e.* eight were *bexA* and *bexB*-positive and one was *bexA*-negative but *bexB*-positive. Finally, we adapted the method for use in a high-throughput DNA hybridization-based microarray method, which showed 98.75 and 97.5% concordance to the PCR method for *bexA* and *bexB,* respectively. In addition, *bexB* showed 84% or greater nucleotide identity among strains containing the capsule locus. In this study, we demonstrate that *bexB* is a reliable proxy for the capsule locus and its detection provides a simple and reliable method for differentiating strains that lack the entire capsule locus from those containing a partial or complete capsule locus.


[1] This work has been published with the following citation: Davis GS, Sandstedt SA, Patel M, Marrs CF, Gilsdorf JR. Use of bexB to detect the capsule locus in *Haemophilus influenzae*. 2011. J. Clin. Microbiol. 49(7): 2594-601.

## Introduction

*Haemophilus influenzae*, gram-negative bacteria that commonly reside as commensals within the human pharynx and may cause respiratory or invasive infections, are variable in presence of a polysaccharide capsule. Historically, *H. influenzae* lacking a capsule were detected by their failure to react with antibodies directed against the six immunologically distinct capsular polysaccharides and were designated as nontypeable, i.e. serotype negative, *H. influenzae* (NTHi) (33). Nontypeable strains predominate among commensal organisms (9, 29) and those causing respiratory infections such as with acute otitis media (AOM), sinusitis, or bronchitis, while serotypeable strains are more often isolated from clinical samples of patients with invasive infections such as bacteremia, meningitis, facial cellulitis and septic arthritis. In addition, multilocus sequence typing has shown that NTHi are genetically widely variable while serotypeable strains are more clonal (25).

The polysaccharide capsule of typeable *H. influenzae* strains is encoded by the *cap*-locus composed of three distinct regions, designated I – III (Figure 2.1b). The genes contained within regions I and III, designated *bexDCBA* and *hcsAB*, respectively, are highly conserved across all capsular types and are required for transport of capsule constituents across the outer membrane (20, 21, 23, 37, 40). Region II genes encode capsule types a through f-specific proteins and, thus, vary by serotype. The organization and genetics of the *cap*-locus are complex; duplications, partial loss, and complete loss of the *cap*-locus have been documented (3, 4, 12, 14-17, 19, 21, 22, 32, 37, 38, 44).

Differentiating typeable from nontypeable *H. influenzae* can be challenging. Type specific serum agglutination has classically been used to confirm the presence and specificity of the *H. influenzae* capsule. A strain may, however, fail to react with typing sera, and thus be classified as "non-typeable," for several reasons. First, inaccuracies in performing and interpreting slide agglutination tests have been well-documented (30, 45). Second, strains with one copy of the *cap* region in which *bexA* is partially deleted (3, 4, 12, 14, 15, 17, 19) are referred to as capsule-deficient variants as they contain a majority of the *cap*-locus; genetically these isolates are *cap*-locus positive, but show altered ability to produce functional capsule and fail to agglutinate with typing serum, *i.e.*, they are nonserotypeable. Third, a previously serotypeable strain could have a deletion of the entire capsule locus, as apparently occurred with

strain Rd, a nontypeable variant of a type d strain that has been extensively passed in the laboratory (37). Finally, a strain may lack the entire *cap*-locus as a consequence of long past evolutionary events, *i.e.*, be a true NTHi.

Detection of *bexA* and capsule specific genes by polymerase chain reaction (PCR) is the most commonly used method of capsule genotyping. PCR methods based on detection of *bexA* alone, however, fail to identify capsule deficient variants because the 5' primer of the standard primer pair hybridizes in the deleted region of *bexA*. This problem was overcome by the addition of PCR analyses to detect individual types a-f capsule specific regions, along with *bexA*, for each strain (8, 26). This approach, however, requires up to seven separate PCR reactions, making it both labor and resource intensive when studying a large number of isolates, especially when clinical or epidemiological data suggest that a significant portion of the strains may be NTHi. This study was motivated by the need for a simple and accurate method to differentiate true NTHi, from genetically *cap* locus positive *H. influenzae* strains.

We hypothesized that *bexB*, which is located within region I of the *cap* locus (Fig 2.1), could serve as a more reliable marker for the capsule locus than *bexA*, based on the observations that region I genes are present across all capsular types (20, 23) and there are no reports of *bexB* partial deletions analogous to those observed in *bexA*. We first validated the PCR method to detect *bexB* using *H. influenzae* strains whose capsule status had been well defined. We then tested 40 *H. influenzae* strains that had been serotype positive when tested in clinical microbiology laboratories. To better understand the prevalence of genetically *cap* region-positive strains that fail to produce a *bexA* amplicon, we then tested the *bexB* PCR method on invasive strains that were serotype negative in the clinical microbiology laboratories and on strains likely to be nontypeable based upon their site of isolation (*i.e*, commensal and otitis media isolates). Finally, we adapted the molecular capsular typing method for use in a high-throughput hybridization format and tested it with a microarray containing the genomic DNA of 546 *H. influenzae* strains. We demonstrate here that screening for *bexB* provides a simple and reliable method for distinguishing *H. influenzae* strains containing a complete or partial capsule locus from those lacking the capsule locus.

## Materials and Methods

**Bacterial strains.** All strains used in this study were defined as *H. influenzae* using current clinical laboratory criteria. *H. influenzae* isolates, maintained with minimal passage, were stored in skim milk at -80 $^{o}$C and, for testing, were cultured on Chocolate II Agar (BD Diagnostics) at 37 $^{o}$C in the presence of 5% $CO_2$.

*H. influenzae* types a, c, and d strains were obtained from ATCC; the fully sequenced *H. influenzae* strain Rd is a lab-adapted strain that lacks the entire capsule locus; the well-studied strain Eagan (1) was the source of chemically purified type b capsule; two type b capsule deficient variant strains were previously described (15, 27); and two fully sequenced nontypeable otitis media strains, 86-028NP (a kind gift from Dr. Lauren Bakaletz, Ohio State University) and R2846 (a kind gift from Dr. Arnold Smith, University of Washington) lack capsule genes. The capsule status of these strains was considered to be well defined (see Table 2.2A). A *H. haemolyticus* strain served as a non-*H. influenzae* control.

*H. influenzae* strains used for the PCR and microarray based analyses were chosen from our large collection of *H. influenzae* clinical and commensal isolates collected over 30 years from Michigan, California, Kentucky, Pittsburgh, Minnesota, and Georgia (a kind gift of Dr. Monica Farley of the Atlanta VA Hospital) and included 40 *H. influenzae* strains previously found to be serotype positive by clinical microbiology laboratories (see Table 2.2B). Also included were throat isolates from healthy children attending day care (9, 39), middle ear isolates of children with acute otitis media from Finland (a kind gift of Dr. Tehri Kilpi) and Israel (a kind gift of Dr. Ron Dagan), all considered likely to be NTHi based on site of isolation, as well as invasive blood or cerebrospinal fluid isolates that failed to react with the type b serotyping serum in the clinical microbiology laboratory (see Table 2.2C).

**PCR based studies.** The design of PCR primers targeting a 567 or 760 base-pair (bp) region of *H. influenzae bexB* (Table 2.1) was based on sequence data available in GenBank. Primers targeting *bexA*, *pepN* (which served as a DNA positive control), and serotype-specific capsule genes have been previously described (Table 2.1) (6, 8, 28). Whole cell lysate, *i.e.*, crude lysate, served as the DNA template source for PCR reactions and was prepared by suspending several bacterial colonies, collected with a sterile cotton swab, from a fresh plate into Tris-EDTA

(TE, pH 7.5), heating for 10 min at 100 $^{o}$C in a thermocycler, centrifuging for 10 min, and removing the supernatant, which was stored at – 20 $^{o}$C until used for PCR. For routine screening, PCR was performed with *Taq* polymerase (New England Biolabs, Ipswich, MA, cat. no. M0267). Each 20 µl PCR reaction consisted of 1X ThermoPol Reaction buffer, 0.2 mM each dNTP, 0.2 µM each primer, 2.0 U *Taq*, and 1 µl crude lysate (template). Reactions were subjected to an initial denaturation step of 2 minutes at 95 $^{o}$C followed by 30 amplification cycles of 95 $^{o}$C for 30 seconds, 54 $^{o}$C for 30 seconds and 72 $^{o}$C for 45 seconds. All *bexB* positive isolates were subjected to further analysis with region II serotype-specific primers (Table 2.1). Amplicons were separated on 1.5% agarose gels in 1X Tris-Acetate-EDTA (TAE) buffer and visualized with ethidium bromide staining.

**Microarray studies.** *H. influenzae* library-on-a-slide (LOS) microarrays, consisting of *H. influenzae* genomic DNA arrayed on a glass slide, were prepared, tested, and analyzed as previously described (36, 46). In brief, total genomic DNA from each *H. influenzae* strain was isolated and spotted in duplicate onto membrane coated slides. First, a digoxygenin labeled genomic DNA quantity control probe consisting of equal amounts of each of the seven *H. influenzae* multilocus sequence typing (MLST) genes (http://Haemophilus.mlst.net) and *pepN* was hybridized to the microarray (36). The intensity of control probe hybridization to each spot was determined using Spotfinder v.3.1.1 (http://www.tm4.org/spotfinder.html) and duplicate spots were averaged using MIDAS v.2.19 (http://www.tm4.org/midas.html) (34). The slides were stripped and then rehybridized with either a *bexA* or *bexB* fluorescein labeled probe (18). The *bexA* and *bexB* probes were "mixed probes" in that they contained equal concentrations of either *bexA* or *bexB* amplified from strains representing each of the six capsular serotypes (type a, strain ATCC9006; type b, strain AA194; type c, strain ATCC 9007; type d, strain ATCC 9008; type e, strains AA52 and M4; type f, strain F2243.7). The mixed probes were PCR amplified from genomic DNA with primers HI-1 and HI-2 (*bexA* probe) or bexB.1 primers (*bexB* probe) (Table 2.1). Signal intensities were calculated as described above and the ratio of the log transformed *bexA* or *bexB* signal to the MLST-*pepN* concentration control signal was calculated for each spot and the distribution of intensities was modeled using programs written in the statistical software "R" (http://www.r-project.org); spots were classified as positive, negative, or uncertain as described in Sandstedt, *et al.* (36).

Concordance of the *bexA* and *bexB* results between the PCR based method and the microarray method was estimated by testing 80 strains for these two genes by both PCR and by hybridization of *bexA* and *bexB* probes on microarray.

**DNA sequence analysis.** To test for the magnitude of DNA sequence identity among a relatively large number of capsule-gene containing *H. influenzae* strains, DNA sequence analysis was conducted on 45 *bexB* positive isolates by first amplifying a 760 bp internal region of *bexB* with the Expand High Fidelity System (Roche, Indianapolis, IN, cat. no. 11732641) using 1X Expand High Fidelity Buffer with 1.5 mM $MgCl_2$, 0.2 mM of each dNTP, 0.2 µM each primer (bexB.FLF and bexB.FLR); 1.3 U Expand High Fidelity Enzyme mix, and 1 µl crude lysate (template) per 25 µl reaction. The resultant amplicon was purified with the Qiagen QIAquick PCR purification kit (QIAGEN, Valencia, CA, cat. no. 28104) following the manufacturer's protocol. Sequencing was performed directly from the purified PCR product, in both directions, at the University of Michigan DNA Sequencing Core. Sequences were analyzed with DNASTAR Lasergene (v8). After removal of primers and poor quality sequences, 666 nucleotides from each strain were aligned using Clustal W (43) and then adjusted manually where necessary. Phylogenetic analyses, using all codon positions, were conducted in MEGA4 (41). Whenever two or more sequences from strains of the same capsule type were 100% identical at the nucleotide level, a single consensus sequence was used for subsequent analyses.

## Results

***bexB* PCR test results.** Both pairs of *bexB* primers, bexB.1F/bexB.1R and bexB.FLF/bexB.FLR produced amplicons of the expected size, 567 bp and 760 bp, respectively, and the *bexA* and *pepN* primers produced amplicons of the expected sizes 343 bp and 918 bp, respectively. Figure 2.1d shows the PCR products of 5 representative strains: the type b strain Eagan, from which type b capsule has been chemically isolated and characterized (1); the true nontypeable strain 86-028NP (13) which lacks the capsule locus by DNA sequence analysis and thus is *bexA* and *bexB* negative; Strain Rd, initially characterized as a type d strain, which lacks the entire capsule locus due to a deletion (11) by DNA sequence analysis and thus is *bexA*-, *bexB*-, and *cap* d-negative; and two previously described strains that are *cap* region-positive by gene analysis, but serotype negative (*i.e.*, capsule deficient variants) (15, 27), and, as expected, are *bexA*-negative, *bexB*-positive, and contained their respective *cap* specific genes.

***bexB* PCR validation results**. Table 2.2a shows the results of the study to validate the *bexB* PCR method using well defined *H. influenzae* strains that were known to be capsule positive or negative by gene analysis or capsule isolation, or were obtained as serotypeable strains from ATCC. As expected, all four serotypeable strains (the ATCC types a, c, and d strains and type b Eagan) were *bexA*- and *bexB*-positive. The two strains previously known to be capsule deficient variants (15, 27) were, as expected, *bexA*-negative and *bexB*-positive, while the four strains known by genomic sequence analysis to lack the *cap* region were *bexA*- and *bexB*-negative. The *H. haemolyticus* strain was *bexA*- and *bexB*-negative.

**Correlation between *bexB* PCR and serotyping.** Among the 40 strains previously serotyped in clinical microbiology laboratories (Table 2.2b), 35 were *bexA*- and *bexB*-positive, consistent with the serotyping results performed at the time of isolation. Five were *bexA*- and *bexB*-negative and lacked region II capsule specific genes, strongly suggesting that these strains are in fact true NTHi lacking the entire capsule locus. This finding is consistent with other reports of false positive serotyping of *H. influenzae* (2, 24, 31, 32), even when tested in reference laboratories.

***bexB* gene among strains likely to be NTHi.** To identify the prevalence of capsule deficient variants among *H. influenzae* populations in which NTHi predominate, a total of 94 *H.*

*influenzae* strains likely to be non-typeable (Table 2.2c) were tested by PCR with *bexA, bexB,* and region II capsule-specific PCR primers. Among 23 commensal *H. influenzae* from the throats of healthy children, 19 were $bexA^-B^-$ (true NTHi), three were $bexA^+B^+$ and PCR positive for the type e-specific region II, and one was $bexA^+B^+$ and PCR positive for the type f-specific region II. Among 55 otitis media isolates, 50 were $bexA^-B^-$ (true NTHi), four were $bexA^+B^+$ and PCR positive for the type f-specific region II, and one was $bexA^-B^+$, a type f capsule deficient variant because it was PCR positive for the type f-specific region II but is unable to export capsule because of the *bexA* deletion. All 16 invasive blood or cerebrospinal isolates found to be serotype b negative in the clinical microbiology laboratories were $bexA^-B^-$, confirming that they were true NTHi (Table 2.2c).

***bexA* and *bexB* microarray results**. To assess the ability of *bexB* probe to hybridize with *H. influenzae* genomic DNA and to expand the capsule characterization of a large number of strains using a high through-put assay, a total of 546 *H. influenzae* strains were screened by whole genome DNA microarray for hybridization with *bexA* and *bexB* (Table 2.3). Only ten of the 546 (1.8%) strains tested by microarray hybridization showed indeterminate results due to technical difficulties, such as, damage to the array surface, too little genomic DNA, etc. Among all strains tested, no $bexA^+B^-$ strains were identified. Of 135 strains originally serotyped as type a – f, two were misclassified by microarray-- a $bexA^+bexB^+$ commensal strain that lacked serotype-specific genes and a $bexA^-bexB^-$ invasive strain that possessed a serotype b-specific gene, that is, a genetically type b capsule-deficient variant, and thus would have been misclassified as NTHi by *bexA* PCR or hybridization alone or by serotyping.

***bexA* and *bexB* PCR-microarray concordance**. The concordance between *bexA* and *bexB* detection by microarray hybridization using gene specific probes and by PCR using gene specific primers was assessed using 80 strains that were analyzed by both methods. One strain was discordant for *bexA* (98.7% concordance) and two strains were discordant for *bexB* (97.5% concordance), one of which was the *bexA* discordant strain, suggesting a possible strain mix-up (Table 2.4).

***bexB* sequencing results and phylogenetic analyses**. To gain insight into *bexB* nucleotide diversity across all capsule types, a 760 base-pair (bp) region of *bexB* was sequenced, in both directions, from 45 *bexB*+ strains as tested by PCR. Following sequence clean-up, a total

of 666 *bexB* bps from each strain, with no gaps, were available for analysis (GenBank accession HQ699679-HQ699723). All 23 type b *bexB* sequences were 100% identical at the nucleotide level. There were three unique type f sequences (type f.1, f.2, and f.3), with three variable sites among 666 bps (0.45%) and three unique type e sequences (type e.1, e.2, e.3) with five variable sites among 666 bps (0.75%). Nucleotide sequence similarities and genetic distances were estimated from a 666 bp region of *bexB* from 45 *bexB*-positive strains (Table 2.5). Pairwise nucleotide similarity ranged from 84.2 – 100% and averaged 91.7% across all sequences. Evolutionary distances, estimated using the maximum composite likelihood method (41, 42), and percent nucleotide similarity were estimated for all unique sequences. Maximum composite likelihood distances ranged from 0 to 0.199 base substitutions per site with an overall average distance across sequences of 0.101.

To address the potential for nucleotide diversity within the bexB.1F/bexB.1R primer annealing regions, the nucleotide sequences within the primer annealing regions were analyzed. Four and two substitutions were identified in the bexB.1F and bexB.1R primers, respectively. With the exception of the forward primer (bexB.1F) and type a *bexB*, none of the substitutions were localized to the 3' end of the primer (Figure 2.2).

Maximum parsimony phylogenetic analysis produced 45 equally most parsimonious unrooted trees (length = 148) that varied only slightly in topology. Based upon the 100% consensus tree (Figure 2.3), type e and f sequences were closely related and remained unresolved; bootstrap branch support for the type e and type f cluster was 99% (based upon 5,000 replicates). Type a *bexB* was sister to the group composed of type e and f *bexB* sequences and the branch support for the type a, e, and f grouping was 99%. Type c and d *bexB* sequences were 100% identical to each other the nucleotide level.

## Discussion

Current molecular capsule typing methods rely, in part, on *bexA*, which encodes a protein important in capsule exportation, as a marker for the *cap* locus; these methods are reliable for the many strains that possess two copies of the *cap* locus—one complete copy and a second copy with a deletion in *bexA*. Some *H. influenzae* strains, however, contain a single *cap* locus with a partially deleted *bexA* (Figure 1c) that is not detected with currently used *bexA* primers. Accurate capsule locus status of these strains requires seven PCR reactions for each strain, one for *bexA* and when negative, then testing for each *cap*-specific region II genes. We have demonstrated that *bexB*, which is located adjacent to *bexA* in region I of the capsule locus and encodes another protein important in capsule exportation, is a more reliable marker of the capsule locus because it can be detected in *H. influenzae* strains that possess a single *cap* locus and a *bexA* mutation in that locus. Importantly, none of the 76 *cap* region positive strains in this study were *bexA*$^{+}$*bexB*-, demonstrating a high concordance between bexB positivity and presence of type-specific cap region genes. The ability to identify true NTHi strains using a single *bexB* PCR reaction, as demonstrated in this study, is especially useful for laboratories specifically interested in capsule characterization of large *H. influenzae* strain collections.

The utility of *bexB* as a capsule type-independent marker for the capsule locus was confirmed in this study using 40 *H. influenzae* strains previously serotyped in clinical microbiology laboratories. We identified strains possessing *cap* region genes, regardless of capsule type or *bexA* status, based upon the presence of a single *bexB* amplicon. The capsule type of each *bexB*+ strain was then determined by *cap* specific PCR. Two previously described type b capsule deficient variants, Aar117 and Aar64 (15, 27), were correctly identified as *bexA*-negative/*bexB*-positive in this study (Figure 2.1d; Table 2.2a). Previous studies (2, 24, 31, 32), have clearly documented the poor sensitivity of serotyping to distinguish *cap* region positive from *cap* region negative *H. influenzae*, even in reference laboratories. The most common serotyping error appears to be positive serotype reactions in strains lacking the *cap* locus. We identified five *bexA*$^{-}$*bexB*$^{-}$ strains that were previously serotype positive in the clinical microbiology laboratory. While a portion of the *cap* locus of these strains could conceivably have been deleted subsequent to serotype analysis in the clinical microbiology laboratory, we feel this is unlikely given that these strains were minimally passaged. Furthermore, these strains

also lacked region II capsule-specific genes as assessed by PCR and, thus, would be unable to express capsule. Thus, we concluded that these five strains (Table 2.2b) reflect false positive serotyping results.

The *bexB* PCR-based test identified an otitis media type f-capsule deficient variant that lacked *bexA* and would not have been identified using only *bexA* primers for the PCR reaction (Table 2.2b). In addition, the *bexB* microarray study (Table 2.2c) identified an invasive type b-capsule deficient variant that lacked *bexA*. Similarly, Cerquetti et al. (5) described four of 41 non-type b invasive strains that failed to generate an amplicon using *bexA* primers but contained type b specific regions. Similarly, Nelson and Smith (31), using a multiplex PCR to amplify *bexA* and the *cap* specific gene regions from *H. influenzae* invasive isolates, identified three strains with absence of *bexA* but presence of *cap* b specific regions. While such capsule deficient variants were rare, their identification is important to fully characterize disease-causing or commensal *H. influenzae* strains.

PCR based screening methods dependent upon "universal" primers may not yield an amplicon if nucleotide polymorphisms, particularly those localized to the 3' primer annealing regions, preclude adequate annealing of the primers to the target gene region. Indeed, sequence diversity of *bexA* has been shown to compromise its detection by PCR (35, 47). To assess this potential in *bexB*, a 666 base-pair internal region of *bexB*, which encompassed the entire bexB.1F and bexB.1R primer annealing sites, of 45 *bexB*$^{+}$ strains were PCR amplified and their sequences analyzed. Nucleotide substitutions were observed within the primer annealing regions across serotypes, but, with the exception of the type a *bexB*, none of the substitutions are localized to the 3' end of the primer (Figure 2.2). Thus we concluded that given our current knowledge of *bexB* sequence diversity, the bexB.1F-R primer pair used in this study will amplify *bexB* across all six capsule types.

For testing large numbers of *H. influenzae* strains, high-throughput hybridization based methods present an attractive and efficient alternative to PCR screening. Furthermore, probe-based hybridization methods, such as dot blots and microarrays, have the potential to improve sensitivity of the test because they rely on a much larger region of probe-target complementarity than do PCR based approaches (567 *versus* 44 base-pairs in the case of *bexB*) and are thus are less sensitive to sequence variation. In theory, hybridization-based approaches can detect

sequences that vary by as much as 15% (L. Zhang, personal communication). To increase the probability of detecting *bexB*-positive strains across all serotypes, we employed a mixed probe composed of *bexB* sequences representing each of the six capsule types (a to f). Using the mixed *bexB* probe, we demonstrated the utility and accuracy (97.5% *bexB* concordance with detection by PCR) of this method by screening a large collection of *H. influenzae* isolates, simultaneously, using a genomic DNA microarray hybridization technique.

DNA sequence analyses revealed two distinct *bexB* sequence groups based upon nucleotide identity, one containing type c, d, and b strains and the other type e and f; this grouping is similar to that of *bexA* (35, 47). Our finding of 100% nucleotide identity in *bexB* of type b strains is consistent with the 100% identity of *bexA* among type b strains reported by Zhou and Sam (35, 47). In contrast to the 100% nucleotide identity of *bexA* sequences among both type e and type f strains reported previously (35, 47), we identified a small amount of nucleotide diversity among *bexB* sequences of both type e and type f strains, 5/666 (0.75%) and 3/666 (0.45%), respectively.

While *bexB* appears to be a reliable marker for the capsule locus, this method is not without limitations. The sequence diversity within the *bexB* primer annealing sites across all six serotypes could compromise its use in PCR assays. We noted, however, that all of the 28 strains that hybridized the mixed *bexB* probe on the microarray assay also amplified using the bexB.1 *bexB* PCR primers (Table 2.4), suggesting that the *bexB* primers we used were sufficient to amplify most, if not all, *H. influenzae bexB* sequences. A second limitation is that targeting *bexB* alone cannot differentiate true NTHi strains from strains, such as Rd (21), from which the entire capsule locus has been deleted. The magnitude of this limitation is difficult to assess, as the frequency of naturally occurring, complete *cap*-locus deletions is simply not known. This limitation is, however, not unique to the approach presented here, as none of the available molecular typing methods can detect complete capsule deletions.

In conclusion, we have described a *bexB*-targeted molecular strategy that rapidly and reliably distinguishes true non-typeable *H. influenzae* (*i.e.*, those lacking the capsule locus) from those containing a complete or partial *cap*-locus (Figure 2.4). Ongoing surveillance for *H. influenzae* is necessary for assessing the organisms associated with *H. influenzae* type b (Hib) vaccine failure and for tracking changes in disease patterns. The ability to accurately classify

strains, while challenging, is essential for any successful *H. influenzae* surveillance program, and incorporating *bexB*-based methods into molecular typing schemes can simplify strain characterization while simultaneously improving the ability to detect genetically typeable *H. influenzae* (i.e., improved sensitivity). Thus *bexB*, used as a stand-alone marker for the *cap*-locus, provides a valuable addition to the armamentarium of *H. influenzae* molecular characterization methods aimed at differentiating genetically typeable strains from true NTHi.

## Acknowledgements

We would like to thank Sarah Sartola and Monica Farley for providing type a strains.

This work was supported with funding from the Interdisciplinary Training Program in Infectious Disease (TA32AI049816), the Molecular Mechanisms of Microbial Pathogenesis Training Program (AI007528) and the National Institute of Health (R01-DC05840 and R01-AI125630).

Table 2.1. Polymerase chain reaction primers used in this study.

| Target | Primer | Primer Sequence (5' to 3') | Expected Amplicon Size (bp) | Reference |
|---|---|---|---|---|
| *bexB* | bexB.1F<br>bexB.1R | GGTGATTAACGCGTTGCTTATGCG<br>TTGTGCCTGTGCTGGAAGGTTATG | 567 | This study |
| *bexB* | bexB.FLF<br>bexB.FLR | TCATTGTGGCTCAACTCCTTTACT<br>AGCTATTCAAGGACGGGTGATTAACGC | 760 | This study |
| *bexA* | HI-1<br>HI-2 | CGTTTGTATGATGTTGATCCAGAC<br>TGTCCATGTCTTCAAAATGATG | 343 | Falla, et al., 1994 (8) |
| *pepN* | pepN_F<br>pepN_R | GATGGTCGCCATTGGGTGG<br>GATCTGCGGTTGGCGGTGTGG | 918 | Ecevit, et al., 2004 (6) |

Table 2.2. Results of *bexA* and *bexB* detection by PCR among *H. influenzae* strains. (a) *bexA* and *bexB* detection among strains with known capsule status. (b) *bexA* and *bexB* detection among *H. influenzae* strains previously serotyped in clinical microbiology laboratories. (c) *bexA* and *bexB* detection among *H. influenzae* isolates likely to be NTHi based on site of isolation or on negative type b serotyping by clinical microbiology laboratories.

a. Validation study

| Strain | type | *bexA*[*] | *bexB*[*] |
|---|---|---|---|
| ATCC 9006 | a | + | + |
| Eagan† | b | + | + |
| ATCC 9007 | c | + | + |
| ATCC 9008 | d | + | + |
| Rd (KW20)[#] | *cap*-minus | 0 | 0 |
| AAr117[|] | b-minus | 0 | + |
| AAr64[|] | b-minus | 0 | + |
| 86-028NP[#] | NTHi | 0 | 0 |
| R2846[#] | NTHi | 0 | 0 |
| R2866[#] | NTHi | 0 | 0 |
| *H. haemolyticus* | non-*H. influenzae* | 0 | 0 |

b. Strains previously serotyped in clinical microbiology laboratories

| Serotype | (n) | *bexA*$^{+}$*B*$^{+}$ | *bexA*$^{-}$*B*$^{+}$ | *bexA*$^{+}$*B*$^{-}$ | *bexA*$^{-}$*B*$^{-}$ |
|---|---|---|---|---|---|
| type a | (7) | 5 | 0 | 0 | 2 |
| type b | (20) | 20 | 0 | 0 | 0 |
| type c | (1) | 1 | 0 | 0 | 0 |
| type d | (3) | 1 | 0 | 0 | 2 |
| type e | (4) | 4 | 0 | 0 | 0 |
| type f | (5) | 4 | 0 | 0 | 1 |

c. Strains likely to be NTHi based on site of isolation or negative type b serotyping

| Strain source | (n) | *bexA*$^{+}$*B*$^{+}$ | *bexA*$^{-}$*B*$^{+}$ | *bexA*$^{+}$*B*$^{-}$ | *bexA*$^{-}$*B*$^{-}$ | *cap* type[&] (n) |
|---|---|---|---|---|---|---|
| Commensal | (23) | 4 | 0 | 0 | 19 | type e (3), type f (1) |
| Otitis media | (55) | 4 | 1 | 0 | 50 | type f (5) |
| Invasive[$] | (16) | 0 | 0 | 0 | 16 | |

[*] + = present, 0 = absent
† type b capsule isolated and chemically characterized
[#] complete genomic sequences available
[∫] lacks cap region by gene analysis, as previously reported (15, 27)
[$] serotype non-b
[&]determined by capsule specific PCR

Table 2.3. Results of the *bexA* and *bexB* microarray hybridization study.

| | *bexA*$^{+}$B$^{+}$ | *bex*A$^{-}$B$^{-}$ | *bex*A$^{-}$B$^{+}$ | Indeterminate[*] |
|---|---|---|---|---|
| Commensal, n = 322 | 15 (10 e, 4 f, 1 NTHi)[**] | 300 | 0 | 7 (7 NTHi)[**] |
| Otitis media, n = 107 | 4 (4 f)[**] | 100 | 0 | 3 (3 NTHi)[**] |
| Invasive strains, n = 119 | 115 | 1 (b)[**] | 1 (b)[**] | 0 |

* Indeterminate results on microarray with *bexA* and *bexB* probes
** *cap* type by PCR

Table 2.4. Concordance of *bexA* and *bexB* by PCR and microarray hybridization.

| ***bexA*** | Microarray positive | Microarray negative | Total |
|---|---|---|---|
| PCR positive | 28 | 1 | 29 |
| PCR negative | 0 | 51 | 51 |
| Total | 28 | 52 | 80 |
| ***bexB*** | | | |
| PCR positive | 28 | 2 | 30 |
| PCR negative | 0 | 50 | 50 |
| Total | 28 | 52 | 80 |

Table 2.5. Pairwise percent nucleotide identities and maximum composite likelihood evolutionary distances among a 666 bp region of *H. influenzae bexB*. Percent nucleotide identities are listed above the diagonal and maximum composite likelihood distances (41, 42) are listed below the diagonal. Of the 45 *bexB* sequences analyzed, ten were unique.

| | type a | type b | type c | type d | type e.1 | type e.2 | type e.3 | type f.1 | type f.2 | type f.3 |
|---|---|---|---|---|---|---|---|---|---|---|
| type a | | 88.1 | 86.2 | 87.5 | 90.7 | 90.2 | 90.5 | 90.4 | 90.2 | 90.7 |
| type b | 0.141 | | 96.7 | 98.8 | 86.2 | 85.7 | 86.0 | 85.9 | 85.7 | 86.2 |
| type c | 0.170 | 0.035 | | 97.6 | 84.7 | 84.2 | 84.5 | 84.4 | 84.2 | 84.7 |
| type d | 0.150 | 0.012 | 0.025 | | 85.6 | 85.1 | 85.4 | 85.3 | 85.1 | 85.6 |
| type e.1 | 0.107 | 0.170 | 0.192 | 0.178 | | 99.4 | 99.8 | 99.7 | 99.5 | 100 |
| type e.2 | 0.113 | 0.177 | 0.199 | 0.186 | 0.006 | | 99.3 | 99.7 | 99.6 | 99.4 |
| type e.3 | 0.109 | 0.172 | 0.194 | 0.180 | 0.002 | 0.008 | | 99.5 | 99.4 | 99.8 |
| type f.1 | 0.111 | 0.175 | 0.197 | 0.183 | 0.003 | 0.003 | 0.005 | | 99.8 | 99.7 |
| type f.2 | 0.113 | 0.177 | 0.199 | 0.185 | 0.005 | 0.005 | 0.006 | 0.002 | | 99.5 |
| type f.3 | 0.107 | 0.170 | 0.192 | 0.178 | 0.000 | 0.006 | 0.002 | 0.003 | 0.005 | |

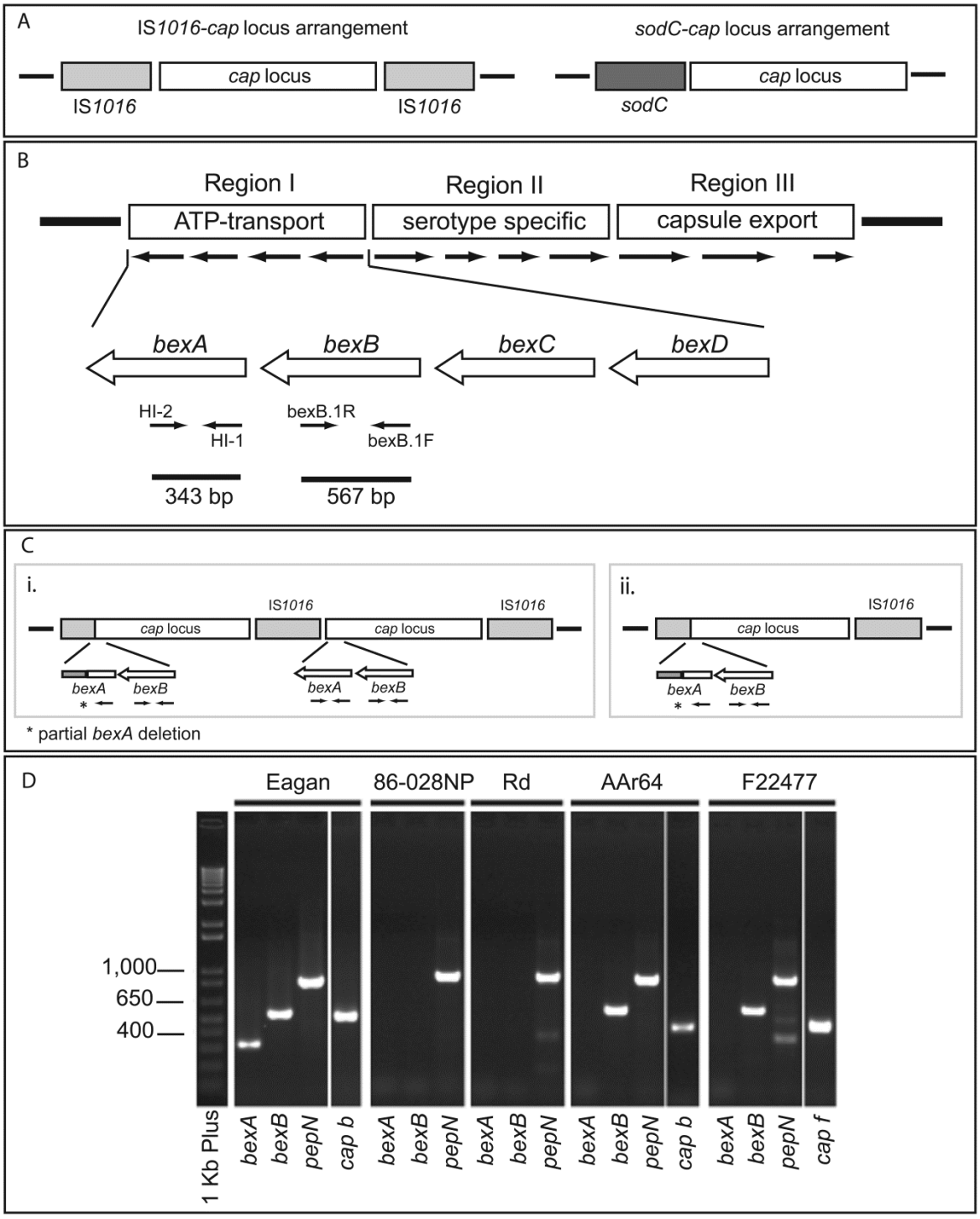

A
IS1016-cap locus arrangement
cap locus
IS1016
IS1016
sodC-cap locus arrangement
cap locus
sodC
B
Region I
Region II
Region III
ATP-transport
serotype specific
capsule export
bexA
bexB
bexC
bexD
HI-2
HI-1
bexB.1R
bexB.1F
343 bp
567 bp
C
i.
IS1016
IS1016
cap locus
cap locus
bexA
bexB
bexA
bexB
*
ii.
IS1016
cap locus
bexA
bexB
*
* partial bexA deletion
D
Eagan
86-028NP
Rd
AAr64
F22477
1,000
650
400
1 Kb Plus
bexA
bexB
pepN
cap b
bexA
bexB
pepN
bexA
bexB
pepN
bexA
bexB
pepN
cap b
bexA
bexB
pepN
cap f

Figure 2.1. Schematic representations of the capsule locus. (A) The capsule locus can be categorized based upon its location in the *H. influenzae* genome relative to either an IS1016 element or *sodC*. In one arrangement the *cap*-locus is flanked by IS1016 insertion element sequences. In the second arrangement, *sodC* is upstream of the *cap*-locus. (B) The *cap*-locus is divided into three regions, region I, II, and III; both *bexA* and *bexB* are located within region I. Previously described *bexA* specific primers produce a 343 bp amplicon (8) and the *bexB* primers used in this study amplify a 567 bp product. (C) Representative *cap*-locus arrangements. (i) A duplicated IS1016-*cap* locus harboring a *bexA* partial deletion in one copy. (ii) A single copy IS1016-*cap* arrangement. (*partial *bexA* deletion, precludes annealing with the standard *bexA* 5' primer). (D) Representative PCR results. Eagan is a type b strain, containing intact *bexA, bexB,* and the serotype b specific *cap* gene; 86-028NP is a NTHi strain lacking *bexA* and *bexB*; Strain Rd is a type d strain in which the entire *cap* region has been deleted; Aar64 is a *bexA*- type b capsule deficient variant; and, F22477 is a *bexA*- type f capsule deficient variant strain. Invitrogen 1 Kb Plus (Invitrogen Corporation, Carlsbad, CA, cat no. 10787-026) served as the DNA ladder.

```
bexB.1F      5′ GGTGATTAACGCGTTGCTTATGCG 3′
type_a             ............A..A........
type_b             ........................
type_c             ............A...........
type_d             ............A...........
type_f.1           ............A..A........
type_f.2           ............A..A........
type_f.3           ............A..A........
type_e.1           ............A..A........
type_e.2           ............A..A........
type_e.3           ............A..A........


bexB.1R      5′ TTGTGCCTGTGCTGGAAGGTTATC 3′
type_a             ...A.......A.........G..
type_b             ........................
type_c             ........................
type_d             ........................
type_f.1           ...A.......A......A.....
type_f.2           ...A.......A......A.....
type_f.3           ...A.......A......A.....
type_e.1           ...A.......A......A.....
type_e.2           ...A.......A......A.....
type_e.3           ...A.......A......A.....
```

Figure 2.2. Substitutions within *bexB* primer annealing regions. The top row of each sequence set represents the actual PCR primer used (Table 2.1); dots represent agreement with the primer sequence. Primers were designed using publicly available type b nucleotide sequences. The bexB.1R and bexB.1F primer set amplified *bexB* from all six capsular types. There is a nucleotide substitution localized to the 3'-end of the forward primer (bexB_1R) relative to type a *bexB*, however, there were no substitutions in the 3'-most region of either primer.

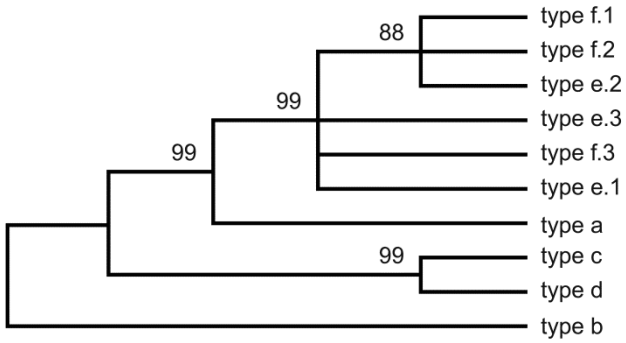



Figure 2.3. *bexB* evolutionary history inferred using the Maximum Parsimony method (7). The 100% consensus tree from the 45 most parsimonious trees (tree length = 148) is shown. The consistency index is 0.965217, the retention index is 0.981651, and the composite index is 0.947507 for all sites and the 98 parsimony-informative sites. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (5000 replicates) are shown next to the branches (10). A total of 666 nucleotides were included in the final dataset and 98 nucleotides were parsimony informative. Phylogenetic analyses were conducted in MEGA4 (41) using all codon positions; there were no gaps in the sequence alignment.

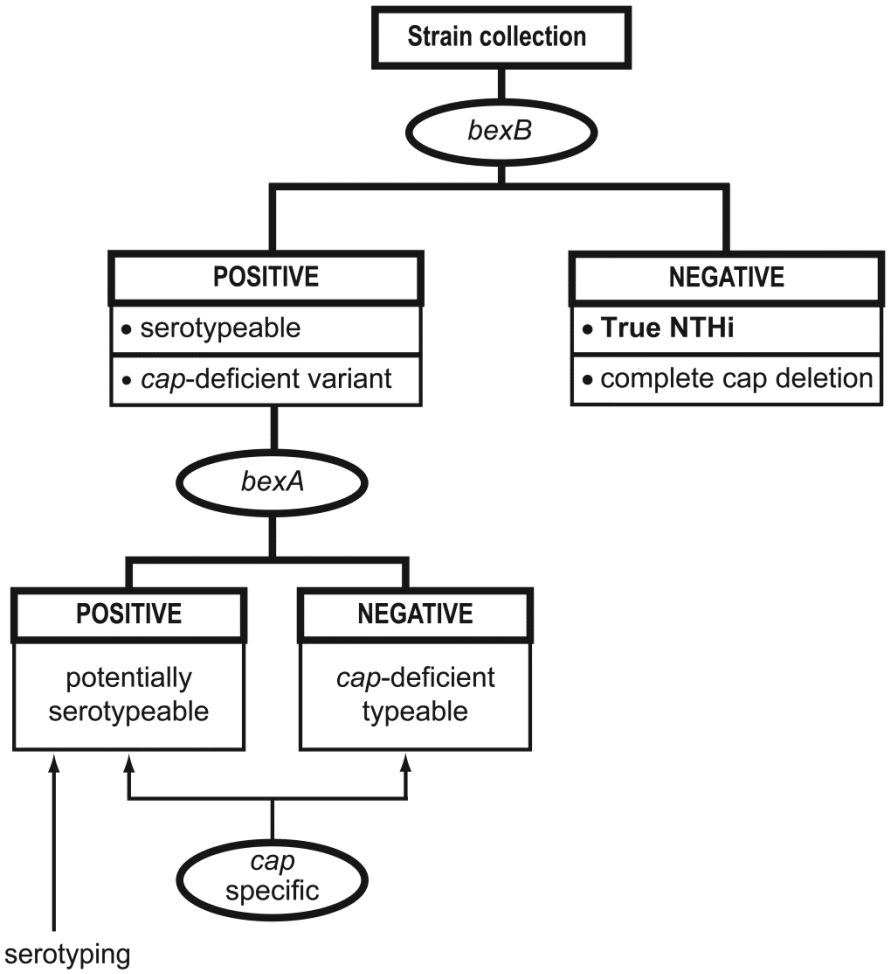



Figure 2.4. Suggested workflow for characterizing *H. influenzae* strain collections with regard to the capsule locus. Gene names listed within an oval represent detection of that gene via a PCR or probe based technique. Identification of "true" NTHi strains can be achieved by screening for the presence of *bexB;* strains lacking *bexB* represent true NTHi or typeable strains in which the entire capsule locus has been deleted. In the second step, presence of *bexA* differentiates potentially serotypeable strains (*bexA*-positive) from capsule deficient but genetically typeable strains (nonserotypeable, *bexA*-minus, *cap* locus-positive). Among *bexB*-positive strains, capsular type can be assigned by screening for capsule specific genes residing within region 2 of the capsule locus (8). Finally, if necessary, production of a functional capsule can be verified with traditional type-specific serotyping (8).

## Literature Cited


1. Anderson, P., J. Pitt, and D. H. Smith. 1976. Synthesis and release of polyribophosphate by *Haemophilus influenzae* type b in vitro. Infect Immun 13:581-9.
2. CDC. 2002. Serotyping discrepancies in *Haemophilus influenzae* type b disease, United States, 1998-1999. MMWR 51:706-707.
3. Cerquetti, M., R. Cardines, M. L. Ciofi Degli Atti, M. Giufre, A. Bella, T. Sofia, P. Mastrantonio, and M. 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 192:819-23.
4. Cerquetti, M., R. Cardines, M. Giufre, T. Sofia, F. D'Ambrosio, P. Mastrantonio, and M. L. Ciofi degli Atti. 2006. Genetic diversity of invasive strains of *Haemophilus influenzae* type b before and after introduction of the conjugate vaccine in Italy. Clin Infect Dis 43:317-9.
5. Cerquetti, M., M. L. Ciofi degli Atti, G. Renna, A. E. Tozzi, M. L. Garlaschi, and P. Mastrantonio. 2000. Characterization of non-type B *Haemophilus influenzae* strains isolated from patients with invasive disease. The HI Study Group. J Clin Microbiol 38:4649-52.
6. Ecevit, I. Z., K. W. McCrea, M. M. Pettigrew, A. Sen, C. F. Marrs, and J. R. Gilsdorf. 2004. Prevalence of the *hifBC*, *hmw1A*, *hmw2A*, *hmwC*, and hia Genes in *Haemophilus influenzae* Isolates. J Clin Microbiol 42:3065-72.
7. Eck, R., and M. O. Dayhoff. 1966. Atlas of Protein Sequence and Structure. National Biomedical Research Foundation, Silver Spring, Maryland.
8. Falla, T. J., D. W. Crook, L. N. Brophy, D. Maskell, J. S. Kroll, and E. R. Moxon. 1994. PCR for capsular typing of *Haemophilus influenzae*. J Clin Microbiol 32:2382-6.
9. Farjo, R. S., B. Foxman, M. J. Patel, L. Zhang, M. M. Pettigrew, S. I. McCoy, C. F. Marrs, and J. R. Gilsdorf. 2004. Diversity and sharing of *Haemophilus influenzae* strains colonizing healthy children attending day-care centers. Pediatr Infect Dis J 23:41-6.
10. Felsenstein, J. 1985. Confidence limits on phylogenies: An approach using the bootstrap. Evolution 39:783-791.
11. Fleischmann, R. D., M. D. Adams, O. White, R. A. Clayton, E. F. Kirkness, A. R. Kerlavage, C. J. Bult, J. F. Tomb, B. A. Dougherty, J. M. Merrick, and et al. 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science 269:496-512.
12. Giufre, M., R. Cardines, P. Mastrantonio, and M. Cerquetti. 2006. Variant IS1016 insertion elements in invasive *Haemophilus influenzae* type b isolates harboring multiple copies of the capsulation b locus. Clin Infect Dis 43:1225-6.
13. Harrison, A., D. W. Dyer, A. Gillaspy, W. C. Ray, R. Mungur, M. B. Carson, H. Zhong, J. Gipson, M. Gipson, L. S. Johnson, L. Lewis, L. O. Bakaletz, and R. S. Munson, Jr. 2005. Genomic sequence of an otitis media isolate of nontypeable *Haemophilus influenzae*: comparative study with *H. influenzae* serotype d, strain KW20. J Bacteriol 187:4627-36.
14. Hoiseth, S. K., C. J. Connelly, and E. R. Moxon. 1985. Genetics of spontaneous, high-frequency loss of b capsule expression in *Haemophilus influenzae*. Infect Immun 49:389-95.

15. Hoiseth, S. K., and J. R. Gilsdorf. 1988. The relationship between type b and nontypable *Haemophilus influenzae* isolated from the same patient. J Infect Dis 158:643-5.
16. Hoiseth, S. K., E. R. Moxon, and R. P. Silver. 1986. Genes involved in *Haemophilus influenzae* type b capsule expression are part of an 18-kilobase tandem duplication. Proc Natl Acad Sci U S A 83:1106-10.
17. Kapogiannis, B. G., S. Satola, H. L. Keyserling, and M. M. Farley. 2005. Invasive infections with *Haemophilus influenzae* serotype a containing an IS1016-bexA partial deletion: possible association with virulence. Clin Infect Dis 41:e97-103.
18. Kong, F., M. Brown, A. Sabananthan, X. Zeng, and G. L. Gilbert. 2006. Multiplex PCR-based reverse line blot hybridization assay to identify 23 Streptococcus pneumoniae polysaccharide vaccine serotypes. J Clin Microbiol 44:1887-91.
19. Kroll, J. S., I. Hopkins, and E. R. Moxon. 1988. Capsule loss in *H. influenzae* type b occurs by recombination-mediated disruption of a gene essential for polysaccharide export. Cell 53:347-56.
20. 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. Mol Microbiol 4:1853-62.
21. Kroll, J. S., B. M. Loynds, and E. R. Moxon. 1991. The *Haemophilus influenzae* capsulation gene cluster: a compound transposon. Mol Microbiol 5:1549-60.
22. Kroll, J. S., and E. R. Moxon. 1988. Capsulation and gene copy number at the cap locus of *Haemophilus influenzae* type b. J Bacteriol 170:859-64.
23. Kroll, J. S., S. Zamze, B. Loynds, and E. R. Moxon. 1989. Common organization of chromosomal loci for production of different capsular polysaccharides in *Haemophilus influenzae*. J Bacteriol 171:3343-7.
24. LaClaire, L. L., M. L. Tondella, D. S. Beall, C. A. Noble, P. L. Raghunathan, N. E. Rosenstein, and T. Popovic. 2003. Identification of *Haemophilus influenzae* serotypes by standard slide agglutination serotyping and PCR-based capsule typing. J Clin Microbiol 41:393-6.
25. Lacross, N. C., C. F. Marrs, M. Patel, S. A. Sandstedt, and J. R. Gilsdorf. 2008. High genetic diversity of nontypeable *Haemophilus influenzae* isolates from two children attending a day care center. J Clin Microbiol 46:3817-21.
26. Maaroufi, Y., J. M. De Bruyne, C. Heymans, and F. Crokaert. 2007. Real-time PCR for determining capsular serotypes of *Haemophilus influenzae*. J Clin Microbiol 45:2305-8.
27. McCrea, K. W., J. L. Sauver, C. F. Marrs, D. Clemans, and J. R. Gilsdorf. 1998. Immunologic and structural relationships of the minor pilus subunits among *Haemophilus influenzae* isolates. Infect Immun 66:4788-96.
28. McCrea, K. W., M. L. Wang, J. Xie, S. A. Sandstedt, G. S. Davis, J. H. Lee, C. F. Marrs, and J. R. Gilsdorf. 2010. Prevalence of the *sodC* Gene in Nontypeable *Haemophilus influenzae* and *Haemophilus haemolyticus* by Microarray-Based Hybridization. J Clin Microbiol 48:714-9.
29. Mukundan, D., Z. Ecevit, M. Patel, C. F. Marrs, and J. R. Gilsdorf. 2007. Pharyngeal colonization dynamics of *Haemophilus influenzae* and *Haemophilus haemolyticus* in healthy adult carriers. J Clin Microbiol 45:3207-17.
30. Murphy, T. F., and M. A. Apicella. 1987. Nontypable *Haemophilus influenzae*: a review of clinical aspects, surface antigens, and the human immune response to infection. Rev Infect Dis 9:1-15.

31. Nelson, K. L., and A. L. Smith. 2009. Determination of capsulation status in *Haemophilus influenzae* by multiplex polymerase chain reaction. Diagn Microbiol Infect Dis.
32. Ogilvie, C., A. Omikunle, Y. Wang, I. J. St Geme, 3rd, C. A. Rodriguez, and E. E. Adderson. 2001. Capsulation loci of non-serotype b encapsulated *Haemophilus influenzae*. J Infect Dis 184:144-9.
33. Pittman, M. 1931. Variation and Type Specificity in the Bacterial Species *Hemophilus Influenzae*. J Exp Med 53:471-492.
34. Saeed, A. I., V. Sharov, J. White, J. Li, W. Liang, N. Bhagabati, J. Braisted, M. Klapa, T. Currier, M. Thiagarajan, A. Sturn, M. Snuffin, A. Rezantsev, D. Popov, A. Ryltsov, E. Kostukovich, I. Borisovsky, Z. Liu, A. Vinsavich, V. Trush, and J. Quackenbush. 2003. TM4: a free, open-source system for microarray data management and analysis. Biotechniques 34:374-8.
35. Sam, I. C., and M. Smith. 2005. Failure to detect capsule gene *bexA* in *Haemophilus influenzae* types e and f by real-time PCR due to sequence variation within probe binding sites. J Med Microbiol 54:453-5.
36. Sandstedt, S. A., L. Zhang, M. Patel, K. W. McCrea, Z. Qin, C. F. Marrs, and J. R. Gilsdorf. 2008. Comparison of laboratory-based and phylogenetic methods to distinguish between *Haemophilus influenzae* and *H. haemolyticus*. J Microbiol Methods 75:369-71.
37. Satola, S. W., P. L. Schirmer, and M. M. Farley. 2003. Complete sequence of the cap locus of *Haemophilus influenzae* serotype b and nonencapsulated b capsule-negative variants. Infect Immun 71:3639-44.
38. Satola, S. W., P. L. Schirmer, and M. M. Farley. 2003. Genetic analysis of the capsule locus of *Haemophilus influenzae* serotype f. Infect Immun 71:7202-7.
39. St Sauver, J., C. F. Marrs, B. Foxman, P. Somsel, R. Madera, and J. R. Gilsdorf. 2000. Risk factors for otitis media and carriage of multiple strains of *Haemophilus influenzae* and *Streptococcus pneumoniae*. Emerg Infect Dis 6:622-30.
40. Sukupolvi-Petty, S., S. Grass, and J. W. St Geme, 3rd. 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. J Bacteriol 188:3870-7.
41. Tamura, K., J. Dudley, M. Nei, and S. Kumar. 2007. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol Biol Evol 24:1596-9.
42. Tamura, K., M. Nei, and S. Kumar. 2004. Prospects for inferring very large phylogenies by using the neighbor-joining method. Proc Natl Acad Sci U S A 101:11030-5.
43. Thompson, J. D., D. G. Higgins, and T. J. Gibson. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22:4673-80.
44. Ueno, K., J. Nishi, N. Imuta, K. Tokuda, and Y. Kawano. 2010. Presence of multiple copies of capsulation loci in invasive *Haemophilus influenzae* type b (Hib) strains in Japan before introduction of the Hib conjugate vaccine. Microbiol Immunol 54:160-3.
45. Wallace, R. J., Jr., D. M. Musher, E. J. Septimus, J. E. McGowan, Jr., F. J. Quinones, K. Wiss, P. H. Vance, and P. A. Trier. 1981. *Haemophilus influenzae* infections in adults: characterization of strains by serotypes, biotypes, and beta-lactamase production. J Infect Dis 144:101-6.

46. Zhang, L., U. Srinivasan, C. F. Marrs, D. Ghosh, J. R. Gilsdorf, and B. Foxman. 2004. Library on a slide for bacterial comparative genomics. BMC Microbiol 4:12.
47. Zhou, J., D. K. Law, M. L. Sill, and R. S. Tsang. 2007. Nucleotide sequence diversity of the *bexA* gene in serotypeable *Haemophilus influenzae* strains recovered from invasive disease patients in Canada. J Clin Microbiol 45:1996-9.

# Chapter 3

## Prevalence, distribution, and sequence diversity of *hmwA* among a diverse collection of commensal and OM NTHi strains[2]


## Abstract

Nontypeable *Haemophilus influenzae* (NTHi) are Gram-negative bacteria that colonize the human pharynx, their only known natural reservoir. Adherence to the host epithelium marks one of the first step in NTHi colonization and pathogenesis. Approximately 75% of NTHi strains encode a pair of high molecular weight (HMW) adhesins that mediate attachment to the host epithelium. HMW adhesins are highly immunogenic, antigenically diverse, and display a wide range of amino acid diversity both within and between strains. In this study, prevalence of *hmwA,* which encodes the HMW adhesin, was determined a collection of 170 NTHi strains collected from the ears (OM strains) or throats (commensal strains) of children from Finland, Israel, and the US. Overall, *hmwA* was detected in 72% of NTHi isolates and was significantly more prevalent (P = 0.038) among OM isolates than commensal isolates; the prevalence ratio comparing *hmwA* prevalence among ear strains with that of commensal strains was 1.23 (95% CI (1.01, 1.50). Among 91 *hmwA*-positive NTHi strains, 93% possessed two *hmw* loci in conserved locations on the chromosome. To extend our understanding of *hmwA* binding domain sequence diversity, we selected 33 strains for sequencing. The average amino acid identity across all *hmwA* sequences was 62%. Phylogenetic analyses of the *hmwA* binding domains revealed four distinct *hmwA* sequence clusters, and the majority of sequences (83%) belonged to one of two


[2] This work is being prepared for submission with the following authors: Davis GS, Patel M, Hammond J, Zhang, L, Marrs CF, Gilsdorf JR

sequence clusters. *hmwA* sequences did not cluster by chromosomal location, geographic region, or disease status. These results contribute to our understanding of HMW adhesin diversity among commensal and OM strains, and provide insight into the mechanisms driving the evolution of NTHi HMW adhesins.

## Introduction

Nontypeable *Haemophilus influenzae* (NTHi) are Gram-negative bacteria that colonize the human pharynx, their only known natural reservoir. Although generally considered a commensal, NTHi is capable of causing localized infections of the upper and lower respiratory tract, for example, acute otitis media, as well as more severe invasive infections. NTHi colonization is initiated by adherence of bacterial cells to the host epithelium, and since disease causing strains arise from the strains colonizing the pharynx, adherence also marks the first step in NTHi pathogenesis. NTHi produce a wide array of surface exposed adhesins that mediate attachment to epithelial cells (50-53, 56). High molecular weight (HMW) adhesins, which are expressed by approximately 75% of NTHi strains (5), are one of the predominant non-pilin NTHi adhesins. HMW adhesins are widely distributed across NTHi phylogenetic groups as defined by MLST (25) or by multi-locus enzyme electrophoresis (55).

The HMW adhesins are a family of paralogous proteins encoded by the *hmw* locus, which is present in two copies on the NTHi chromosome (9). Both the gene content and the chromosomal locations of the *hmw* loci, designated *hmw1* and *hmw2*, are conserved across strains (6, 9). Each *hmw* locus encodes three proteins, *hmwA*, *hmwB*, and *hmwC* (6). The functional HMW adhesins are encoded by *hmwA; hmwB* and *hmwC* encode proteins required for HMW adhesin maturation, glycosylation, and secretion (5, 6, 31, 32, 54). The *hmw1* and *hmw2* loci are located downstream of HI1679 and HI1598, respectively, as specified with respect to *H. influenzae* strain Rd (5).

HMW adhesins display a wide range of amino acid diversity both within and between strains (5, 9, 18, 29). The amino acid diversity is not evenly distributed throughout the mature HMW adhesin, but is localized to the region of the adhesin that interacts with the host epithelium, that is, the binding domain (5, 18, 21, 30). Amino acid diversity within the HMW binding domain likely contributes to NTHi tissue tropism. While five distinct NTHi *in vitro* adherence patterns have been reported (62), most strains analyzed to date display adherence patterns similar to that of the prototypic NTHi strain 12, with one adhesin conferring HMW1-like adherence and the other HMW2-like adherence (9, 53, 62).

Because the HMW adhesins are exposed to the external environment and directly interact with host tissues, amino acid diversity likely plays a significant role in immune evasion. NTHi outer membrane proteins, and specifically the HMW adhesins, stimulate an antibody mediated immune response in their hosts (4, 42). In fact, HMW adhesins are the immunodominant outer membrane proteins (4). The antibodies directed against the HMW adhesin, however, offer varying degrees of cross protection against heterologous NTHi strains (4, 62). That antibodies targeting the HMW adhesins are not broadly protective highlights the potential of amino acid diversity to mediate NTHi immune evasion.

All current, publicly available, *hmwA* DNA sequence data come exclusively from disease isolates (5, 9, 21, 30, 34, 62), but, disease isolates represent only a subset of the total NTHi population, much of which resides as commensal strains in the pharynges of healthy individuals. In this study, we utilized a well characterized collection of NTHi strains collected from the middle ears of children with otitis media (OM strains) or from the throats of healthy children (commensal strains) from Finland, Israel, and the US (36). We determined that the overall *hmwA* prevalence was 72% and that OM strains were significantly more likely to possess *hmwA* than were commensal strains. We then sequenced the binding domain region of $hmwA_{1679}$ and $hmwA_{1598}$ from 33 strains that, based upon prior multilocus sequence typing (MLST) analyses (36), were phylogenetically distributed and included both OM and commensal strains. *hmwA* sequence analysis revealed a wide range of genetic diversity both within and between NTHi strains from this collection and phylogenetic analyses suggested that *hmwA* binding domains form four distinct *hmwA* sequence clusters.

## Materials and Methods

**Strain collection.** The NTHi isolates used in this study have been described in detail elsewhere (36). In brief, the collection contains 170 NTHi strains isolated between 1994 – 2002 from children seven years of age or under living in Finland, Israel, or the United States. Ninety-five of the isolates were collected from the middle ears of children with acute otitis media ("OM" strains) and the remaining 75 were collected from the throats or nasopharynges of otherwise healthy children ("commensal" strains).

***hmwA* nomenclature.** The HMW adhesins were first characterized in NTHi strain 12, which is considered the prototypic HMW-producing strain (4-6). The HMW adhesins differed in molecular weight; HMW1 (encoded by $hmwA_{1679}$) migrated at approximately 125 kDa and HMW2 (encoded by $hmwA_{1598}$) migrated at approximately 120 kDa. This was the basis of the original naming system in which the genes encoding the 125 kDa and 120 kDa adhesins were named *hmw1A* and *hmw2A* (5). In a later study, Buscher, *et al.*, determined the exact chromosomal location of the *hmw1* and *hmw2* loci in NTHi strain 12 (9) and defined them with reference to the respective ORFs in the fully sequenced *H. influenzae* strain Rd (27). The NTHi strain 12 *hmw1* locus resides adjacent to Rd ORF HI1679, contains *hmw1A*, and produces the HMW1 adhesin, the strain 12 *hmw2* locus resides adjacent to Rd ORF HI1598, contains *hmw2A*, and produces the HMW2 adhesin (9).

In subsequent studies, various different criteria were used to assign names to the *hmwA* genes in newly characterized strains. Buscher, *et al.*, determined the *in vitro* adherence profiles of the HMW adhesins for several strains and categorized them as HMW1-like or HMW2-like using the NTHi strain 12 HMW adherence patterns as a reference (9). Most NTHi strains analyzed to date contain two *hmw* loci – one locus encodes a HMW1-like adhesin, which mediates adherence via interaction with α 2,3-sialic acid, and the other an HMW2-like adhesion, whose adherence interactions are not α 2,3-sialic acid and remain unknown (9, 30). While *hmw* loci chromosomal locations are conserved, the specific adhesin encoded at each locus can vary by strain, for example, *hmwA* adjacent to HI1679 in NTHi strain 12 confers HMW1-like adherence whereas in strain 5 the *hmwA* adjacent to HI1679 encode an HMW with HMW2-like adherence (9). Other groups refer to *hmwA* genes with reference to their chromosomal location in strain 12, *e.g.*, considering *hmw1A* as being adjacent to Rd ORF HI1598 regardless of its

adherence characteristics (30). This later approach is necessary for tracking specific *hmwA* genes when their HMW adherence characteristics are unknown.

In this study, we assign each *hmwA* a locus-specific designation, $hmwA_{1598}$ or $hmwA_{1679}$, that corresponds its chromosomal location relative to Rd ORF HI1598 or Rd ORF HI1679, respectively. Since we did not determine HMW adherence profiles in this study, this allows us to track each *hmwA* locus without reference to the respective proteins' adherence characteristics. With reference to NTHi strain 12, our $hmwA_{1679}$ and $hmwA_{1598}$ designations correspond to strain 12 *hmw1A* and *hmw2A*, respectively.

***hmwA*-leader PCR screen.** To determine *hmwA* prevalence, whole cell lysates of all 170 strains were interrogated with a single pair of PCR primers that annealed to a conserved region of the $hmwA_{1679}$ and $hmwA_{1598}$ signal sequences (Table 3.1). Whole cell lysate was prepared by suspending several bacterial colonies, collected with a sterile cotton swab, from a fresh plate into Tris-EDTA (TE, pH 7.5), heating for 10 min at 100 $^{o}$C in a thermocycler, centrifuging for 10 min, and removing the supernatant, which was stored at – 20 $^{o}$C until used for PCR. For routine *hmwA*-leader screening, PCR was performed with *Taq* polymerase (New England Biolabs, Ipswich, MA, cat. no. M0267). Each 20 µl PCR reaction consisted of 1X ThermoPol Reaction buffer, 0.2 mM of each dNTP, 0.2 µM of each primer, 2.0 units of *Taq* polymerase, and 2 µl crude lysate (template). Reactions were subjected to an initial denaturation step of 2 minutes at 95 $^{o}$C followed by 30 amplification cycles of 95 $^{o}$C for 30 seconds, 58 $^{o}$C for 30 seconds and 72 $^{o}$C for 1 minute. Amplicons were separated on 1.5% agarose gels in 1X Tris-Acetate-EDTA (TAE) buffer and visualized with ethidium bromide staining.

***$hmwA_{1679}$* and *$hmwA_{1598}$* PCR amplification.** Total genomic DNA was isolated for all *hmwA*-leader PCR-positive NTHi isolates and subjected to further analysis with $hmwA_{1679}$ and $hmwA_{1598}$ specific PCR primer pairs (Table 3.1). In a majority of strains tested, primers A1FL5.1 and *hmwA*core_1R successfully amplified $hmwA_{1679}$ and A2FL5.2 and *hmwA*core_1R successfully amplified $hmwA_{1598}$. The strains listed in Table 3.S1, however, required alternative primer pairs (Table 3.1) for successful *hmwA* amplification. Genomic DNA was isolated using the Wizard genomic DNA purification kit (Promega, Madison, WI., cat. no. A1120) according to the manufacturer's instructions, resuspended in 1x Tris-EDTA buffer (10 mM Tris-HCL and 1mM EDTA at pH 8), and stored at 4 $^{o}$C until needed. For *hmwA*-specific reactions, PCR was

performed with Phusion High-Fidelity DNA polymerase (New England Biolabs, Ispwich, MA, cat. no. M0530). Each 25 µl PCR assay consisted of 1X Phusion HF buffer, 125 ng of genomic DNA template, 0.2 mM of each dNTP, and 0.2 µM of each primer, and 0.5 units of Phusion polymerase. Reactions were subjected to an initial denaturation step of 2 minutes at 98 $^{o}$C followed by 30 amplification cycles of 98 $^{o}$C for 15 seconds, 65 $^{o}$C for 15 seconds and 72 $^{o}$C for 3 minutes. Amplicons were separated on 1.0% agarose gels in 1X Tris-Acetate-EDTA (TAE) buffer and visualized under UV illumination following ethidium bromide staining.

***hmwA* core binding domain amplification, cloning, and sequencing.** The $hmwA_{1679}$ and $hmwA_{1598}$ binding domain regions were sequenced from a subset of the NTHi isolates used in this study (Table 3.2). Binding domain regions were amplified separately for each strain using Phusion amplified $hmwA_{1679}$ and $hmwA_{1598}$ amplicons as DNA template. Prior to re-amplification, each primary *hmwA* amplicon was purified using a QIAquick PCR purification kit (QIAGEN, Valencia, CA, cat. no. 28104) following the manufacturer's protocol. Each *hmwA* binding domain was PCR amplified with primers that bound to a conserved region of both $hmwA_{1679}$ and $hmwA_{1598}$, hmwAseq1 and hmwAseq4 (Table 3.1), with Phusion High-Fidelity DNA polymerase (New England Biolabs, Ispwich, MA, cat. no. M0530). Each 25 µl PCR assay consisted of 1X Phusion HF buffer, 10 ng of purified *hmwA* PCR product, 0.2 mM of each dNTP, 0.2 µM of each primer, and 0.5 units of Phusion polymerase. Reactions were subjected to an initial denaturation step of 1 minute at 98 $^{o}$C followed by 30 amplification cycles of 98 $^{o}$C for 15 seconds, 65 $^{o}$C for 15 seconds and 72 $^{o}$C for 1.25 minutes. Amplicons were separated on 1.5% agarose gels in 1X Tris-Acetate-EDTA (TAE) buffer and visualized with ethidium bromide staining.

The resultant *hwmA* binding domain amplicons were cloned into Invitrogen's pCR4 sequencing vector using the Zero Blunt TOPO PCR cloning kit (Invitrogen, Carlsbad, CA, cat. no. K2875) according to the manufacturer's protocols. Plasmid was isolated from a bacterial culture, grown overnight at 37 $^{o}$C in the presence of 50 µg/ml kanamycin, with a GenCatch Plus plasmid DNA miniprep kit (Epoch Life Sciences, Sugar Land, TX; cat. no. 21-60050) following manufacturer's protocols. *hmwA* insert size was verified by restriction enzyme digestion followed by separation on 1.5% agarose gels in 1X Tris-Acetate-EDTA (TAE) buffer and visualized with ethidium bromide staining.

Sanger DNA Sequencing was performed at the University of Michigan DNA Sequencing Core. Multiple sequencing reactions were required per *hmwA* core binding domain and additional sequencing primers, 400 in total, were designed as needed. PCR primers were designed with IDT PrimerQuest (http://www.idtdna.com/Primerquest/Home/Index), sequence editing and contig assembly was performed with SeqMan Pro v8 (DNASTAR, Madison, WI).

Fifteen publicly available NTHi *hmwA* nucleotide sequences, obtained from GenBank, were also included in the analyses (Table 3.2).

***hmwA* sequence analysis.** *hmwA* sequence alignments were performed in two stages. First, *hmwA* nucleotide sequences were translated to amino acids using the computer package MEGA5 and aligned using the MUSCLE algorithm (23). Aligned amino acids were converted back to their original nucleotide sequence and neighbor joining (NJ) tree was estimated in MEGA5 (59). This first alignment was performed simply to estimate a NJ tree to serve as an initial guide tree for the following alignment steps. Using the NJ tree as a guide tree, *hmwA* nucleotide sequences were converted to amino acids, aligned with the computer program PRANKSTER, and then converted back to the original nucleotides (37). The process was then repeated, using a NJ tree estimated from the first PRANKSTER alignment, and this second alignment was used for all subsequent analyses. This two-stage alignment process is suggested when the phylogenetic relationships among sequences are uncertain.

Evolutionary analyses were conducted with the computer program MEGA5 (59). Pairwise amino acid p-distances – the proportion of amino acids that vary between two sequences – were calculated for all sequence pairs using the pairwise gap deletion option. Maximum composite likelihood distances (65) were estimated from the *hmwA* nucleotide sequences. Rate heterogeneity between was modeled using a discrete gamma distribution with five rate categories; the gamma parameter value was estimated for each *hmwA* alignment under a general time reversible substitution model (41). Gaps were deleted from nucleotide alignments prior to analyses. Standard errors of the distance measures, both amino acid and nucleotide, were estimated using a bootstrap method with 1000 replicates (26).

Sequence alignments were tested for evidence of recombination with the PHI test (7), implemented in SplitsTree4 (35). To account for the effects of recombination, relationships

between *hmwA* sequences were assessed using the Neighbor-Net (8) method implemented in SplitsTree4 (35). *hmwA* phylogenies were also estimated using a maximum likelihood approach using maximum composite likelihood distances with rate heterogeneity among sites modeled with a five category discrete gamma distribution and implemented in MEGA5 (59).

## Results

***hmwA* prevalence.** All 170 strains in the collection (36) were screened for the presence of *hmwA* by PCR using hmwA_leader PCR primers (Table 3.1) that annealed to conserved regions of the signal sequence of both $hmwA_{1598}$ and $hmwA_{1679}$; these primers did not differentiate between the two *hmwA* loci. Overall, 123 (72%) strains were *hmwA*-leader PCR positive (Fig 3.1). *hmwA* prevalence varied by geographic region, (Table 3.3) and was significantly higher among strains from Finland (85%) than from Israel (66%) (P = 0.0231) or the US (66%) (P = 0.0196).

Overall, OM strains were significantly (P = 0.038) more likely to possess *hmwA* (69%) than were commensal strains (31%) and the prevalence ratio comparing *hmwA* prevalence among ear strains with that of commensal strains was 1.23 (95% CI (1.01, 1.50); *i.e.*, OM strains were 1.23 times more likely to be *hmwA*-positive than were commensal strains. Within each geographic region, OM strains were more likely than commensal strains to contain *hmwA*, but, these differences were not statistically significant ($\alpha = 0.05$) for any region (Table 3.3).

Ninety-one *hmwA*-positive strains were interrogated by PCR assay to determine if they possessed two separate *hmw* loci adjacent to Rd ORFs HI1679 and HI1598. Eighty-five (93.4%) of the *hmwA*-leader positive strains tested contained both $hmwA_{1679}$ and $hmwA_{1598}$. One strain was PCR-negative for $hmw_{1679}$ (G1222,), two strains were PCR-negative for $hmwA_{1598}$ (G1522 and I167), and three strains lacked both *hmwA* loci (G123, G522, and F1158).

***hmwA* sequencing.** The $hmwA_{1679}$ and $hmwA_{1598}$ binding domain regions were sequenced from 33 strains selected based upon their phylogenetic relationships estimated by MLST analysis (Fig. 3.1) (36). By selecting strains widely distributed across the MLST phylogeny, it was our intention to capture a wide range of *hmwA* diversity. Of the 33 strains, 12 were from Finland, 11 from Israel, 10 from the US, and, in total, 17 were OM strains (Table 3.2).

The *hmwA* binding regions ranged in size from 2652 to 3573 nucleotides (Table 3.2) which is consistent with previous studies (9, 18, 29, 55, 62). Sequence data was obtained for both $hmwA_{1679}$ and $hmwA_{1598}$ from 26/33 of the strains in our study collection and for the remaining seven strains *hmwA* sequence was obtained for one of the two loci (Table 3.2). Five of the seven

*hmwA* sequences were excluded due to poor quality and two additional *hmwA* sequences were excluded because they encoded in-frame stop codons.

Pairwise evolutionary distances estimated from the *hmwA* nucleotide sequence data ranged from 0.00 to 1.29 base substitutions per site with an overall average genetic distance of 0.7531 (SE = 0.04327). Pairwise *hmwA* amino acid p-distances, *i.e.*, the proportion of amino acids that vary between two sequences, ranged from 0.0000 to 0.4898 and the overall average pairwise amino acid distance was 0.3838 (SE = 0.0093). This means that, on average, the *hmwA* binding region sequences shared approximately 62% amino acid identity.

Four pairs of *hmwA* binding region sequences were 100% identical at the nucleotide level (F2865 $hmwA_{1679}$ and F894 $hmwA_{1598}$; I280 $hmwA_{1679}$ and I280 $hmwA_{1598}$; K15RE $hmwA_{1598}$ and H0421 $hmwA_{1679}$; I283 $hmwA_{1598}$ and G423 $hmwA_{1598}$), but only strain I280 contained identical *hmwAs* at both loci. When considering the translated *hmwA* nucleotide sequences, the binding regions at both loci, $hmwA_{1679}$ and $hmwA_{1598}$, within the strain varied by no more than three amino acids in four strains (I280, F651, F1015, and F11242); each of these strains encoded identical or nearly identical HMW adhesin binding domains at each *hmw* locus.

***hmwA* evolutionary analyses.** The PHI test (7) detected evidence of recombination ($P < 0.001$) within the *hmwA* sequences. Since recombination can preclude representing phylogenetic relationships among sequences with a single phylogeny, we employed a network approach to visualize *hmwA* relationships (Fig. 3.2). The Neighbor-Net LSfit index was 0.993, meaning that the phenetic distances between sequences on the network captured 99.3% of the maximum composite likelihood genetic distances. The network topology (Fig. 3.2) was identical to the bootstrap consensus tree estimated by the maximum likelihood method (Fig. 3.3) using maximum composite likelihood distances; this should not be surprising given the high LSfit index which means that, despite the evidence for recombination, the dataset is still relatively tree-like.

The *hmwA* phylogeny revealed four *hmwA* sequence clusters, but, the majority of *hmwA* sequences (83%) fell into either Clusters 1 or 2 (Fig. 3.3). Interestingly, of the 31 strains for which both *hmwA* loci were sequenced, 29 (94%) possessed at least one locus that belonged to Cluster 1 (Figs. 3.2 and 3.3). In two strains (F651 and I280) both genes, $hmwA_{1679}$ and $hmwA_{1598}$,

were placed in Cluster 1; in one strain (F1015) both genes were placed in Cluster 4; and in one strain (F1124-2) both genes were placed in Cluster 2.

Evolutionary distances and amino acid p-distances were estimated for Cluster 1 (Table 3.S2) and Cluster 2 (Table 3.S3) sequences separately whereas Clusters 3 and 4 sequences were combined into a single cluster for analysis (referred to as Cluster 3-4; Table 3.S4). The average evolutionary distances ranged from 0.502 to 0.636 nucleotide substitutions per site for Cluster 3-4 and Cluster 1, respectively. Average amino acid p-distances ranged from 0.3341 for Cluster 3-4 to 0.3467 for Cluster 1; *i.e.*, on average, any two Cluster 1 sequences differ at approximately 35% of their amino acid sites.

*hmwA* sequences did not cluster by locus, geographic region, or by disease (OM or commensal). There are, however, some examples of OM and commensal strains of the same MLST sequence type (ST) also having closely related *hmwA* sequences at both loci (Fig 3.3), *e.g.*, H0421 and K15RE2.6 are both ST-11, F1646 and F1942 are both ST-3, F894 and F2865 are both ST-155, and F1296-5 and K19RE2.4 are both ST-57; I202 is also a ST-57, the I202 Cluster 2 $hmwA_{1679}$ is similar to the other Cluster 2 ST-57s but the I202 Cluster 1 $hmwA_{1598}$ is more distantly related. With the exception of the ST-57 strains, each ST pair with closely related *hmwA* loci represent one OM and one commensal strain isolated from the same geographic region. The two ST-57s with similar *hmwA* sequences at both loci are both OM strains, but, one was isolated in Finland and the other in the US.

## Discussion

In this study, we determined the prevalence of *hmwA* in a collection of 170 well characterized NTHi strains (36) isolated from healthy children and children with otitis media between 1994 and 2002 from disparate geographic regions. The overall *hmwA* prevalence in this collection was 72% and OM strains were significantly more likely to be *hmwA*-positive than were commensal strains. When stratified by region, strains collected from Finland were significantly more likely to be *hmwA*-positive than were strains from Israel or the US. We further characterized 33 of the *hmwA*-positive strains by sequencing the binding regions of $hmwA_{1679}$ and $hmwA_{1598}$ and documented examples of both highly conserved and markedly diverse *hmwA* sequences within and between strains. Finally, we conducted phylogenetic analyses that demonstrated *hmwA* binding region sequences form four distinct sequence clusters independent of chromosomal location, disease status, or geographic location.

With the exception of the relatively high *hmwA* prevalence in Finland (85%), our findings are consistent with previous studies that report *hmw* prevalence estimates ranging from 38 – 80% (2, 22, 24, 25, 55, 62). In a diverse collection of NTHi strains, Ecevit *et al.*, reported an *hmw1A* prevalence of 55% and *hmw*-associated genes were significantly more likely to be present in ear strains when compared to throat strains (22). In a collection of 58 NTHi strains, 41% were defined as HMW-positive and the prevalence of HMW was significantly greater among disease isolates, OM and COPD, than throat isolates (62). Two additional studies suggest that approximately 38 - 59% of NTHi strains possess *hmwA* as detected with a PCR based screen (24, 25). When screened by Western blotting with antibodies raised against recombinant strain 12 HMW1 protein, approximately 75% of 125 epidemiologically unlinked NTHi strains were HMW-positive (5). Thus, while the overall *hmwA* prevalence in this collection is within the ranges reported from other strain collections, the high *hmwA* prevalence among strains from Finland is an interesting finding and warrants further investigation.

The phylogenetic relationships among the strains in this collection have been recently characterized using MLST analysis (36). *hmwA*-positive strains were distributed throughout the MLST phylogeny and, consistent with the phylogenetic relationships among strains as determined with MLST (36), do not cluster by geographic region or disease status (Fig. 3.1). The wide distribution of *hmw* across phylogenetic clusters is consistent with the observation of Erwin

*et al.*, that identified *hmw*-positive strains in nearly all major NTHi clades as defined by MLST (25). Similarly, *hmwA* was distributed across all major NTHi groupings in a collection of disease isolates characterized by multi-locus enzyme electrophoresis (55). Among the 170 NTHi strains included in this study, there were 109 different STs and 23 of these STs were represented by two or more strains. In 16 of the 23 STs the strains were concordant for the presence, or absence, of *hmwA*. Of particular note, the two most abundant STs, ST57 and ST34, both of which consist entirely of OM strains, were variable for *hmwA* presence; 17/18 (94%) of ST57s and 4/5 (80%) of ST34s were *hmwA*-positive. The fact that strains of the same ST can be discordant with respect to *hmwA* presence suggests that the MLST genes and *hmwA* loci circulate independently.

We interrogated a subset of our strain collection with $hmwA_{1679}$ and $hmwA_{1598}$ specific PCR assays to determine if *hmwA*-positive strains contained two *hmw* loci at the chromosomal locations defined in NTHi strain 12 (9). The majority of strains tested (93%) possessed two *hmwA* loci. Multiple PCR primer pairs, however, were required to successfully amplify one or both of the *hmwAs* from a small number of strains (Table 3.S1). The failure to detect one, or both, *hmwA* in some strains is likely due to false-negative PCR results owing to nucleotide diversity and not the absence of the target gene; false-positive PCR results have been reported previously (30, 62). There have, however, been reports of strains possessing a single *hmwA* locus when using PCR (10) and DNA probes specific to the variable regions of NTHi strain 12 $hmwA_{1679}$ and $hmwA_{1598}$ (22). In contrast, studies that rely on Southern hybridization with a probe that anneals to conserved regions of both $hmwA_{1679}$ and $hmwA_{1598}$ generally report two *hmwA* per cell (9). Thus, further examination of this strain collection via Southern blotting with a conserved *hmwA* probe is warranted.

The *hmwA* DNA sequence data currently available comes exclusively from NTHi disease isolates. To extend our understanding of *hmwA* sequence diversity among commensal and OM strains, we sequenced the *hmwA* binding regions from a subset of strains from our collection. Strains were chosen for sequencing with two primary objectives in mind. First, we sought to capture a wide range of the *hmwA* sequence diversity; thus, we selected strains that were widely distributed across the MLST phylogeny and from all three geographic regions. Secondly, we wished compare *hmwA* distribution and diversity by disease status; thus, we chose approximately equal numbers of OM and commensal strains.

Targeted sequencing of the *hwmA* binding regions of 33 strains revealed a wide range of diversity in both amino acid composition and size of the binding region. The overall pairwise amino acid identity among the translated *hmwA*s sequenced from this strain collection, independent of *hmwA* locus, was approximately 62%. Evecit *et al.*, reported pairwise amino acid identities slightly higher, 66 – 77%, than those reported here, but, their analysis included the conserved signal sequences and thus would be expected to be higher (21). A study comparing the entire mature binding protein reported only 52% identity, on average, when comparing across both *hmwA* loci from three NTHi strains (9). Finally, although Giufre *et al*., reported higher diversity within the *hmwA* core binding domain, their study focused on a smaller portion of the most diverse region of *hmwA* (30). Interestingly, they also reported four core-binding domains from invasive isolates that were 100% identical at the nucleotide level to strain 12 *hmw1A* and two core-binding domains that were 100% identical to *hmw2A* (30); we did not have any sequences represented more than twice in our collection. The *hmwA* binding domains also varied in length (Table 3.2). This size variation is consistent with other studies (5, 9, 18, 21) and is also evident by Western blotting (4, 9, 11, 29, 62).

Two of the *hmwA* binding domains sequenced in this study contained an in-frame stop codon. In a previous study of *hmwA* sequence diversity, Giufre *et al*., identified a stop codon within the strain 72 HMW2 nucleotide sequence resulting from a single base pair deletion (30). Additionally, they failed to detect strain 72 HMW2 protein by Western blotting, but, it is unclear if that is a result of phase variation, lack of antibody cross-reactivity, or a consequence of the premature stop codon (30). Both examples in our study appear to be the result of single nucleotide deletions within homopolymeric tracts. Homopolymeric tracts of more than eight nucleotides (*e.g.* nine or more consecutive A's) are considered simple sequence repeats (SSRs) (44). Variation in SSRs within the *hmwA* coding region could introduce frameshifts resulting in phase variable HMW adhesin expression. A survey of four fully sequenced *H. influenzae* genomes identified between 13 and 17 homopolymeric SSRs per genome, but, none of the SSRs were predicted to be phase variable (44). Six *hmwA* sequences in the current study showed a total of ten homopolymeric SSRs (seven adenine (A) and three guanine (G)) ranging in length from nine to ten repeats each. One of the frameshifted *hmwA* sequences in our study (I168 $hmwA_{1598}$) had a single A-deletion in one of the nine-repeat tracts and the second (G423 $hmwA_{1679}$) had a single G-deletion in one of the ten-repeat tracts. These findings raise the possibility that

homopolymeric SSRs within the *hmwA* coding sequence represents a second phase variable mechanism affecting HWM adhesin expression. It is also possible, however, that the insertions we observed represent PCR artifacts. Further studies are necessary to determine if these SSRs truly mediate phase variable HMW adhesin expression.

Four strains in this collection possessed nearly identical *hmwA* sequences at both *hmw* loci and these strains possessed *hmwA* sequences that fell into three different sequence clusters. One possible explanation for the similarity of the sequences within a single strain is that they represent recent gene conversion events. The gastric pathogen *Helicobacter pylori* has a diverse array of adhesins, including *babA* and *babB*, which are members of a paralogous family of outer membrane proteins that interact with the gastric epithelium (48). During the course of infection, *H. pylori babB*-mediated adherence is altered by both gene conversion events, resulting in the loss of *babA* and the duplication of *babB* and by the loss of *babA* adherence via changes in dinucleotide repeats (48). *H. pylori* intragenomic gene conversion events have also been demonstrated in outer membrane proteins, *sabA* and *sabB*, and duplication of *sabA* increases SabA production and SabA-mediated adherence (57). Thus, a single NTHi strain that encodes two nearly identical HMW proteins may be restricted to a more limited environmental niche while displaying increased adherence within that niche. Furthermore, a strain with limited HMW diversity could potentially face decreased fitness in the face of antibody mediated immunity, but, such a cost could potentially be mitigated by *hmwA* phase variation mediated by simple sequence repeats. While the *hmwA* sequences described here display a high degree of amino acid identity in the binding region, their binding regions differ at a small number of sites and these minor differences could potentially affect adherence properties and/or immunogenicity of the encoded adhesin.

NTHi is naturally competent and sequence analysis has demonstrated evidence of both horizontal gene transfer and recombination (12-14, 16, 25, 33, 36, 38-40, 43 , 47). Recombination confounds phylogenetic reconstructions because no single phylogeny can represent the evolutionary history of a recombinant gene. We therefore employed a network based approach, the Neighbor-Net method (35), to visualize relationships among *hmwA* core binding regions. In a Neighbor-Net, conflicting phylogenetic signals are represented by boxes along the pathways connecting nodes. Since conflicting phylogenetic signals can arise from

processes other than homologous recombination, such as, evolutionary rate heterogeneity between sequences, we specifically tested for the presence of *hmwA* recombination. We found strong evidence for recombination leading us to conclude that the inconsistencies reflected in the *hmwA* phylogenetic network (Fig. 3.2) were attributable, at least in part, to homologous recombination. Although the network topology for the *hmwA*-family core binding region revealed four sequence clusters, most sequences (84%) fell into either Cluster 1 or Cluster 2 (Fig. 3.2). There was no evidence for disease-associated sequence clusters; all four clusters contain *hmwA* binding sequences from both OM and commensal strains.

Interestingly, the majority of strains (94%) for which both loci were sequenced possess an *hmwA* that falls within Cluster 1. The HMW adhesins of NTHi strain 12 are known to differ in their *in vitro* adherence properties (9, 53, 62). NTHi strain 12 HMW1 interacts with α-2,3 N-linked sialic acid moieties on glycoproteins of the respiratory epithelium (49), whereas the receptor for strain 12 HMW2 remains uncharacterized. Previous phylogenetic analyses demonstrate that *hmwA* sequences cluster into groups based upon their *in vitro* adherence properties (9). Thus, the sequences in Cluster 1 might be expected to have binding characteristics similar to the NTHi strain 12 HMW2. While investigating adherence characteristics in a diverse collection of NTHi isolates, van Schilfgaarde *et al*., identified a small number of NTHi isolates that displayed HMW adherence properties distinct from the strain 12 HMW1- and HMW2-type adherence (62). In total, four distinct adherence patterns were described, which is interesting given that we identified four *hmwA* sequence clusters. It is possible, therefore, that the Cluster 3-4 *hmwA* sequences encode HMW adhesins with novel adherence profiles.

Characterization of a diverse collection of NTHi strains with an electrophoretic typing scheme suggests that OM strains are more clonal than COPD isolates (60). A second study noted that, when tested against a panel of monoclonal antibodies generated against NTHi Strain 12 HMW adhesins, OM strains expressed less diverse HMW adhesins than did COPD strains or strains isolated from the throats of healthy individuals (62). Using the *hmwA* binding sequences, we asked whether the OM strains and commensal strains within each cluster exhibited similar levels of genetic diversity. Commensal strains from Cluster 1 and Cluster 2 display a wider range of genetic diversity, with mean genetic distances of 0.652 (± 0.220) and 0.583 (± 2.609), respectively, than OM strains, with mean genetic distances of 0.621 (± 0.020) and 0.562 (±

0.295). The fact that *hmwA* core binding sequences from OM strains are less diverse than throat strains is consistent with the general view that OM strains represent a limited subset of the much larger population of commensal strains.

It is important to exercise caution when evaluating the role of HMW adhesins during OM pathogenesis based solely on prevalence data. Prevalence data alone do not provide any information regarding expression of functional HMW adhesins and simply reflect the presence of the adhesin genes. The relationship between *hmwA* presence and HMW expression is complicated by the fact that both *hmwA* loci contain tandem repeats located within their promoter regions and these repeats, which are gained and lost via slipped strand mispairing, affect the level HMW adhesin production (17). High numbers of repeats are associated with reduced HMW production and reduced NTHi adherence. Since the HMW adhesins are immunogenic, phase variation provides a population-level mechanism that allows NTHi to avoid clearance by HMW-specific adaptive immunity. Indeed, mathematical modeling suggests that HMW-specific antibody mediated immune pressure can shape the within-host NTHi population structure, selecting for bacterial cells with a high repeat number and low levels of surface expressed HMW (Dissertation Chapter 5). Variation in HMW adhesin expression has implications for NTHi pathogenesis during both OM and COPD. For example, NTHi isolates collected from the ears of children with AOM have higher numbers of *hmwA*–associated simple sequence repeats than matched isolates collected from the throat (17) and serial isolates collected over time from COPD patients reveal increases in repeat numbers over time, which is associated with decreased HMW production and reduced NTHi *in-vitro* adherence (11). As these examples illustrate, the simple presence or absence of *hmwA* only tells part of the story. The fact that *hmwA* is more prevalent among OM isolates, however, suggests that *hmwA*-containing strains are more likely to survive in the middle ear space and cause disease than their *hmwA*-negative counterparts and supports a role for HMW adhesins in NTHi virulence.

Recently a pneumococcal-NTHi protein D conjugate vaccine, composed pneumococcal polysaccharide capsule antigens conjugated to NTHi protein D, has been approved for use in children in Canada and several European countries (15, 20, 46). Protein D is a highly conserved outer membrane lipoprotein present in both typeable and nontypeable *H. influenzae* (28). The conjugate vaccine effectively protects against invasive pneumococcal disease (19) but is

ineffective against NTHi as it does not reduce NTHi nasopharyngeal colonization (45, 61). Thus, efforts to develop an NTHi vaccine are ongoing. Ideally, an effective vaccine would specifically target strains with increased pathogenic potential, leaving behind a community of "commensal" NTHi in the pharynx. If *hmwA* sequences from OM strains formed a distinct phylogenetic cluster then those sequences could be used specifically for a vaccine designed to target only a subset of virulent NTHi strains.

HMW has long been championed as an attractive candidate for inclusion in a multivalent vaccine since it is both surface exposed and highly immunogenic (1, 5, 63, 64). As highlighted in the current study, the observed amino acid diversity coupled with the lack of any apparent disease-specific *hmwA* sequences likely precludes HMW's usefulness as a vaccine component. Furthermore, mathematical modeling of the interaction between *hmwA* phase variation and the host adaptive immune response suggests that a rapid and strong anti-HMW immune response, as might be induced by secondary exposure following successful vaccination, may actually increase the duration of time that the colonizing NTHi population remains adherent and, as a consequence, may increase NTHi transmissibility (Chapter 5 of dissertation). It should be emphasized, however, that while interesting, the modeling results are preliminary and warrant further analysis. Taken together, these findings cast a long shadow over HMW's potential as one component of a multi-component NTHi vaccine.

In summary, the prevalence of *hmwA* among this strain collection was 72% and prevalence varied by geographic region. Overall, OM strains were significantly more likely to possess *hmwA* than were commensal strains. The HMW adhesins are presumably costly to produce both from an energetic perspective given their large size, especially when compared to the relatively small size of the NTHi chromosome, and from a fitness perspective since they are an immunodominant target of the host's adaptive immune system (3). The fact that a majority of strains interrogated for $hmwA_{1697}$ and $hmwA_{1598}$ possessed an *hmwA* at both loci suggests that having two copies confers a real fitness advantage over having a single *hmwA*, even when both copies possess identical, or nearly identical, binding domains. Phylogenetic network analysis suggests that *hmwA* sequences form four distinct sequence clusters, but, nearly all strains possess an *hmwA* that falls within Cluster 1 raising the possibility that Cluster 1 sequences may represent the ancestral adhesin and that, following duplication, the second copy has taken on expanded

adherence properties possibly defining distinct niches within the pharynx. Future studies, aimed at defining the adherence profiles of Cluster 3 and 4 sequences, will help determine if these proteins display novel adherence characteristics.

Table 3.1. PCR primers used in this study.

| Primer Name | Primer Sequence (5' to 3') | Reference |
|---|---|---|
| *hmwA*_leader_F | GCCAATTTCCGCTTCACCCTCTTT | this study |
| *hmwA*_leader_R | AACAACTCCGCCGTATTTAACCGC | this study |
| A1FL5.2 | YGATAGSGTAGATCTCCCCGCCTTTGC | this study |
| *hmwA*1F_ORF5-1 | TGGAACTTCTTTTGCTGTGGCTGATGC | (30) |
| A2FL5.2 | TTGTGGCAATTCAATACCTATTTGTGG | this study |
| *hmwA*2F_ORF5-1 | CCTCTTAATTGGGCATTAGTTGG | (30) |
| *hmwA*_core_1R | CCGGTGATATTCACGCTGCTTGAGG | this study |
| *hmwA*R_HMWB3R | GATGAAGAAGCCAGGCCAAGCAATAC | (30). |
| *hmwA*seqR_*Not*I | ATGATCAGCGGCCGCGGSGTGATRTTYACDYTRCTTGAGG | this study |
| *hmwA*seqF_*Spe*I | CAGTCAGACTAGTAAGGYAAAAAYGGYATTCAATTAGC | this study |

Table 3.2. NTHi strains selected for *hmwA* sequencing.

| Strain | ST | Region | Site of isolation | *hmwA*$_{1679}$ (nt)[a] | *hmwA*$_{1679}$ cluster[b] | *hmwA*$_{1598}$ (nt)[a] | *hmwA*$_{1598}$ cluster[b] |
|---|---|---|---|---|---|---|---|
| F164-6 | 3 | Finland | Ear | 2850 | 3 | 2739 | 1 |
| F286-5 | 155 | Finland | Ear | 3060 | 1 | 3567 | 2 |
| F1124-2 | 12 | Finland | Ear | 3486 | 2 | 3477 | 2 |
| F1152-8 | 40 | Finland | Ear | 2796 | 1 | 3345 | 2 |
| F1296-5 | 57 | Finland | Ear | 3030 | 2 | 2949 | 1 |
| F120 | 472 | Finland | Throat | 3132 | 2 | 2880 | 1 |
| F441 | 253 | Finland | Throat | 2967 | 2 | ND | |
| F651 | 852 | Finland | Throat | 3069 | 1 | 3069 | 1 |
| F894 | 155 | Finland | Throat | 3516 | 2 | 3111 | 1 |
| F938 | 853 | Finland | Throat | 2934 | 1 | ND | |
| F1015 | 245 | Finland | Throat | 3414 | 4 | 3414 | 4 |
| F1942 | 3 | Finland | Throat | 2853 | 1 | 2850 | 3 |
| I168 | 12 | Israel | Ear | 2784 | 1 | STOP | |
| I198 | 877 | Israel | Ear | 2697 | 2 | 2652 | 1 |
| I202 | 57 | Israel | Ear | 2664 | 2 | 2868 | 1 |
| I208 | 893 | Israel | Ear | 2847 | 1 | 3045 | 2 |
| I213 | 244 | Israel | Ear | 3408 | 3 | 3354 | 1 |
| I218 | 238 | Israel | Ear | 3186 | 3 | 2871 | 1 |
| I249 | 859 | Israel | Throat | 3042 | 3 | ND | |
| I256 | 764 | Israel | Throat | 3573 | 2 | 2868 | 1 |
| I280 | 892 | Israel | Throat | 3000 | 1 | 3009 | 1 |
| I283 | 165 | Israel | Throat | 3009 | 1 | 3306 | 2 |
| I336 | 866 | Israel | Throat | 2865 | 1 | 2949 | 2 |
| G423 | 34 | US | Ear | STOP | | 3384 | 2 |
| G922 | 203 | US | Ear | 2778 | 1 | 3213 | 4 |
| G1822 | 34 | US | Ear | ND | | 3180 | 1 |
| K15RE2.6 | 11 | US | Ear | 3060 | 1 | 3243 | 2 |
| K19RE2.4 | 57 | US | Ear | 2877 | 1 | 3030 | 2 |
| K32RE2.5 | 143 | US | Ear | 3387 | 2 | 2919 | 1 |
| 63.4-22 | 196 | US | Throat | 3117 | 1 | 3516 | 4 |
| H04.2.1 | 11 | US | Throat | 3243 | 2 | 3060 | 1 |
| J06.2.2 | 880 | US | Throat | 2844 | 1 | 3042 | 2 |
| N02.2.3 | 881 | US | Throat | ND | | 2973 | 1 |
| **Sequences obtained from GenBank** | | | | | | | |
| strain 12 | | | | 3114 | 2 | 2934 | 1 |
| strain 5 | | | | 3297 | 2 | 3306 | 1 |
| strain 72 | | | | 2910 | 2 | STOP | |
| AAr96 | | | | NA | | 3045 | 2 |
| PittEE | | | | 3246 | 3 | 3267 | 1 |
| 86-028NP | | | | 2979 | 3 | 3129 | 1 |
| A950006 | | | | 3174 | 3 | NA | |
| G822 | | | | 2949 | 1 | NA | |
| AAr105 | | | | 2973 | 1 | NA | |

[a] NA = not available, ND = not determined, STOP = in-frame stop codon

[b] Phylogenetic Cluster to which each *hmwA* belongs (see Fig. 3.3)

Table 3.3. *hmwA* prevalence as determined by PCR.

| All (n = 170) | *hmwA*$^{+}$ (%) | *hmwA*$^{-}$ (%) | PR (95% CI)[a] | P-value[b] |
|---|---|---|---|---|
| OM | 75 (61) | 20 (43) | 1.23 (1.01, 1.51) | 0.038 |
| Commensal | 48 (39) | 27 (57) | | |
| Total | 123 (100) | 47 (100) | | |
| Finland (n = 55) | | | | |
| OM | 26 (55) | 3 (37) | 1.11 (0.89, 1.39) | 0.455 |
| Commensal | 21 (45) | 5 (63) | | |
| Total | 47 (100) | 8 (100) | | |
| Israel (n = 50) | | | | |
| OM | 21 (64) | 7 (41) | 1.38 (0.89, 2.13) | 0.147 |
| Commensal | 12 (36) | 10 (59) | | |
| Total | 33 (100) | 17 (100) | | |
| US (n = 65) | | | | |
| OM | 28 (67) | 10 (45) | 1.33 (0.90, 1.95) | 0.184 |
| Commensal | 15 (33) | 12 (55) | | |
| Total | 43 (100) | 22 (100) | | |

[a] Prevalence ratio (OM/commensal) and 95% confidence intervals

[b] Fisher's exact test two-tailed P-value

Table 3.4. *hmwA* loci specific PCR results among a subset of the strain collection.

| *hmwA* locus | OM (n = 59) | Commensal (n = 32) | Total (n = 91) |
|---|---|---|---|
| both | 54 | 31 | 85 |
| $hmwA_{1679}$ only | 1 | 0 | 1 |
| $hmwA_{1598}$ only | 2 | 0 | 2 |
| neither | 2 | 1 | 3 |

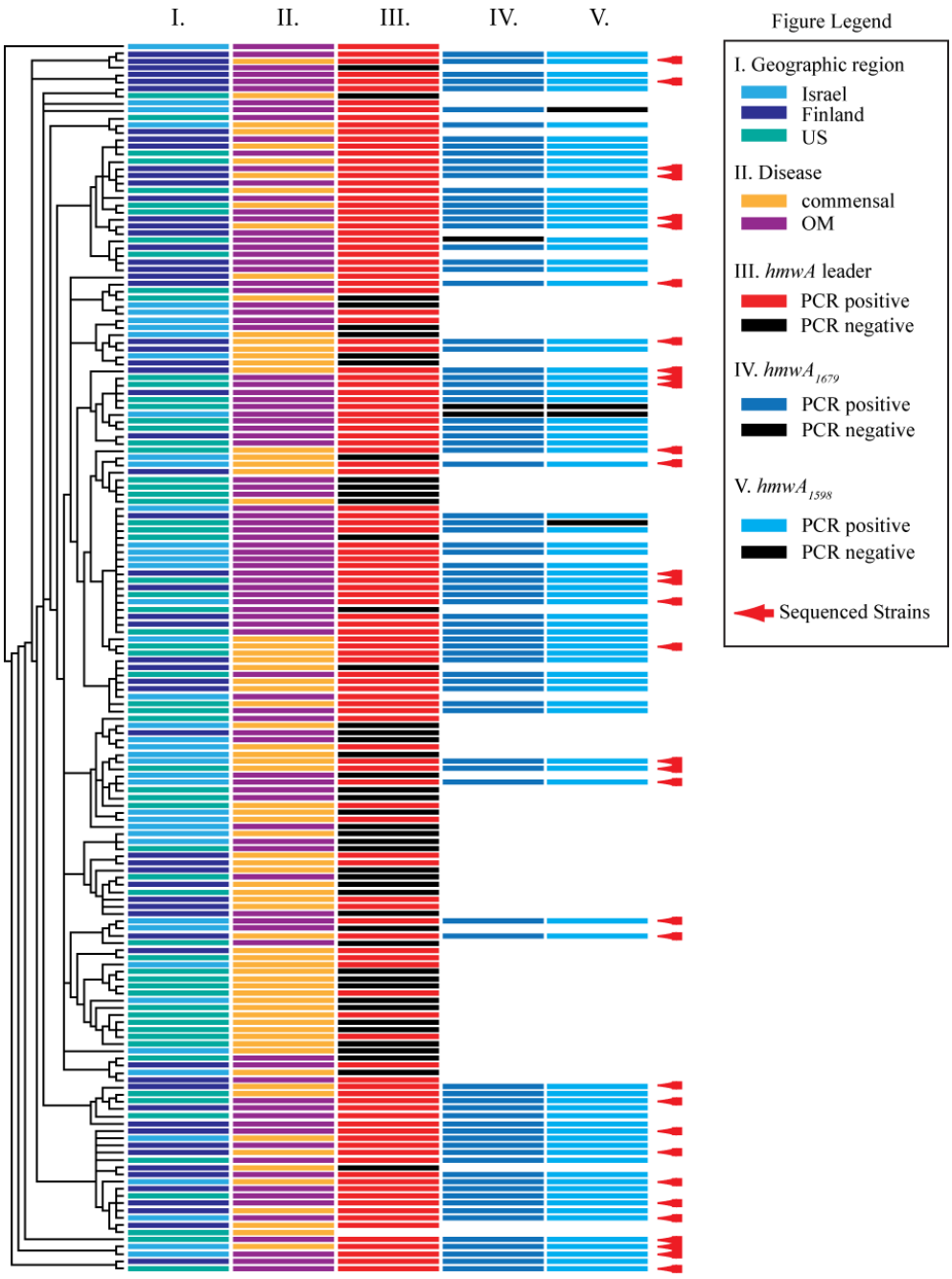



Figure 3.1. MLST phylogeny for the 170 strains used in this study (phylogeny taken from (36)) . The first two columns are color coded according to the geographic region the strain was collected from (II) and the disease status of the subject (II). Of the 123 *hmwA*-leader positive strains (III), 91 were selected for interrogation by PCR assay for the presence of $hmwA_{1679}$ (IV) and $hmwA_{1598}$ (V). Strains chosen for *hmwA* binding domain DNA sequence analysis are identified by the red arrows.

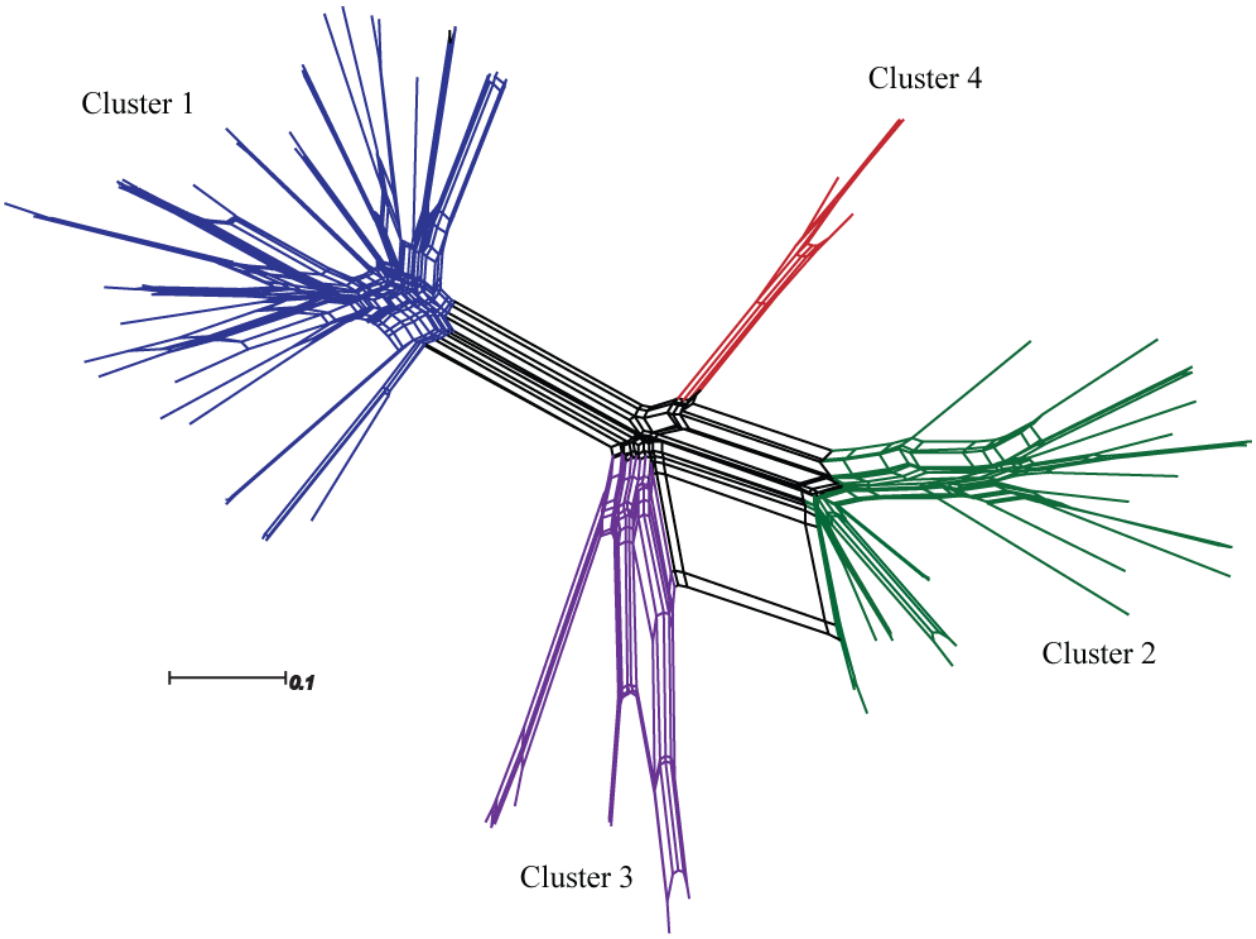



Figure 3.2. Neighbor-Net phylogenetic network for all *hmwA* binding domain sequences. The network was constructed from genetic distances estimated using the maximum composite likelihood model with rate variation among sites was modeled with a gamma distribution and five rate categories (shape parameter = 0.567)(58). Boxes in the network reflect inconsistencies in the phylogenetic signal such as those arising from recombination; LSfit = 0.9927. There are four distinct *hmwA* sequence clusters, Cluster 1(blue), Cluster 2 (green), Cluster 3 (purple), and Cluster 4(red). The majority of strains (94%) possess an *hmwA* that belongs to Cluster 1 with the second *hmwA* from any given strain falling within the Clusters 2, 3, or 4. All positions containing gaps and missing data were eliminated from the analysis resulting in a total of 1809 positions in the final dataset.

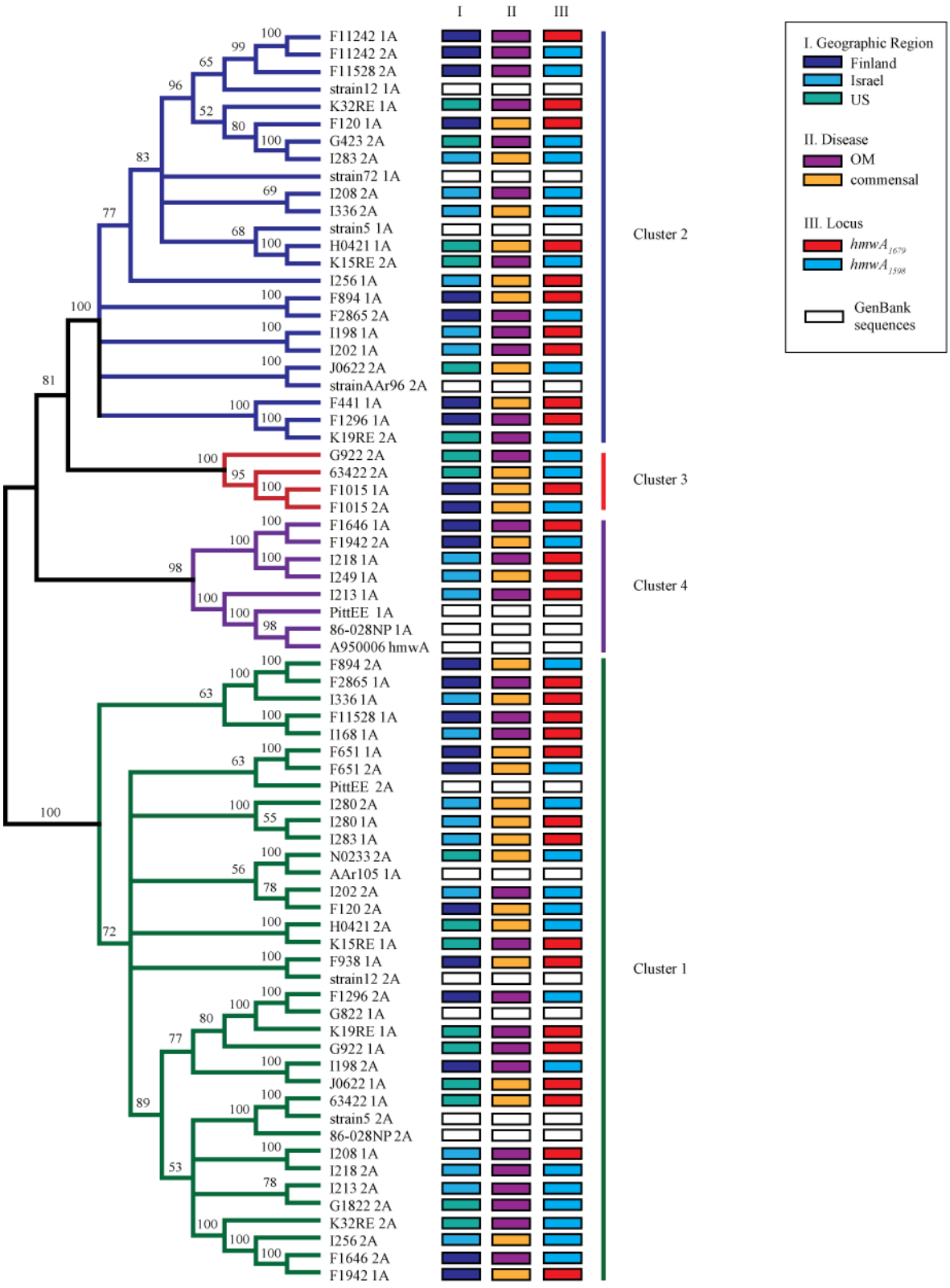

I
II
III
F11242 1A
F11242 2A
F11528 2A
strain12 1A
K32RE 1A
F120 1A
G423 2A
I283 2A
strain72 1A
I208 2A
I336 2A
strain5 1A
H0421 1A
K15RE 2A
I256 1A
F894 1A
F2865 2A
I198 1A
I202 1A
J0622 2A
strainAAr96 2A
F441 1A
F1296 1A
K19RE 2A
G922 2A
63422 2A
F1015 1A
F1015 2A
F1646 1A
F1942 2A
I218 1A
I249 1A
I213 1A
PittEE 1A
86-028NP 1A
A950006 hmwA
F894 2A
F2865 1A
I336 1A
F11528 1A
I168 1A
F651 1A
F651 2A
PittEE 2A
I280 2A
I280 1A
I283 1A
N0233 2A
AAr105 1A
I202 2A
F120 2A
H0421 2A
K15RE 1A
F938 1A
strain12 2A
F1296 2A
G822 1A
K19RE 1A
G922 1A
I198 2A
J0622 1A
63422 1A
strain5 2A
86-028NP 2A
I208 1A
I218 2A
I213 2A
G1822 2A
K32RE 2A
I256 2A
F1646 2A
F1942 1A
Cluster 2
Cluster 3
Cluster 4
Cluster 1
I. Geographic Region
Finland
Israel
US
II. Disease
OM
commensal
III. Locus
hmwA1679
hmwA1598
GenBank sequences

Figure 3.3. *hmwA* core binding region evolutionary relationships were estimated using the maximum likelihood approach. *hmwA* sequences form four distinct sequence clusters and most strains (94%) possess an *hmwA* that falls within Cluster 1. The sequences do not cluster by geographic region (I), disease status (II), or by *hmwA* locus (III). The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (1000 replicates) are shown next to the branches (26); branches recovered in less than 50% of the bootstrap replicates were collapsed. Differences in the evolutionary rates among sites were modeled with a five category discrete gamma distribution (shape parameter = 0.6044). All positions containing gaps and missing data were eliminated from the analysis resulting in a total of 1809 positions in the final dataset. Evolutionary analyses were conducted in MEGA5 (59).

## Supplemental Tables and Figures

Table 3.S1. *hmwA* amps for G strains (see Table 2 for primer sequences).

| | ***hmwA*$_{1679}$** | | ***hmwA*$_{1598}$** | |
|---|---|---|---|---|
| **Strain** | **forward primer** | **reverse primer** | **forward primer** | **reverse primer** |
| G322 | hmwA1F_ORF5-1 | hmwAR_HMWB3R | hmwA2F_ORF5-1 | hmwAR_HMWB3R |
| G423 | hmwA1F_ORF5-1 | hmwA_core_1R | A2FL5.2 | hmwA_core_1R |
| G822 | hmwA1F_ORF5-1 | hmwAR_HMWB3R | hmwA2F_ORF5-1 | hmwAR_HMWB3R |
| G922 | hmwA1F_ORF5-1 | hmwA_core_1R | hmwA2F_ORF5-1 | hmwA_core_1R |
| G1222 | no amplification | no amplification | A2FL5.2 | hmwA_core_1R |
| G1322 | hmwA1F_ORF5-1 | hmwA_core_1R | A2FL5.2 | hmwA_core_1R |
| G1522 | hmwA1F_ORF5-1 | hmwAR_HMWB3R | no amplification | no amplification |
| G1822 | A1FL5.2 | hmwAR_HMWB3R | A2FL5.2 | hmwA_core_1R |

Table 3.S2. Cluster 1 pairwise amino acid p-distances (above the diagonal) and pairwise maximum composite likelihood genetic distances (below the diagonal).

| | | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|
| | **Strain_locus** | 63422_1A | F120_2A | F651_1A | F651_2A | F894_2A | F938_1A |
| 1 | 63422_1A | | 0.4120 | 0.4398 | 0.4398 | 0.3832 | 0.3775 |
| 2 | F120_2A | 0.8467 | | 0.3407 | 0.3375 | 0.3438 | 0.3462 |
| 3 | F651_1A | 0.9792 | 0.5506 | | 0.0029 | 0.3678 | 0.3194 |
| 4 | F651_2A | 0.9785 | 0.5418 | 0.0020 | | 0.3698 | 0.3226 |
| 5 | F894_2A | 0.7802 | 0.5957 | 0.7435 | 0.7517 | | 0.3409 |
| 6 | F938_1A | 0.7642 | 0.5678 | 0.4997 | 0.5080 | 0.5720 | |
| 7 | F1296_2A | 0.5008 | 0.6907 | 0.6316 | 0.6416 | 0.8974 | 0.5268 |
| 8 | F1646_2A | 0.5627 | 0.7288 | 0.7688 | 0.7813 | 0.5948 | 0.5566 |
| 9 | F1942_1A | 0.5755 | 0.7192 | 0.8144 | 0.8013 | 0.6230 | 0.5549 |
| 10 | F2865_1A | 0.7906 | 0.5955 | 0.7551 | 0.7549 | 0.0000 | 0.5824 |
| 11 | F11528_1A | 0.8228 | 0.7295 | 0.8540 | 0.8525 | 0.7054 | 0.6421 |
| 12 | G922_1A | 0.5219 | 0.7249 | 0.7511 | 0.7629 | 0.6581 | 0.4835 |
| 13 | G1822_2A | 0.5836 | 0.6184 | 0.7918 | 0.8031 | 0.4980 | 0.4959 |
| 14 | H0421_2A | 0.8602 | 0.5172 | 0.6449 | 0.6548 | 0.6185 | 0.4735 |
| 15 | I168_1A | 0.8069 | 0.7307 | 0.8491 | 0.8382 | 0.7030 | 0.6429 |
| 16 | I198_2A | 0.7683 | 0.7769 | 0.7741 | 0.7613 | 0.8579 | 0.7271 |
| 17 | I202_2A | 0.6161 | 0.4981 | 0.5864 | 0.5935 | 0.7861 | 0.6194 |
| 18 | I208_1A | 0.7161 | 0.5542 | 0.6728 | 0.6632 | 0.5735 | 0.5436 |
| 19 | I213_2A | 0.6834 | 0.5454 | 0.8183 | 0.8271 | 0.5906 | 0.4603 |
| 20 | I218_2A | 0.7034 | 0.5836 | 0.6876 | 0.6981 | 0.6046 | 0.4665 |
| 21 | I256_2A | 0.5761 | 0.6980 | 0.7885 | 0.7759 | 0.6503 | 0.5487 |
| 22 | I280_1A | 0.7344 | 0.5793 | 0.5623 | 0.5621 | 0.6945 | 0.6716 |
| 23 | I280_2A | 0.7343 | 0.5789 | 0.5710 | 0.5618 | 0.7035 | 0.6811 |
| 24 | I283_1A | 0.7415 | 0.5931 | 0.5679 | 0.5746 | 0.6970 | 0.6775 |
| 25 | I336_1A | 0.6643 | 0.6652 | 0.5817 | 0.5877 | 0.4324 | 0.6734 |
| 26 | J0622_1A | 0.7565 | 0.7656 | 0.7506 | 0.7622 | 0.9160 | 0.6949 |
| 27 | K15RE_1A | 0.8694 | 0.5258 | 0.6648 | 0.6658 | 0.6365 | 0.4947 |
| 28 | K19RE_1A | 0.5400 | 0.6613 | 0.6120 | 0.6220 | 0.7255 | 0.3479 |
| 29 | K32RE_2A | 0.6712 | 0.6567 | 0.6730 | 0.6809 | 0.5944 | 0.5027 |
| 30 | N0233_2A | 0.6896 | 0.5066 | 0.7030 | 0.6921 | 0.6152 | 0.4830 |
| 31 | G822_1A | 0.5202 | 0.7462 | 0.6477 | 0.6579 | 0.8984 | 0.5352 |
| 32 | 86-028NP_2A | 0.3912 | 0.6643 | 0.7980 | 0.8065 | 0.6100 | 0.5135 |
| 33 | strain 5_2A | 0.3222 | 0.7059 | 0.6127 | 0.6215 | 0.7277 | 0.7148 |
| 34 | strain 12_2A | 0.7973 | 0.5740 | 0.5291 | 0.5356 | 0.5965 | 0.0376 |
| 35 | AAr105_1A | 0.6586 | 0.4930 | 0.6378 | 0.6473 | 0.5417 | 0.5846 |

Table 3.S2. Cluster 1 distances (con't).

| | | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|
| | **Strain_locus** | F1296_2A | F1646_2A | F1942_1A | F2865_1A | F11528_1A | G922_1A |
| 1 | 63422_1A | 0.3016 | 0.3379 | 0.3409 | 0.3852 | 0.3901 | 0.3230 |
| 2 | F120_2A | 0.3803 | 0.3801 | 0.3807 | 0.3448 | 0.3949 | 0.3825 |
| 3 | F651_1A | 0.3710 | 0.3851 | 0.3985 | 0.3711 | 0.3956 | 0.3945 |
| 4 | F651_2A | 0.3742 | 0.3884 | 0.3952 | 0.3711 | 0.3956 | 0.3978 |
| 5 | F894_2A | 0.3975 | 0.3455 | 0.3539 | 0.0000 | 0.3589 | 0.3600 |
| 6 | F938_1A | 0.3222 | 0.3189 | 0.3209 | 0.3449 | 0.3558 | 0.3144 |
| 7 | F1296_2A | | 0.3330 | 0.3510 | 0.4025 | 0.3785 | 0.3209 |
| 8 | F1646_2A | 0.5538 | | 0.0143 | 0.3488 | 0.3592 | 0.3464 |
| 9 | F1942_1A | 0.5912 | 0.0094 | | 0.3550 | 0.3682 | 0.3506 |
| 10 | F2865_1A | 0.9174 | 0.6066 | 0.6274 | | 0.3593 | 0.3645 |
| 11 | F11528_1A | 0.8033 | 0.6585 | 0.6868 | 0.7028 | | 0.4074 |
| 12 | G922_1A | 0.4939 | 0.5626 | 0.5613 | 0.6729 | 0.8413 | |
| 13 | G1822_2A | 0.6859 | 0.5121 | 0.5472 | 0.5034 | 0.6821 | 0.6448 |
| 14 | H0421_2A | 0.6740 | 0.5909 | 0.6468 | 0.6261 | 0.6588 | 0.7397 |
| 15 | I168_1A | 0.8124 | 0.6699 | 0.6919 | 0.6924 | 0.0077 | 0.8376 |
| 16 | I198_2A | 0.5171 | 0.5834 | 0.5883 | 0.8639 | 0.8667 | 0.7365 |
| 17 | I202_2A | 0.5731 | 0.7210 | 0.7500 | 0.8010 | 0.6619 | 0.5656 |
| 18 | I208_1A | 0.6803 | 0.6405 | 0.6408 | 0.5681 | 0.5938 | 0.6308 |
| 19 | I213_2A | 0.6873 | 0.4685 | 0.4862 | 0.6091 | 0.6073 | 0.6660 |
| 20 | I218_2A | 0.6843 | 0.5858 | 0.6145 | 0.6166 | 0.5774 | 0.6004 |
| 21 | I256_2A | 0.5879 | 0.0474 | 0.0364 | 0.6540 | 0.6702 | 0.5660 |
| 22 | I280_1A | 0.6715 | 0.6668 | 0.7346 | 0.7072 | 0.7144 | 0.8487 |
| 23 | I280_2A | 0.6812 | 0.6771 | 0.7334 | 0.7071 | 0.7179 | 0.8605 |
| 24 | I283_1A | 0.6786 | 0.6731 | 0.7422 | 0.7139 | 0.7124 | 0.8555 |
| 25 | I336_1A | 0.7992 | 0.7205 | 0.7735 | 0.4328 | 0.7065 | 0.7771 |
| 26 | J0622_1A | 0.3629 | 0.5341 | 0.5735 | 0.9439 | 0.8727 | 0.7054 |
| 27 | K15RE_1A | 0.6971 | 0.6187 | 0.6662 | 0.6317 | 0.6535 | 0.7617 |
| 28 | K19RE_1A | 0.0945 | 0.6000 | 0.6037 | 0.7396 | 0.8056 | 0.3209 |
| 29 | K32RE_2A | 0.6392 | 0.2469 | 0.2674 | 0.6030 | 0.4884 | 0.5803 |
| 30 | N0233_2A | 0.6865 | 0.6720 | 0.6724 | 0.6168 | 0.5394 | 0.6513 |
| 31 | G822_1A | 0.0217 | 0.5517 | 0.5911 | 0.9186 | 0.8368 | 0.4888 |
| 32 | 86-028NP_2A | 0.7369 | 0.5579 | 0.5828 | 0.6261 | 0.5093 | 0.5818 |
| 33 | strain 5_2A | 0.6015 | 0.5687 | 0.6158 | 0.7385 | 0.7770 | 0.6305 |
| 34 | strain 12_2A | 0.5249 | 0.5815 | 0.5827 | 0.6163 | 0.5989 | 0.5060 |
| 35 | AAr105_1A | 0.6766 | 0.6781 | 0.7029 | 0.5485 | 0.7016 | 0.6494 |

Table 3.S2. Cluster 1 distances (con't).

| | | 13 | 14 | 15 | 16 | 17 | 18 | 19 |
|---|---|---|---|---|---|---|---|---|
| | **Strain_locus** | G1822_2A | H0421_2A | I168_1A | I198_2A | I202_2A | I208_1A | I213_2A |
| 1 | 63422_1A | 0.3432 | 0.3909 | 0.3860 | 0.3859 | 0.3340 | 0.3723 | 0.3821 |
| 2 | F120_2A | 0.3582 | 0.3221 | 0.3942 | 0.4089 | 0.3144 | 0.3584 | 0.3291 |
| 3 | F651_1A | 0.3937 | 0.3421 | 0.3960 | 0.3977 | 0.3311 | 0.3684 | 0.3976 |
| 4 | F651_2A | 0.3957 | 0.3453 | 0.3949 | 0.3966 | 0.3322 | 0.3662 | 0.3996 |
| 5 | F894_2A | 0.2951 | 0.3502 | 0.3605 | 0.4027 | 0.3783 | 0.3468 | 0.3553 |
| 6 | F938_1A | 0.3175 | 0.3172 | 0.3563 | 0.3736 | 0.3448 | 0.3405 | 0.3006 |
| 7 | F1296_2A | 0.3685 | 0.3598 | 0.3802 | 0.3150 | 0.3272 | 0.3671 | 0.3704 |
| 8 | F1646_2A | 0.3263 | 0.3330 | 0.3603 | 0.3424 | 0.3632 | 0.3642 | 0.3115 |
| 9 | F1942_1A | 0.3390 | 0.3522 | 0.3684 | 0.3471 | 0.3699 | 0.3682 | 0.3223 |
| 10 | F2865_1A | 0.2962 | 0.3522 | 0.3595 | 0.4049 | 0.3820 | 0.3456 | 0.3613 |
| 11 | F11528_1A | 0.3517 | 0.3646 | 0.0109 | 0.4031 | 0.3618 | 0.3400 | 0.3377 |
| 12 | G922_1A | 0.3662 | 0.3836 | 0.4079 | 0.3841 | 0.3405 | 0.3711 | 0.3487 |
| 13 | G1822_2A | | 0.3347 | 0.3511 | 0.3984 | 0.3758 | 0.3404 | 0.2863 |
| 14 | H0421_2A | 0.5498 | | 0.3662 | 0.4055 | 0.3124 | 0.3488 | 0.3512 |
| 15 | I168_1A | 0.6751 | 0.6617 | | 0.4048 | 0.3595 | 0.3397 | 0.3424 |
| 16 | I198_2A | 0.7927 | 0.7821 | 0.8581 | | 0.3965 | 0.3671 | 0.3596 |
| 17 | I202_2A | 0.7121 | 0.5269 | 0.6518 | 0.8022 | | 0.3678 | 0.3865 |
| 18 | I208_1A | 0.5595 | 0.5955 | 0.5840 | 0.6628 | 0.6804 | | 0.3461 |
| 19 | I213_2A | 0.3525 | 0.5764 | 0.6156 | 0.6078 | 0.7171 | 0.5258 | |
| 20 | I218_2A | 0.5226 | 0.5227 | 0.5698 | 0.6745 | 0.6786 | 0.0679 | 0.5273 |
| 21 | I256_2A | 0.5183 | 0.6260 | 0.6820 | 0.6036 | 0.7402 | 0.6453 | 0.4909 |
| 22 | I280_1A | 0.7098 | 0.7107 | 0.6968 | 0.8582 | 0.6627 | 0.7940 | 0.7300 |
| 23 | I280_2A | 0.7209 | 0.7203 | 0.6985 | 0.8535 | 0.6710 | 0.7965 | 0.7394 |
| 24 | I283_1A | 0.7170 | 0.7165 | 0.7089 | 0.8652 | 0.6673 | 0.8044 | 0.7207 |
| 25 | I336_1A | 0.7397 | 0.6728 | 0.7029 | 0.8992 | 0.7843 | 0.7494 | 0.7504 |
| 26 | J0622_1A | 0.7978 | 0.7014 | 0.8715 | 0.0530 | 0.7461 | 0.7231 | 0.6563 |
| 27 | K15RE_1A | 0.5703 | 0.0090 | 0.6585 | 0.7789 | 0.5424 | 0.5951 | 0.5941 |
| 28 | K19RE_1A | 0.6667 | 0.6876 | 0.8030 | 0.5807 | 0.5457 | 0.6131 | 0.6231 |
| 29 | K32RE_2A | 0.5622 | 0.5096 | 0.4873 | 0.6913 | 0.5269 | 0.4551 | 0.4048 |
| 30 | N0233_2A | 0.4982 | 0.4743 | 0.5327 | 0.7560 | 0.4925 | 0.5626 | 0.5173 |
| 31 | G822_1A | 0.6830 | 0.7016 | 0.8359 | 0.5222 | 0.6013 | 0.7081 | 0.7010 |
| 32 | 86-028NP_2A | 0.4884 | 0.5609 | 0.5075 | 0.7881 | 0.6166 | 0.4154 | 0.4479 |
| 33 | strain 5_2A | 0.5676 | 0.7729 | 0.7768 | 0.7889 | 0.6991 | 0.6118 | 0.7492 |
| 34 | strain 12_2A | 0.5276 | 0.4858 | 0.5997 | 0.6980 | 0.6109 | 0.5600 | 0.4693 |
| 35 | AAr105_1A | 0.5609 | 0.5655 | 0.6982 | 0.7238 | 0.4959 | 0.6231 | 0.5898 |

Table 3.S2. Cluster 1 distances (con't)

| | | 20 | 21 | 22 | 23 | 24 | 25 |
|---|---|---|---|---|---|---|---|
| | **Strain_locus** | I218_2A | I256_2A | I280_1A | I280_2A | I283_1A | I336_1A |
| 1 | 63422_1A | 0.3659 | 0.3344 | 0.3732 | 0.3732 | 0.3753 | 0.3445 |
| 2 | F120_2A | 0.3585 | 0.3754 | 0.3456 | 0.3456 | 0.3508 | 0.3837 |
| 3 | F651_1A | 0.3579 | 0.3974 | 0.3379 | 0.3400 | 0.3400 | 0.3386 |
| 4 | F651_2A | 0.3611 | 0.3942 | 0.3379 | 0.3379 | 0.3421 | 0.3397 |
| 5 | F894_2A | 0.3449 | 0.3573 | 0.3513 | 0.3532 | 0.3522 | 0.2643 |
| 6 | F938_1A | 0.3226 | 0.3139 | 0.3661 | 0.3680 | 0.3680 | 0.3705 |
| 7 | F1296_2A | 0.3639 | 0.3459 | 0.3695 | 0.3715 | 0.3725 | 0.3874 |
| 8 | F1646_2A | 0.3462 | 0.0548 | 0.3526 | 0.3548 | 0.3548 | 0.3693 |
| 9 | F1942_1A | 0.3594 | 0.0432 | 0.3659 | 0.3658 | 0.3679 | 0.3821 |
| 10 | F2865_1A | 0.3490 | 0.3584 | 0.3543 | 0.3543 | 0.3563 | 0.2637 |
| 11 | F11528_1A | 0.3326 | 0.3650 | 0.3647 | 0.3654 | 0.3622 | 0.3681 |
| 12 | G922_1A | 0.3559 | 0.3498 | 0.3987 | 0.4007 | 0.4007 | 0.3921 |
| 13 | G1822_2A | 0.3312 | 0.3256 | 0.3682 | 0.3701 | 0.3711 | 0.3751 |
| 14 | H0421_2A | 0.3302 | 0.3461 | 0.3701 | 0.3720 | 0.3720 | 0.3602 |
| 15 | I168_1A | 0.3311 | 0.3662 | 0.3584 | 0.3583 | 0.3605 | 0.3681 |
| 16 | I198_2A | 0.3660 | 0.3497 | 0.4064 | 0.4062 | 0.4050 | 0.4204 |
| 17 | I202_2A | 0.3603 | 0.3660 | 0.3519 | 0.3540 | 0.3540 | 0.3877 |
| 18 | I208_1A | 0.0685 | 0.3715 | 0.3932 | 0.3941 | 0.3951 | 0.3983 |
| 19 | I213_2A | 0.3379 | 0.3259 | 0.3809 | 0.3827 | 0.3766 | 0.3851 |
| 20 | I218_2A | | 0.3607 | 0.3941 | 0.3960 | 0.3960 | 0.3949 |
| 21 | I256_2A | 0.6225 | | 0.3693 | 0.3691 | 0.3734 | 0.3850 |
| 22 | I280_1A | 0.8151 | 0.7373 | | 0.0000 | 0.0030 | 0.3699 |
| 23 | I280_2A | 0.8261 | 0.7362 | 0.0000 | | 0.0060 | 0.3719 |
| 24 | I283_1A | 0.8215 | 0.7525 | 0.0020 | 0.0044 | | 0.3730 |
| 25 | I336_1A | 0.7722 | 0.7525 | 0.6939 | 0.7037 | 0.7004 | |
| 26 | J0622_1A | 0.6870 | 0.5797 | 0.8230 | 0.8345 | 0.8296 | 0.8493 |
| 27 | K15RE_1A | 0.5413 | 0.6399 | 0.7195 | 0.7292 | 0.7348 | 0.6887 |
| 28 | K19RE_1A | 0.5998 | 0.6139 | 0.7934 | 0.8043 | 0.7998 | 0.8011 |
| 29 | K32RE_2A | 0.4591 | 0.2560 | 0.7513 | 0.7603 | 0.7561 | 0.6495 |
| 30 | N0233_2A | 0.5168 | 0.6428 | 0.6615 | 0.6608 | 0.6767 | 0.7069 |
| 31 | G822_1A | 0.6928 | 0.6012 | 0.7096 | 0.7197 | 0.7157 | 0.8318 |
| 32 | 86-028NP_2A | 0.3873 | 0.5426 | 0.7260 | 0.7354 | 0.7309 | 0.7712 |
| 33 | strain 5_2A | 0.6295 | 0.6028 | 0.6157 | 0.6246 | 0.6223 | 0.6660 |
| 34 | strain 12_2A | 0.4865 | 0.5756 | 0.6972 | 0.7064 | 0.7010 | 0.6923 |
| 35 | AAr105_1A | 0.6200 | 0.6993 | 0.6189 | 0.6282 | 0.6247 | 0.6365 |

Table 3.S2. Cluster 1 distances (con't).

| | | 26 | 27 | 28 | 29 | 30 | 31 |
|---|---|---|---|---|---|---|---|
| | **Strain_locus** | J0622_1A | K15RE_1A | K19RE_1A | K32RE_2A | N0233_2A | G822_1A |
| 1 | 63422_1A | 0.3687 | 0.3920 | 0.3246 | 0.3605 | 0.3503 | 0.3067 |
| 2 | F120_2A | 0.3972 | 0.3242 | 0.3698 | 0.3684 | 0.3302 | 0.3934 |
| 3 | F651_1A | 0.3940 | 0.3474 | 0.3596 | 0.3541 | 0.3695 | 0.3753 |
| 4 | F651_2A | 0.3972 | 0.3474 | 0.3628 | 0.3552 | 0.3665 | 0.3785 |
| 5 | F894_2A | 0.4104 | 0.3542 | 0.3641 | 0.3402 | 0.3435 | 0.3996 |
| 6 | F938_1A | 0.3573 | 0.3227 | 0.2401 | 0.3101 | 0.3236 | 0.3243 |
| 7 | F1296_2A | 0.2473 | 0.3651 | 0.0972 | 0.3488 | 0.3723 | 0.0244 |
| 8 | F1646_2A | 0.3231 | 0.3385 | 0.3492 | 0.2276 | 0.3675 | 0.3352 |
| 9 | F1942_1A | 0.3417 | 0.3543 | 0.3532 | 0.2370 | 0.3716 | 0.3543 |
| 10 | F2865_1A | 0.4158 | 0.3532 | 0.3686 | 0.3432 | 0.3444 | 0.4046 |
| 11 | F11528_1A | 0.4025 | 0.3650 | 0.3785 | 0.3057 | 0.3245 | 0.3831 |
| 12 | G922_1A | 0.3727 | 0.3891 | 0.2503 | 0.3469 | 0.3696 | 0.3220 |
| 13 | G1822_2A | 0.3889 | 0.3387 | 0.3624 | 0.3544 | 0.3114 | 0.3675 |
| 14 | H0421_2A | 0.3787 | 0.0049 | 0.3592 | 0.3226 | 0.3203 | 0.3661 |
| 15 | I168_1A | 0.4018 | 0.3666 | 0.3791 | 0.3017 | 0.3262 | 0.3848 |
| 16 | I198_2A | 0.0770 | 0.4055 | 0.3372 | 0.3850 | 0.3927 | 0.3242 |
| 17 | I202_2A | 0.3782 | 0.3166 | 0.3243 | 0.3068 | 0.3096 | 0.3359 |
| 18 | I208_1A | 0.3754 | 0.3467 | 0.3444 | 0.3185 | 0.3461 | 0.3716 |
| 19 | I213_2A | 0.3628 | 0.3532 | 0.3404 | 0.2933 | 0.3191 | 0.3693 |
| 20 | I218_2A | 0.3645 | 0.3354 | 0.3381 | 0.3179 | 0.3379 | 0.3650 |
| 21 | I256_2A | 0.3388 | 0.3482 | 0.3481 | 0.2313 | 0.3565 | 0.3491 |
| 22 | I280_1A | 0.4053 | 0.3725 | 0.3854 | 0.3521 | 0.3683 | 0.3737 |
| 23 | I280_2A | 0.4072 | 0.3744 | 0.3874 | 0.3541 | 0.3681 | 0.3757 |
| 24 | I283_1A | 0.4072 | 0.3764 | 0.3874 | 0.3541 | 0.3734 | 0.3757 |
| 25 | I336_1A | 0.4096 | 0.3637 | 0.3874 | 0.3524 | 0.3686 | 0.3907 |
| 26 | J0622_1A | | 0.3841 | 0.3322 | 0.3671 | 0.4017 | 0.2591 |
| 27 | K15RE_1A | 0.7242 | | 0.3644 | 0.3268 | 0.3203 | 0.3713 |
| 28 | K19RE_1A | 0.5650 | 0.7126 | | 0.3413 | 0.3524 | 0.0899 |
| 29 | K32RE_2A | 0.6511 | 0.5340 | 0.6232 | | 0.3400 | 0.3477 |
| 30 | N0233_2A | 0.8320 | 0.4811 | 0.6125 | 0.5746 | | 0.3798 |
| 31 | G822_1A | 0.3739 | 0.7269 | 0.0810 | 0.6473 | 0.7258 | |
| 32 | 86-028NP_2A | 0.8051 | 0.5901 | 0.6753 | 0.3994 | 0.5209 | 0.7637 |
| 33 | strain 5_2A | 0.7557 | 0.7988 | 0.6671 | 0.6052 | 0.7657 | 0.6245 |
| 34 | strain 12_2A | 0.6937 | 0.5016 | 0.3649 | 0.5299 | 0.4757 | 0.5397 |
| 35 | AAr105_1A | 0.7655 | 0.5873 | 0.5801 | 0.6633 | 0.0893 | 0.6888 |

Table 3.S2. Cluster 1 distances (con't)

| | | 32 | 33 | 34 | 35 |
|---|---|---|---|---|---|
| | **Strain_locus** | 86-028NP_2A | strain 5_2A | strain 12_2A | AAr105_1A |
| 1 | 63422_1A | 0.2773 | 0.2426 | 0.3828 | 0.3472 |
| 2 | F120_2A | 0.3724 | 0.3849 | 0.3473 | 0.3160 |
| 3 | F651_1A | 0.3976 | 0.3600 | 0.3312 | 0.3509 |
| 4 | F651_2A | 0.3996 | 0.3630 | 0.3333 | 0.3540 |
| 5 | F894_2A | 0.3514 | 0.3673 | 0.3481 | 0.3265 |
| 6 | F938_1A | 0.3200 | 0.3612 | 0.0470 | 0.3456 |
| 7 | F1296_2A | 0.3687 | 0.3337 | 0.3159 | 0.3691 |
| 8 | F1646_2A | 0.3289 | 0.3344 | 0.3256 | 0.3679 |
| 9 | F1942_1A | 0.3379 | 0.3511 | 0.3273 | 0.3796 |
| 10 | F2865_1A | 0.3563 | 0.3696 | 0.3543 | 0.3292 |
| 11 | F11528_1A | 0.2934 | 0.3768 | 0.3381 | 0.3615 |
| 12 | G922_1A | 0.3326 | 0.3450 | 0.3199 | 0.3641 |
| 13 | G1822_2A | 0.3185 | 0.3285 | 0.3237 | 0.3248 |
| 14 | H0421_2A | 0.3477 | 0.3709 | 0.3193 | 0.3340 |
| 15 | I168_1A | 0.2914 | 0.3762 | 0.3385 | 0.3626 |
| 16 | I198_2A | 0.3861 | 0.3870 | 0.3599 | 0.3867 |
| 17 | I202_2A | 0.3389 | 0.3494 | 0.3394 | 0.3009 |
| 18 | I208_1A | 0.2975 | 0.3606 | 0.3481 | 0.3682 |
| 19 | I213_2A | 0.3100 | 0.3910 | 0.3017 | 0.3432 |
| 20 | I218_2A | 0.2824 | 0.3565 | 0.3333 | 0.3577 |
| 21 | I256_2A | 0.3298 | 0.3481 | 0.3192 | 0.3678 |
| 22 | I280_1A | 0.3660 | 0.3381 | 0.3764 | 0.3482 |
| 23 | I280_2A | 0.3679 | 0.3401 | 0.3784 | 0.3503 |
| 24 | I283_1A | 0.3690 | 0.3411 | 0.3784 | 0.3503 |
| 25 | I336_1A | 0.3839 | 0.3532 | 0.3727 | 0.3483 |
| 26 | J0622_1A | 0.3885 | 0.3730 | 0.3540 | 0.3908 |
| 27 | K15RE_1A | 0.3518 | 0.3759 | 0.3237 | 0.3397 |
| 28 | K19RE_1A | 0.3515 | 0.3455 | 0.2463 | 0.3469 |
| 29 | K32RE_2A | 0.2856 | 0.3434 | 0.3206 | 0.3610 |
| 30 | N0233_2A | 0.3225 | 0.3730 | 0.3107 | 0.0862 |
| 31 | G822_1A | 0.3728 | 0.3367 | 0.3232 | 0.3745 |
| 32 | 86-028NP_2A | | 0.2625 | 0.3231 | 0.3435 |
| 33 | strain 5_2A | 0.3453 | | 0.3674 | 0.3506 |
| 34 | strain 12_2A | 0.5217 | 0.7360 | | 0.3359 |
| 35 | AAr105_1A | 0.6238 | 0.6796 | 0.5818 | |

Table 3.S3. Cluster 2 pairwise amino acid p-distances (above diagonal) and pairwise maximum composite likelihood genetic distances (below diagonal).

| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|---|
| | **Strain_locus** | G423_2A | I283_2A | F120_1A | F11528_2A | strain 12_1A | F11242_1A | F11242_2A |
| 1 | G423_2A | | 0.0000 | 0.2975 | 0.3020 | 0.2856 | 0.3241 | 0.3278 |
| 2 | I283_2A | 0.0000 | | 0.2975 | 0.3020 | 0.2856 | 0.3241 | 0.3278 |
| 3 | F120_1A | 0.3896 | 0.3896 | | 0.2338 | 0.2854 | 0.3311 | 0.3350 |
| 4 | F11528_2A | 0.3649 | 0.3649 | 0.3364 | | 0.2510 | 0.2811 | 0.2818 |
| 5 | strain 12_1A | 0.4699 | 0.4699 | 0.4019 | 0.2984 | | 0.2660 | 0.2697 |
| 6 | F11242_1A | 0.4293 | 0.4293 | 0.4771 | 0.2230 | 0.3027 | | 0.0026 |
| 7 | F11242_2A | 0.4307 | 0.4307 | 0.4786 | 0.2239 | 0.3038 | 0.0005 | |
| 8 | K32RE_1A | 0.5001 | 0.5001 | 0.4282 | 0.3662 | 0.3597 | 0.4308 | 0.4322 |
| 9 | I256_1A | 0.7040 | 0.7040 | 0.4851 | 0.4320 | 0.4161 | 0.4880 | 0.4894 |
| 10 | strain 72_1A | 0.5622 | 0.5622 | 0.5622 | 0.4240 | 0.4349 | 0.5106 | 0.5121 |
| 11 | J0622_2A | 0.6274 | 0.6274 | 0.7009 | 0.5006 | 0.5653 | 0.6551 | 0.6569 |
| 12 | AAr96_2A | 0.6563 | 0.6563 | 0.7440 | 0.5443 | 0.5847 | 0.6886 | 0.6904 |
| 13 | F894_1A | 0.6795 | 0.6795 | 0.6299 | 0.5861 | 0.6569 | 0.6725 | 0.6743 |
| 14 | F2865_2A | 0.6824 | 0.6824 | 0.6326 | 0.5887 | 0.6540 | 0.6696 | 0.6714 |
| 15 | I198_1A | 0.7357 | 0.7357 | 0.6987 | 0.6434 | 0.7021 | 0.7328 | 0.7347 |
| 16 | I202_1A | 0.7884 | 0.7884 | 0.8380 | 0.6504 | 0.7233 | 0.7574 | 0.7593 |
| 17 | F441_1A | 0.6207 | 0.6207 | 0.6920 | 0.6025 | 0.6382 | 0.6276 | 0.6293 |
| 18 | F1296_1A | 0.6191 | 0.6191 | 0.7951 | 0.6334 | 0.6561 | 0.6366 | 0.6383 |
| 19 | K19RE_2A | 0.6199 | 0.6199 | 0.8027 | 0.6258 | 0.6552 | 0.6340 | 0.6357 |
| 20 | I208_2A | 0.5901 | 0.5901 | 0.4839 | 0.4952 | 0.5041 | 0.5072 | 0.5087 |
| 21 | strain 5_1A | 0.5717 | 0.5717 | 0.5093 | 0.4930 | 0.5357 | 0.5155 | 0.5170 |
| 22 | I336_2A | 0.6188 | 0.6188 | 0.5776 | 0.5783 | 0.6162 | 0.6617 | 0.6634 |
| 23 | H0421_1A | 0.6094 | 0.6094 | 0.4883 | 0.4935 | 0.5151 | 0.5243 | 0.5258 |
| 24 | K15RE_2A | 0.6094 | 0.6094 | 0.4883 | 0.4935 | 0.5151 | 0.5243 | 0.5258 |

Table 3.S3. Cluster 2 distances (con't).

| | | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
|---|---|---|---|---|---|---|---|---|
| | **Strain_locus** | K32RE_1A | I256_1A | strain 72_1A | J0622_2A | AAr96_2A | F894_1A | F2865_2A |
| 1 | G423_2A | 0.3471 | 0.3992 | 0.3682 | 0.3902 | 0.3976 | 0.3680 | 0.3660 |
| 2 | I283_2A | 0.3471 | 0.3992 | 0.3682 | 0.3902 | 0.3976 | 0.3680 | 0.3660 |
| 3 | F120_1A | 0.3047 | 0.3387 | 0.3607 | 0.3829 | 0.3925 | 0.3585 | 0.3564 |
| 4 | F11528_2A | 0.2886 | 0.3568 | 0.3310 | 0.3571 | 0.3724 | 0.3798 | 0.3808 |
| 5 | strain 12_1A | 0.2759 | 0.3011 | 0.3046 | 0.3518 | 0.3591 | 0.3456 | 0.3413 |
| 6 | F11242_1A | 0.3093 | 0.3485 | 0.3404 | 0.3725 | 0.3836 | 0.3719 | 0.3681 |
| 7 | F11242_2A | 0.3119 | 0.3522 | 0.3427 | 0.3737 | 0.3849 | 0.3739 | 0.3701 |
| 8 | K32RE_1A | | 0.3168 | 0.3053 | 0.3569 | 0.3692 | 0.3484 | 0.3494 |
| 9 | I256_1A | 0.4883 | | 0.3802 | 0.3517 | 0.3605 | 0.3039 | 0.3020 |
| 10 | strain 72_1A | 0.4895 | 0.6754 | | 0.2684 | 0.2829 | 0.3760 | 0.3737 |
| 11 | J0622_2A | 0.6557 | 0.5553 | 0.3702 | | 0.0473 | 0.3513 | 0.3480 |
| 12 | AAr96_2A | 0.6953 | 0.5775 | 0.4292 | 0.0317 | | 0.3531 | 0.3520 |
| 13 | F894_1A | 0.7787 | 0.5290 | 0.6767 | 0.4599 | 0.4814 | | 0.0034 |
| 14 | F2865_2A | 0.7820 | 0.5266 | 0.6797 | 0.4607 | 0.4837 | 0.0005 | |
| 15 | I198_1A | 0.7850 | 0.5137 | 0.7160 | 0.5455 | 0.5580 | 0.3453 | 0.3436 |
| 16 | I202_1A | 0.5799 | 0.6713 | 0.6530 | 0.5090 | 0.5409 | 0.4415 | 0.4394 |
| 17 | F441_1A | 0.7734 | 0.6207 | 0.5762 | 0.3600 | 0.3867 | 0.3507 | 0.3525 |
| 18 | F1296_1A | 0.8105 | 0.6594 | 0.6331 | 0.3839 | 0.4292 | 0.3884 | 0.3902 |
| 19 | K19RE_2A | 0.8197 | 0.6567 | 0.6303 | 0.3797 | 0.4245 | 0.3854 | 0.3872 |
| 20 | I208_2A | 0.5669 | 0.6302 | 0.5070 | 0.5627 | 0.6091 | 0.5242 | 0.5265 |
| 21 | strain 5_1A | 0.5598 | 0.6226 | 0.4869 | 0.5634 | 0.6157 | 0.5390 | 0.5414 |
| 22 | I336_2A | 0.6112 | 0.6402 | 0.6231 | 0.7564 | 0.7711 | 0.5827 | 0.5853 |
| 23 | H0421_1A | 0.5505 | 0.4526 | 0.4303 | 0.5638 | 0.5967 | 0.5151 | 0.5175 |
| 24 | K15RE_2A | 0.5505 | 0.4526 | 0.4303 | 0.5638 | 0.5967 | 0.5151 | 0.5175 |

Table 3.S3. Cluster 2 pairwise distances (con't).

| | | 15 | 16 | 17 | 18 | 19 | 20 | 21 |
|---|---|---|---|---|---|---|---|---|
| | **Strain_locus** | I198_1A | I202_1A | F441_1A | F1296_1A | K19RE_2A | I208_2A | strain 5_1A |
| 1 | G423_2A | 0.3889 | 0.3872 | 0.3599 | 0.3382 | 0.3393 | 0.3420 | 0.3479 |
| 2 | I283_2A | 0.3889 | 0.3872 | 0.3623 | 0.3382 | 0.3393 | 0.3420 | 0.3479 |
| 3 | F120_1A | 0.3834 | 0.3978 | 0.3681 | 0.3765 | 0.3787 | 0.3203 | 0.3341 |
| 4 | F11528_2A | 0.3729 | 0.3635 | 0.3638 | 0.3623 | 0.3612 | 0.3302 | 0.3232 |
| 5 | strain 12_1A | 0.3818 | 0.3731 | 0.3647 | 0.3662 | 0.3673 | 0.3132 | 0.3199 |
| 6 | F11242_1A | 0.3818 | 0.3803 | 0.3670 | 0.3584 | 0.3595 | 0.3144 | 0.3158 |
| 7 | F11242_2A | 0.3819 | 0.3817 | 0.3682 | 0.3607 | 0.3618 | 0.3167 | 0.3169 |
| 8 | K32RE_1A | 0.3873 | 0.3383 | 0.3876 | 0.3692 | 0.3715 | 0.3349 | 0.3273 |
| 9 | I256_1A | 0.3445 | 0.3676 | 0.3638 | 0.3630 | 0.3641 | 0.3585 | 0.3693 |
| 10 | strain 72_1A | 0.3843 | 0.3516 | 0.3790 | 0.3863 | 0.3853 | 0.3463 | 0.3397 |
| 11 | J0622_2A | 0.3634 | 0.3294 | 0.3309 | 0.3187 | 0.3167 | 0.3479 | 0.3418 |
| 12 | AAr96_2A | 0.3646 | 0.3318 | 0.3403 | 0.3350 | 0.3330 | 0.3709 | 0.3600 |
| 13 | F894_1A | 0.2679 | 0.2804 | 0.3023 | 0.3048 | 0.3037 | 0.3360 | 0.3540 |
| 14 | F2865_2A | 0.2657 | 0.2781 | 0.3001 | 0.3027 | 0.3016 | 0.3326 | 0.3529 |
| 15 | I198_1A | | 0.0623 | 0.3076 | 0.2731 | 0.2720 | 0.3791 | 0.3641 |
| 16 | I202_1A | 0.0768 | | 0.3152 | 0.2636 | 0.2624 | 0.3781 | 0.3734 |
| 17 | F441_1A | 0.3937 | 0.4455 | | 0.2031 | 0.2031 | 0.3356 | 0.3722 |
| 18 | F1296_1A | 0.3697 | 0.3931 | 0.1891 | | 0.0059 | 0.3536 | 0.3624 |
| 19 | K19RE_2A | 0.3693 | 0.3926 | 0.1875 | 0.0023 | | 0.3558 | 0.3624 |
| 20 | I208_2A | 0.5629 | 0.6424 | 0.4651 | 0.5881 | 0.5874 | | 0.3041 |
| 21 | strain 5_1A | 0.5987 | 0.7223 | 0.6051 | 0.5776 | 0.5782 | 0.3922 | |
| 22 | I336_2A | 0.6288 | 0.8231 | 0.5873 | 0.7161 | 0.7133 | 0.4700 | 0.5392 |
| 23 | H0421_1A | 0.6190 | 0.7680 | 0.6138 | 0.6713 | 0.6719 | 0.3816 | 0.3462 |
| 24 | K15RE_2A | 0.6190 | 0.7680 | 0.6138 | 0.6713 | 0.6719 | 0.3816 | 0.3462 |

Table 3.S3. Cluster 2 pairwise distances (con't).

| | | 22 | 23 | 24 |
|---|---|---|---|---|
| | **Strain_locus** | I336_2A | H0421_1A | K15RE_2A |
| 1 | G423_2A | 0.3580 | 0.3715 | 0.3715 |
| 2 | I283_2A | 0.3580 | 0.3715 | 0.3715 |
| 3 | F120_1A | 0.3454 | 0.3383 | 0.3383 |
| 4 | F11528_2A | 0.3525 | 0.3680 | 0.3680 |
| 5 | strain 12_1A | 0.3525 | 0.3216 | 0.3216 |
| 6 | F11242_1A | 0.3680 | 0.3597 | 0.3597 |
| 7 | F11242_2A | 0.3692 | 0.3628 | 0.3628 |
| 8 | K32RE_1A | 0.3522 | 0.3337 | 0.3337 |
| 9 | I256_1A | 0.3818 | 0.3239 | 0.3239 |
| 10 | strain 72_1A | 0.3592 | 0.3040 | 0.3040 |
| 11 | J0622_2A | 0.4013 | 0.3477 | 0.3477 |
| 12 | AAr96_2A | 0.4013 | 0.3519 | 0.3519 |
| 13 | F894_1A | 0.3429 | 0.3246 | 0.3246 |
| 14 | F2865_2A | 0.3417 | 0.3226 | 0.3226 |
| 15 | I198_1A | 0.3513 | 0.3621 | 0.3621 |
| 16 | I202_1A | 0.3803 | 0.3804 | 0.3804 |
| 17 | F441_1A | 0.3613 | 0.3623 | 0.3623 |
| 18 | F1296_1A | 0.3874 | 0.3764 | 0.3764 |
| 19 | K19RE_2A | 0.3874 | 0.3764 | 0.3764 |
| 20 | I208_2A | 0.3134 | 0.2869 | 0.2869 |
| 21 | strain 5_1A | 0.3467 | 0.2903 | 0.2903 |
| 22 | I336_2A | | 0.3118 | 0.3118 |
| 23 | H0421_1A | 0.5200 | | 0.0000 |
| 24 | K15RE_2A | 0.5200 | 0.0000 | |

Table 3.S4. Cluster 3-4 pairwise amino acid p-distances (above diagonal) and pairwise maximum composite likelihood genetic distances (below diagonal).

| | | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|
| | **Strain_locus** | F1646_1A | F1942_2A | I218_1A | I249_1A | F1015_1A | F1015_2A |
| 1 | F1646_1A | | 0.0105 | 0.3018 | 0.2994 | 0.3956 | 0.3923 |
| 2 | F1942_2A | 0.003 | | 0.3083 | 0.3004 | 0.3967 | 0.3989 |
| 3 | I218_1A | 0.381 | 0.387 | | 0.1507 | 0.4208 | 0.4178 |
| 4 | I249_1A | 0.354 | 0.354 | 0.110 | | 0.4426 | 0.4426 |
| 5 | F1015_1A | 0.705 | 0.696 | 0.742 | 0.791 | | 0.0026 |
| 6 | F1015_2A | 0.695 | 0.706 | 0.731 | 0.789 | 0.002 | |
| 7 | 63422_2A | 0.703 | 0.694 | 0.670 | 0.761 | 0.180 | 0.184 |
| 8 | G922_2A | 0.667 | 0.659 | 0.643 | 0.748 | 0.214 | 0.219 |
| 9 | I213_1A | 0.531 | 0.524 | 0.577 | 0.625 | 0.564 | 0.573 |
| 10 | PittEE_1A | 0.551 | 0.558 | 0.668 | 0.661 | 0.568 | 0.562 |
| 11 | 86-028NP_1A | 0.534 | 0.541 | 0.653 | 0.679 | 0.557 | 0.551 |
| 12 | A950006_A | 0.535 | 0.541 | 0.655 | 0.681 | 0.559 | 0.552 |

Table 3.S4. Cluster 3-4 distances (con't).

| | | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|
| | **Strain_locus** | 63422_2A | G922_2A | I213_1A | PittEE_1A | 86-028NP_1A | A950006_A |
| 1 | F1646_1A | 0.4125 | 0.4124 | 0.3749 | 0.3863 | 0.3719 | 0.3709 |
| 2 | F1942_2A | 0.4147 | 0.4135 | 0.3727 | 0.3897 | 0.3753 | 0.3743 |
| 3 | I218_1A | 0.3949 | 0.3956 | 0.3568 | 0.4180 | 0.4047 | 0.4020 |
| 4 | I249_1A | 0.4270 | 0.4395 | 0.3937 | 0.4217 | 0.4217 | 0.4124 |
| 5 | F1015_1A | 0.1867 | 0.2390 | 0.3742 | 0.3764 | 0.3623 | 0.3615 |
| 6 | F1015_2A | 0.1885 | 0.2420 | 0.3761 | 0.3744 | 0.3602 | 0.3595 |
| 7 | 63422_2A | | 0.2236 | 0.3902 | 0.3992 | 0.3871 | 0.3838 |
| 8 | G922_2A | 0.188 | | 0.3563 | 0.3913 | 0.3877 | 0.3873 |
| 9 | I213_1A | 0.597 | 0.511 | | 0.2225 | 0.2035 | 0.2134 |
| 10 | PittEE_1A | 0.626 | 0.587 | 0.149 | | 0.0353 | 0.0322 |
| 11 | 86-028NP_1A | 0.601 | 0.567 | 0.146 | 0.015 | | 0.0010 |
| 12 | A950006_A | 0.603 | 0.568 | 0.147 | 0.015 | 0.001 | |

## Literature Cited


1. Barenkamp, S. J. 1996. Immunization with high-molecular-weight adhesion proteins of nontypeable *Haemophilus influenzae* modifies experimental otitis media in chinchillas. Infect Immun 64:1246-51.
2. Barenkamp, S. J. 1992. Outer membrane proteins and lipopolysaccharides of nontypeable *Haemophilus influenzae*. J Infect Dis 165 Suppl 1:S181-4.
3. Barenkamp, S. J. 1986. Protection by serum antibodies in experimental nontypable Haemophilus influenzae otitis media. Infect Immun 52:572-8.
4. Barenkamp, S. J., and F. F. Bodor. 1990. Development of serum bactericidal activity following nontypable *Haemophilus influenzae* acute otitis media. Pediatr Infect Dis J 9:333-9.
5. Barenkamp, S. J., and E. Leininger. 1992. Cloning, expression, and DNA sequence analysis of genes encoding nontypeable *Haemophilus influenzae* high-molecular-weight surface-exposed proteins related to filamentous hemagglutinin of *Bordetella pertussis*. Infect Immun 60:1302-13.
6. Barenkamp, S. J., and J. W. St Geme, 3rd. 1994. Genes encoding high-molecular-weight adhesion proteins of nontypeable *Haemophilus influenzae* are part of gene clusters. Infect Immun 62:3320-8.
7. Bruen, T. C., H. Philippe, and D. Bryant. 2006. A simple and robust statistical test for detecting the presence of recombination. Genetics 172:2665-81.
8. Bryant, D., and V. Moulton. 2004. Neighbor-net: an agglomerative method for the construction of phylogenetic networks. Mol Biol Evol 21:255-65.
9. Buscher, A. Z., K. Burmeister, S. J. Barenkamp, and J. W. St Geme, 3rd. 2004. Evolutionary and functional relationships among the nontypeable *Haemophilus influenzae* HMW family of adhesins. J Bacteriol 186:4209-17.
10. Cardines, R., M. Giufre, P. Mastrantonio, M. L. Ciofi degli Atti, and M. Cerquetti. 2007. Nontypeable Haemophilus influenzae meningitis in children: phenotypic and genotypic characterization of isolates. Pediatr Infect Dis J 26:577-82.
11. Cholon, D. M., D. Cutter, S. K. Richardson, S. Sethi, T. F. Murphy, D. C. Look, and J. W. St Geme, 3rd. 2008. Serial isolates of persistent *Haemophilus influenzae* in patients with chronic obstructive pulmonary disease express diminishing quantities of the HMW1 and HMW2 adhesins. Infect Immun 76:4463-8.
12. Clemans, D. L., C. F. Marrs, M. Patel, M. Duncan, and J. R. Gilsdorf. 1998. Comparative analysis of Haemophilus influenzae hifA (pilin) genes. Infect Immun 66:656-63.
13. Cody, A. J., D. Field, E. J. Feil, S. Stringer, M. E. Deadman, A. G. Tsolaki, B. Gratz, V. Bouchet, R. Goldstein, D. W. Hood, and E. R. Moxon. 2003. High rates of recombination in otitis media isolates of non-typeable Haemophilus influenzae. Infect Genet Evol 3:57-66.

14. Connor, T. R., J. Corander, and W. P. Hanage. 2012. Population subdivision and the detection of recombination in non-typable Haemophilus influenzae. Microbiology 158:2958-64.
15. Croxtall, J. D., and G. M. Keating. 2009. Pneumococcal polysaccharide protein D-conjugate vaccine (Synflorix; PHiD-CV). Paediatr Drugs 11:349-57.
16. Davis, J., A. L. Smith, W. R. Hughes, and M. Golomb. 2001. Evolution of an autotransporter: domain shuffling and lateral transfer from pathogenic Haemophilus to Neisseria. J Bacteriol 183:4626-35.
17. Dawid, S., S. J. Barenkamp, and J. W. St Geme, 3rd. 1999. Variation in expression of the *Haemophilus influenzae* HMW adhesins: a prokaryotic system reminiscent of eukaryotes. Proc Natl Acad Sci U S A 96:1077-82.
18. Dawid, S., S. Grass, and J. W. St Geme, 3rd. 2001. Mapping of binding domains of nontypeable *Haemophilus influenzae* HMW1 and HMW2 adhesins. Infect Immun 69:307-14.
19. De Wals, P., B. Lefebvre, F. Defay, G. Deceuninck, and N. Boulianne. 2012. Invasive pneumococcal diseases in birth cohorts vaccinated with PCV-7 and/or PHiD-CV in the province of Quebec, Canada. Vaccine 30:6416-20.
20. Dinleyici, E. C., and Z. A. Yargic. 2009. Pneumococcal conjugated vaccine: PHiD-CV. Expert Rev Anti Infect Ther 7:1063-74.
21. Ecevit, I. Z., K. W. McCrea, C. F. Marrs, and J. R. Gilsdorf. 2005. Identification of new hmwA alleles from nontypeable *Haemophilus influenzae*. Infect Immun 73:1221-5.
22. Ecevit, I. Z., K. W. McCrea, M. M. Pettigrew, A. Sen, C. F. Marrs, and J. R. Gilsdorf. 2004. Prevalence of the *hifBC, hmw1A, hmw2A, hmwC*, and hia Genes in *Haemophilus influenzae* Isolates. J Clin Microbiol 42:3065-72.
23. Edgar, R. C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792-7.
24. Erwin, A. L., K. L. Nelson, T. Mhlanga-Mutangadura, P. J. Bonthuis, J. L. Geelhood, G. Morlin, W. C. Unrath, J. Campos, D. W. Crook, M. M. Farley, F. W. Henderson, R. F. Jacobs, K. Muhlemann, S. W. Satola, L. van Alphen, M. Golomb, and A. L. Smith. 2005. Characterization of genetic and phenotypic diversity of invasive nontypeable Haemophilus influenzae. Infect Immun 73:5853-63.
25. Erwin, A. L., S. A. Sandstedt, P. J. Bonthuis, J. L. Geelhood, K. L. Nelson, W. C. Unrath, M. A. Diggle, M. J. Theodore, C. R. Pleatman, E. A. Mothershed, C. T. Sacchi, L. W. Mayer, J. R. Gilsdorf, and A. L. Smith. 2008. Analysis of genetic relatedness of *Haemophilus influenzae* isolates by multilocus sequence typing. J Bacteriol 190:1473-83.
26. Felsenstein, J. 1985. Confidence limits on phylogenies: An approach using the bootstrap. Evolution 39:783-791.
27. Fleischmann, R. D., M. D. Adams, O. White, R. A. Clayton, E. F. Kirkness, A. R. Kerlavage, C. J. Bult, J. F. Tomb, B. A. Dougherty, J. M. Merrick, and et al. 1995.

Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science 269:496-512.
28. Forsgren, A., K. Riesbeck, and H. Janson. 2008. Protein D of Haemophilus influenzae: a protective nontypeable H. influenzae antigen and a carrier for pneumococcal conjugate vaccines. Clin Infect Dis 46:726-31.
29. Giufre, M., A. Carattoli, R. Cardines, P. Mastrantonio, and M. Cerquetti. 2008. Variation in expression of HMW1 and HMW2 adhesins in invasive nontypeable *Haemophilus influenzae* isolates. BMC Microbiol 8:83.
30. Giufre, M., M. Muscillo, P. Spigaglia, R. Cardines, P. Mastrantonio, and M. Cerquetti. 2006. Conservation and diversity of HMW1 and HMW2 adhesin binding domains among invasive nontypeable *Haemophilus influenzae* isolates. Infect Immun 74:1161-70.
31. Grass, S., A. Z. Buscher, W. E. Swords, M. A. Apicella, S. J. Barenkamp, N. Ozchlewski, and J. W. St Geme, 3rd. 2003. The *Haemophilus influenzae* HMW1 adhesin is glycosylated in a process that requires HMW1C and phosphoglucomutase, an enzyme involved in lipooligosaccharide biosynthesis. Mol Microbiol 48:737-51.
32. Grass, S., and J. W. St Geme, 3rd. 2000. Maturation and secretion of the non-typable *Haemophilus influenzae* HMW1 adhesin: roles of the N-terminal and C-terminal domains. Mol Microbiol 36:55-67.
33. Hiltke, T. J., A. T. Schiffmacher, A. J. Dagonese, S. Sethi, and T. F. Murphy. 2003. Horizontal transfer of the gene encoding outer membrane protein P2 of nontypeable Haemophilus influenzae, in a patient with chronic obstructive pulmonary disease. J Infect Dis 188:114-7.
34. Hogg, J. S., F. Z. Hu, B. Janto, R. Boissy, J. Hayes, R. Keefe, J. C. Post, and G. 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 Biol 8:R103.
35. Huson, D. H., and D. Bryant. 2006. Application of phylogenetic networks in evolutionary studies. Mol Biol Evol 23:254-67.
36. Lacross, N. C., C. F. Marrs, and J. R. Gilsdorf. 2013. Population Structure in Nontypeable Haemophilus influenzae. Infect Genet Evol.
37. Loytynoja, A., and N. Goldman. 2010. webPRANK: a phylogeny-aware multiple sequence aligner with interactive alignment browser. BMC Bioinformatics 11:579.
38. Martin, K., G. Morlin, A. Smith, A. Nordyke, A. Eisenstark, and M. Golomb. 1998. The tryptophanase gene cluster of Haemophilus influenzae type b: evidence for horizontal gene transfer. J Bacteriol 180:107-18.
39. Meats, E., E. J. Feil, S. Stringer, A. J. Cody, R. Goldstein, J. S. Kroll, T. Popovic, and B. G. Spratt. 2003. Characterization of encapsulated and noncapsulated *Haemophilus influenzae* and determination of phylogenetic relationships by multilocus sequence typing. J Clin Microbiol 41:1623-36.

40. Musser, J. M., S. J. Barenkamp, D. M. Granoff, and R. K. Selander. 1986. Genetic relationships of serologically nontypable and serotype b strains of *Haemophilus influenzae*. Infect Immun 52:183-91.
41. Nei, M., and S. Kumar. 2000. Molecular Evolution and Phylogenetics. Oxford University Press, New York.
42. Pichichero, M. E., R. Kaur, J. R. Casey, A. Sabirov, M. N. Khan, and A. Almudevar. 2010. Antibody response to Haemophilus influenzae outer membrane protein D, P6, and OMP26 after nasopharyngeal colonization and acute otitis media in children. Vaccine 28:7184-92.
43. Poulsen, K., J. Reinholdt, and M. Kilian. 1992. A comparative genetic study of serologically distinct Haemophilus influenzae type 1 immunoglobulin A1 proteases. J Bacteriol 174:2913-21.
44. Power, P. M., W. A. Sweetman, N. J. Gallacher, M. R. Woodhall, G. A. Kumar, E. R. Moxon, and D. W. Hood. 2009. Simple sequence repeats in *Haemophilus influenzae*. Infect Genet Evol 9:216-28.
45. Prymula, R., I. Hanovcova, M. Splino, P. Kriz, J. Motlova, V. Lebedova, P. Lommel, E. Kaliskova, T. Pascal, D. Borys, and L. Schuerman. 2011. Impact of the 10-valent pneumococcal non-typeable Haemophilus influenzae Protein D conjugate vaccine (PHiD-CV) on bacterial nasopharyngeal carriage. Vaccine 29:1959-67.
46. Schuerman, L., D. Borys, B. Hoet, A. Forsgren, and R. Prymula. 2009. Prevention of otitis media: now a reality? Vaccine 27:5748-54.
47. Shen, K., P. Antalis, J. Gladitz, S. Sayeed, A. Ahmed, S. Yu, J. Hayes, S. Johnson, B. Dice, R. Dopico, R. Keefe, B. Janto, W. Chong, J. Goodwin, R. M. Wadowsky, G. Erdos, J. C. Post, G. D. Ehrlich, and F. Z. Hu. 2005. Identification, distribution, and expression of novel genes in 10 clinical isolates of nontypeable *Haemophilus influenzae*. Infect Immun 73:3479-91.
48. Solnick, J. V., L. M. Hansen, N. R. Salama, J. K. Boonjakuakul, and M. Syvanen. 2004. Modification of Helicobacter pylori outer membrane protein expression during experimental infection of rhesus macaques. Proc Natl Acad Sci U S A 101:2106-11.
49. St Geme, J. W., 3rd. 1994. The HMW1 adhesin of nontypeable *Haemophilus influenzae* recognizes sialylated glycoprotein receptors on cultured human epithelial cells. Infect Immun 62:3881-9.
50. St Geme, J. W., 3rd, and D. Cutter. 1995. Evidence that surface fibrils expressed by Haemophilus influenzae type b promote attachment to human epithelial cells. Mol Microbiol 15:77-85.
51. St Geme, J. W., 3rd, and D. Cutter. 1996. Influence of pili, fibrils, and capsule on in vitro adherence by Haemophilus influenzae type b. Mol Microbiol 21:21-31.
52. St Geme, J. W., 3rd, M. L. de la Morena, and S. Falkow. 1994. A *Haemophilus influenzae* IgA protease-like protein promotes intimate interaction with human epithelial cells. Mol Microbiol 14:217-33.

53. St Geme, J. W., 3rd, S. Falkow, and S. J. Barenkamp. 1993. High-molecular-weight proteins of nontypable *Haemophilus influenzae* mediate attachment to human epithelial cells. Proc Natl Acad Sci U S A 90:2875-9.
54. St Geme, J. W., 3rd, and S. Grass. 1998. Secretion of the Haemophilus influenzae HMW1 and HMW2 adhesins involves a periplasmic intermediate and requires the HMWB and HMWC proteins. Mol Microbiol 27:617-30.
55. St Geme, J. W., 3rd, V. V. Kumar, D. Cutter, and S. J. Barenkamp. 1998. Prevalence and distribution of the *hmw* and *hia* genes and the HMW and Hia adhesins among genetically diverse strains of nontypeable *Haemophilus influenzae*. Infect Immun 66:364-8.
56. Swords, W. E., B. A. Buscher, K. Ver Steeg Ii, A. Preston, W. A. Nichols, J. N. Weiser, B. W. Gibson, and M. A. Apicella. 2000. Non-typeable Haemophilus influenzae adhere to and invade human bronchial epithelial cells via an interaction of lipooligosaccharide with the PAF receptor. Mol Microbiol 37:13-27.
57. Talarico, S., S. E. Whitefield, J. Fero, R. Haas, and N. R. Salama. 2012. Regulation of Helicobacter pylori adherence by gene conversion. Mol Microbiol 84:1050-61.
58. Tamura, K., M. Nei, and S. Kumar. 2004. Prospects for inferring very large phylogenies by using the neighbor-joining method. Proc Natl Acad Sci U S A 101:11030-5.
59. Tamura, K., D. Peterson, N. Peterson, G. Stecher, M. Nei, and S. Kumar. 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol 28:2731-9.
60. van Alphen, L., D. A. Caugant, B. Duim, M. O'Rourke, and L. D. Bowler. 1997. Differences in genetic diversity of nonecapsulated *Haemophilus influenzae* from various diseases. Microbiology 143 ( Pt 4):1423-31.
61. van den Bergh, M. R., J. Spijkerman, K. M. Swinnen, N. A. Francois, T. G. Pascal, D. Borys, L. Schuerman, E. P. Ijzerman, J. P. Bruin, A. van der Ende, R. H. Veenhoven, and E. A. Sanders. 2013. Effects of the 10-Valent Pneumococcal Nontypeable Haemophilus influenzae Protein D-Conjugate Vaccine on Nasopharyngeal Bacterial Colonization in Young Children: A Randomized Controlled Trial. Clin Infect Dis 56:e30-9.
62. van Schilfgaarde, M., P. van Ulsen, P. Eijk, M. Brand, M. Stam, J. Kouame, L. van Alphen, and J. Dankert. 2000. Characterization of adherence of nontypeable *Haemophilus influenzae* to human epithelial cells. Infect Immun 68:4658-65.
63. Winter, L. E., and S. J. Barenkamp. 2006. Antibodies specific for the high-molecular-weight adhesion proteins of nontypeable *Haemophilus influenzae* are opsonophagocytic for both homologous and heterologous strains. Clin Vaccine Immunol 13:1333-42.
64. Winter, L. E., and S. J. Barenkamp. 2010. Construction and immunogenicity of recombinant adenovirus vaccines expressing the HMW1/HMW2 or Hia adhesion proteins of nontypeable Haemophilus influenzae. Clin Vaccine Immunol.
65. Yang, Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J Mol Evol 39:306-14.

## Chapter 4

# Nontypeable *Haemophilus influenzae* High Molecular Weight Adhesin Molecular Evolution is Driven by Positive Selection[3]


## Abstract

**Background.** Nontypeable *Haemophilus influenzae* (NTHi) is a human-specific commensal that resides within the pharynx. NTHi high molecular weight adhesins (HMWs) mediate attachment to the respiratory epithelium where they interact with the host immune system eliciting a strong humoral response. The *hmw* locus is present in two copies on the NTHi chromosome. *hmwA*, which encodes HMW, displays marked sequence diversity, undergoes recombination, and is phase variable. These mechanisms presumably interact, as the bacterium attempts to evade host immune responses. In this study, we specifically tested for evidence of positive selection acting on *hmwA*.

**Methods.** Two different methods for measuring selective pressure operating on a gene, both of which account for recombination, were employed: (1) MEME, a phylogenetic based maximum likelihood method and (2) omegaMap, a population genetics approximation of the coalescent with recombination. First, we analyzed the HMW-family of adhesins, that is, all *hmw1A* and *hmw2A* sequences combined, for evidence of positive selection in a single analysis. Next, two *hmwA* sequence clusters, Cluster 1 and Cluster 2-3, were defined based upon phylogenetic network analyses and analyzed separately for evidence of positive selection. Positively selected amino acids (AAs) were mapped onto predicted 3-dimensional models of the prototypic NTHi strain 12 HMW1 and HMW2 adhesins.


[3] This work is being prepared for submission with the following authors: Davis GS, Marrs CF, and Gildforf JR.

**Results.** Phylogenetic network analysis identified three distinct *hmwA* sequence clusters. Within the HMW-family of adhesins, considering all *hmwA* sequence clusters in a single analysis, approximately 50 AAs were identified as being under positive selection. All positively selected AAs were located within the mature HMW protein, and, the majority were within the binding domain region. When analyzed by *hmwA* sequence Cluster, approximately 22 and 30 AAs were under positive selection within the Cluster 1 and Cluster 2-3 *hmwA* sequences, respectively. The majority of AAs under positive selection were in regions predicted to be hydrophilic and consisting of either turns or loops.

**Conclusion.** The evolution of NTHi *hmwA* is driven, in part, by positive selection. The positively selected amino acids are restricted to the mature by HMW protein and are largely localized to the core-binding domain.

## Introduction

Nontypeable *Haemophilus influenzae* (NTHi), a human restricted commensal that normally resides within the pharynx, is capable of causing localized respiratory infections, such as acute otitis media (AOM) and sinusitis, among otherwise healthy individuals as well as invasive infections and bronchitis in patients with chronic obstructive pulmonary disease. One of the first steps in NTHi colonization, and pathogenesis, is adherence to the host epithelium, which is mediated by a number of surface exposed pili and non-pilus adhesins, including the high molecular weight (HMW) adhesins (43-46, 49).

The HMW adhesins are encoded by the *hmw* locus which consists of *hmwA, hmwB,* and *hmwC* (2, 3). The functional HMW adhesin is encoded by *hmwA* whereas *hmwB* and *hmwC* encode proteins necessary for adhesin maturation, glycosylation, and secretion (2, 3, 20, 22, 47). HMWs, as detected by Western blotting, are present in approximately 75% of all NTHi and, in nearly all strains tested, the *hmw* locus is present in two copies at conserved but unlinked locations on the NTHi chromosome (4, 8). The *hmw* loci, first characterized in NTHi strain 12, are designated *hmw1* and *hmw2* and their chromosomal locations are specified with respect to *H. influenzae* strain Rd as being downstream of HI1679 and HI1598, respectively (2, 8) (Fig. 4.1).

Mature HMW adhesins display marked amino acid diversity both within and between strains (2, 8, 11, 18). The pairwise amino acid similarity of mature HMW adhesin varies from approximately 63 – 72% and amino acid identity from 52 – 62%, and the adhesins can also exhibit length differences of approximately 17% (8, 11). The HMW binding domain, which is localized to an approximately 400 amino acid region of the mature protein and directly interacts with epithelial cells (11), displays even greater amino acid diversity with pairwise similarities and identities as low as 50% and 35%, respectively (8, 11).

HMW amino acid diversity potentially serves two roles. First, it defines tissue tropism. The two copies of HMW encoded by a single strain confer different *in vitro* adherence characteristics when tested against a panel of human epithelial cell lines (8, 46). The HMW-family of adhesins have been characterized by their differing *in vitro* adherence patterns relative to the proteins encoded by prototypic NTHi strain 12 *hmw1,* located at the HI1679 position, and

*hmw2*, located at the HI1598 position (chromosomal locations are with reference to Strain Rd) (8, 11).

The HMW adhesins, when present, are the immunodominant outer membrane protein (1). Amino acid diversity, therefore, maintains population level HMW adhesin diversity which may play a role in immune evasion, for example, anti-HMW antibodies offer varying degrees of cross protection against heterologous NTHi strains (52). The HMW adhesin sequence diversity may thus be a consequence of immune driven selective pressure, such as for diversifying selection.

The purpose of this study was to test for evidence of positive (*e.g.*, diversifying) selection acting on the mature HMW adhesins. Using two different methods for measuring selective pressure operating on a gene, both of which account for recombination, we first tested for evidence of positive selection acting within the HMW-family of adhesins. We then divided the HMW-family of adhesins into two mutually exclusive sequence clusters and tested each cluster separately to allow for variation in selective pressure between clusters. We found evidence of positive selection in the HMW adhesin family as well as in both *hmwA* sequence clusters. In all three analyses, the majority of positively selected codons were localized to the regions in and around the binding domain of the mature HMW adhesin. Our results suggest that HMW adhesin diversity is due, at least in part, to the action of positive selection.

## Materials and Methods

***hmwA* nucleotide sequences.** Nontypeable *Haemophilus influenzae hmwA* nucleotide sequences used in this study (Table 4.1) were obtained from GenBank. Sequences were limited to those that covered either the full-length *hmwA*, including the signal sequence region, or the entire mature binding protein.

**Sequence alignments.** *hmwA* sequence alignments were done in two stages. First, *hmwA* nucleotide sequences were converted to amino acids using the computer package MEGA v.5 (51) and aligned using the MUSCLE algorithm (14), converted back to the original nucleotide sequence and then used to estimate a neighbor joining (NJ) tree in MEGA v.5 (51). Next, *hmwA* sequences were aligned with the computer program PRANKSTER (36) using the NJ tree as the initial guide tree; again, nucleotides (nts) were converted to amino acids, aligned, and then converted back to nts. Finally, a second PRANKSTER (36) alignment was conducted using a guide tree estimated from the previous alignment; this second alignment was used for all subsequent analyses. Four sequence alignments were generated: (1) thirteen full-length *hmwA* sequences, (2) fifteen *hmwA* mature peptide sequences, (3) eight *hmwA* mature peptide sequences defined as Cluster 2-3 sequences, and (4) seven *hmwA* mature peptide sequences defined as a Cluster 1 sequences. For all subsequent analyses, gaps were deleted from the nucleotide alignments.

**Genetic distances and tests for recombination.** Genetic distances were estimated using the maximum composite likelihood (MCL) method (50) with rate variation among sites modeled with a gamma distribution. Standard error estimates of MCL genetic distance were estimated by a bootstrapping method using 1000 replicates. The value of the gamma distribution shape parameter, genetic distances, and variances were estimated in MEGA5 (51). Prior to conducting phylogenetic analyses, sequence alignments were tested for substitution saturation using the computer program DAMBE (57, 58). Sequence alignments were tested for evidence of recombination with the PHI test (6), implemented in SplitsTree4 (27).

**Phylogenetic analyses.** Evolutionary relationships among *hmwA* were inferred using the maximum likelihood method (38) assuming a general time reversible (GTR) model of sequence evolution with evolutionary rate differences between sites modeled with a discrete gamma

distribution (+G) (59) with allowance for invariant sites (+I) (16, 17). Consensus trees were inferred from 1000 bootstrapping replicates (15). All codon positions were included for analysis, alignment gaps were completely removed, and maximum likelihood phylogenetic analyses were implemented in MEGA v.5 (51). To account for inconsistencies in the phylogenetic signals, such as those introduced by recombination events, phylogenetic networks, based upon MCL distances, were constructed using the Neighbor-Net method (7) implemented in SplitsTree4 (27).

**Positive selection.** Positive selection is inferred when the ratio of the rates of nonsynonymous (dN) to synonymous (dS) substitutions, defined as omega ($\omega = dN/dS$), at a codon is significantly greater than one. Two different methods for measuring selective pressure operating on a gene, both of which account for recombination, were employed: (1) a phylogenetic based maximum likelihood method (37) and (2) a population genetics approximation of the coalescent with recombination (53).

Within the maximum likelihood framework, a mixed effects model of evolution (MEME) was used that allowed for rate variation from site-to-site along the sequence as well as variation among lineages at each site (37). Using MEME, it is possible to identify amino acids that are subject to episodic positive selection. Prior to implementing MEME, sequence alignments were screened for recombination breakpoints using the genetic algorithms for recombination detection (GARD) method (30, 31). GARD identifies potential recombination breakpoints in the sequences and defines non-recombinant blocks of sequence. To account for recombination, MEME analyzes non-recombinant blocks, identified by GARD, individually for evidence of positive selection. Both GARD and MEME were accessed via the Datamonkey public server (www.datamonkey.org).

To detect positive selection, we used omegaMap (53), a Bayesian method based on an approximation of the coalescent that can detect positive selection in the presence of recombination. Analyses were conducted on all four sequence alignments. Equilibrium codon frequencies, excluding stop codons, were estimated from NTHi 86-028NP complete genome. Ten product of approximate conditionals likelihood orderings were chosen for each analysis (53). Analyses were run for 1,000,000 iterations and a burn-in of 200,000 iterations were removed from each run; three runs were completed per sequence alignment. Objective priors were specified following recommendations provided in the omegaMap documentation; improper

inverse distributions were specified on parameters $\mu$, $\kappa$, and $\varphi$ and inverse distributions for $\omega$ and $\rho$. Window sizes of rBlock = 30 and oBlock = 5 were used for all analyses except for the HMW2 analysis that was run with rBlock = oBlock = 5. Output files were assessed for convergence and analyzed with the program R (www.r-project.org). Summary files were generated using the computer program *summarize* (53).

**Hydrophathy plots.** Kyte & Doolittle hydrophobicity plots (33) and Hopp-Wood hydrophilicity plots (26), using a window size of nine amino acids, were generated for NTHi strain 12 HMW1 and HMW2 mature proteins using the ProtScale tool available at the ExPASy Bioinformatics Resource Portal (www.expasy.org). Amino acids with a Kyte-Doolittle score > 0 are considered hydrophobic whereas amino acids with a Hopp-Woods score > 0 are considered hydrophilic.

**Protein structure predictions.** The I-TASSER server (40, 60) was used to predict three-dimensional structural models of NTHi strain 12 HMW1 and HMW2 mature proteins. Models were visualized with the the molecular graphics and modeling programs PyMOL version 1.3 (http://www.pymol.org/pymol) (41) and YASARA View (www.yasara.org) (32).

## Results

**Full length *hmwA* analyses.** Thirteen full-length *hmwA* sequences in GenBank (Table 4.1) that ranged in length from 4407 to 4803 nts with an average length of 4599 were used in this analysis. After aligning the sequences and removing gaps, 3609 nts remained in the final full-length *hmwA* alignment. The overall mean genetic distance, using an estimated gamma shape parameter = 0.3135, was 0.481 (S.E. = 0.022) base substitutions per site and pairwise distances ranged from 0.004 to 0.710 base substitutions per site (Table 4.S1). The observed index of saturation, $I_{ss}$, was significantly smaller than the $I_{ss}$ critical value, $I_{ss.c}$, ($P < 0.001$) (57, 58) indicating that the sequences have not reached substitution saturation and were suitable for phylogenetic analyses. The PHI test for recombination found statistically significant ($P < 0.01$) evidence of *hmwA* recombination. The *hmwA* alignment was also screened for evidence of recombination with the GARD method and 15 statistically significant ($P < 0.1$) recombination break points were identified (Table 4.S2).

The presence of recombination means that *hmwA* relationships cannot be accurately represented with a single phylogenetic tree. Instead, phylogenetic networks were constructed using the Neighbor-Net method (7). The *hmwA* Neighbor-Net displays the relationships among sequences while allowing for the visualization of conflicting phylogenetic signals, such as that arising from recombination, as boxes in the network. The network (Fig. 4.S1) was constructed from MCL distances (Table 4.S1) and suggests a conflicting signal in the *hmwA* sequences, a finding consistent with the statistical tests for recombination. The Neighbor-Net LSfit index was 0.9924; *i.e.*, the phenetic distances represented on the Neighbor-Net capture 99.24% of the estimated MCL genetic distances meaning that, despite the evidence for recombination, the dataset is still quite tree-like. Three distinct *hmwA* sequence clusters are evident in the network (Fig. 4.S1). Cluster 1 contains the *hmw2A* from each of the four strains for which sequence data are available for both *hmwA* loci. Estimation of a maximum likelihood consensus tree (based on 1000 bootstrap replicates) resulted in a phylogeny with the same overall topology; this is not surprising given the high Neighbor-Net LSfit index.

The 13 full-length *hmwA* sequence alignment was used specifically to test for evidence of positive selection operating within the first 441 amino acids of the HMW preprotein, which

encodes the *hmwA* signal sequence (11, 22). The GARD method detected five statistically significant ($P < 0.1$) recombination breakpoints within the signal sequence (Table 4.S2); the breakpoint information was taken into account when implementing the MEME method. There was no evidence of positive selection, using both the MEME method and omegaMap, operating within the *hmwA* signal sequence region. To the contrary, there was strong evidence of purifying selection ($\omega < 1$) operating within the signal sequence region (Fig. 4.S2). This was not unexpected given that the signal sequence is not exposed on the cell surface and, more importantly, may be subject to functional constraints since it must interact with both the Sec-dependent machinery, HMWC, and HMWB during HMW maturation and cell surface localization (22).

**HMW-family mature proteins.** The entire HMW adhesin mature protein (corresponding to amino acids 442 to 1536 of NTHi strain 12 HMW1A (11)) was represented in 15 *hmwA* sequences. *hmwA* nucleotide sequences were 60.1% A-T, ranged in length from 2976 to 3540 nts, and averaged 3275 nts in length. The overall mean genetic distance, assuming a gamma shape parameter = 0.5736, was 0.647 (S.E. = 8.668) base substitutions per site and pairwise genetic distances ranged from 0.001 to 0.985 base substitutions per site (Table 4.S3). The observed index of saturation, $I_{ss}$, was significantly smaller than the $I_{ss}$ critical value, $I_{ss.c}$, ($P < 0.001$) (57, 58) suggesting that the sequences have not experienced substitution saturation and thus are suitable for phylogenetic analyses. The PHI test for recombination found significant ($P < 0.01$) evidence of *hmwA* recombination and the GARD analysis identified 13 statistically significant ($P < 0.1$) recombination breakpoints (Table 4.S2). The MEME method was conducted on non-recombinant regions of the sequences defined by the GARD analysis.

The Neighbor-Net topology of the 15 HMW mature protein sequences (Fig. 4.2) was the same as that for the 13 full-length *hmwA* sequences (Fig. 4.S1). Three *hmwA* clusters were recovered in the Neighbor-Net and the NTHi strain 15 *hmw1A* and *hmw2A* sequences were placed in Clusters 2 and 3, respectively (Fig. 4.2). Consistent with tests for recombination, there was evidence of conflicting phylogenetic signal in the network. The Neighbor-Net LSfit index was 0.996 and the topology of the Neighbor-Net (Fig. 4.2) was identical to the bootstrap consensus tree (Fig. 4.S3) estimated using the maximum likelihood method and GTR+G+I (G = 1.3398, I = 30.50%) model of sequence evolution.

The MEME method, taking into account recombination, identified 55 codons as being subjected to episodic diversifying selection ($P < 0.05$) (Table 4.2, Fig. 4.S2). Using the glycosylated amino acids of NTHi strain 12 HMW1A as a reference (23), MEME identified three glycosylated amino acids that were also subjected to positive selection (asparagines AA3, AA471, AA505). Sixteen of the PSSs were predicted to be in hydrophobic regions (KD > 0) and 28 were located in regions predicted to be hydrophilic (HW > 0); one PSS (AA681) was predicted to be both hydrophobic (KD = 0.256) and hydrophilic (HW = 0.022) (Table 4.2**)**.

omegaMap identified 51 of positively selected amino acids (Table 4.2, Fig. 4.S2), ten of which were identified by both MEME and omegaMap as being positively selected. Seven of the PSSs identified by omegaMap mapped to hydrophobic regions (KD > 0) and 30 sites mapped to regions predicted to be hydrophilic (HW > 0). In both analyses, codons predicted to be positively selected were localized to the HMW binding domain region. None of the codons encoding glycosylated amino acids, relative to NTHi strain 12 HMW1 (23), were predicted to be positively selected.

***hmwA* sequence clusters.** The *hmw1* and *hmw2* loci apparently arose as the result of a gene duplication event and thus the two *hmwA* sequences from a given strain are paralogs. This provided the rationale for dividing *hmwA* sequences into two mutually exclusive sequence clusters designated Cluster 1 and Cluster 2-3. Sequences were allocated into one of the two clusters such that each of the five strains for which both *hmwA* copies were available had one *hmwA* in each cluster (Fig. 4.2). Cluster 1 contained seven *hmwA* sequences and included the NTHi Strain 12 *hmw2A* and Cluster 2-3 contained eight sequences and included strain 12 *hmw1A.* The strain 12, 5 and 15 *hmwA* nucleotide sequences clustered in the same way as previously reported when HMW amino acid sequences were subjected to phylogenetic analyses (8). This suggests that the clusters formed by *hmwA* nucleotide sequences reflect the *in vitro* adherence characteristics of their respective HMW adhesin. This leads to the prediction that Cluster 1 and Cluster 2 sequences may confer HMW2-like and HMW1-like adherence properties, respectively.

Clusters 2-3 sequences were more genetically diverse than Cluster 1 sequences. Cluster 2-3 *hmwA* sequences ranged in length from 3087 – 3474 nts with an average length of 3290 nts. Cluster 2-3 *hmwA* pairwise MCL genetic distances (gamma shape parameter = 0.6711) ranged

from 0.001 to 0.843 base substitutions per site and the overall mean genetic distance was 0.566 (S.E. = 10.860) base substitutions per site (Table 4.S4a). Cluster 1 *hmwA* sequences ranged in length from 2976 – 3483 nts and averaged 3227 nts. Cluster 1 pairwise MCL genetic distances (gamma shape parameter = 0.4925), ranged from 0.263 to 0.693 base substitutions per site and the overall mean genetic distance was 0.546 (S.E. = 0.022) base substitutions per site (Table 4.S4b). The PHI test for recombination identified statistically significant ($P < 0.01$) evidence of *hmwA* recombination in both clusters. GARD analysis identified four and five statistically significant ($P < 0.1$) recombination breakpoints in the Cluster 2-3 and Cluster 1 sequences, respectively (Table 4.S2); MEME analyses were conducted on non-recombinant regions of each alignment.

To allow for the possibility of selective pressures acting differently within each sequence Cluster, each Cluster was analyzed separately using MEME and omegaMap. Among Cluster 1 sequences, 25 and 37 amino acids were identified as under positive selection using MEME and omegaMap, respectively (Table 4.3, Figs. 4.3b and 4.S4). Cluster 2-3 sequences encoded 23 and 20 positively selected sites as predicted by MEME and omegaMap, respectively (Table 4.3, Figs. 4.3a and 4.S4). In both sequence clusters, the majority of the positively selected amino acids were localized to the HMW binding region (Fig 4.S4). While the number of positively selected codons identified with each method were similar for both *hmwA* Clusters, the exact codons predicted to be under positive selection differed by the method used (Table 4.3, Fig. 4.S4). In general, MEME identified singleton amino acids under selection whereas omegaMap was more likely to identify multiple contiguous amino acids as positively selected.

**Predicted HMW adhesin three-dimensional model.** To gain insight into the relative localization of positively selected amino acids, three dimensional models were predicted based on NTHi strain 12 HMW1 and HMW2 mature protein sequences (the 441 amino acid signal sequence was removed). I-TASSER provides a confidence score, C-score, that ranges from [-5, 2] and can be used to assess the quality of a predicted model; models with higher confidence have a higher scores. The C-score for the best fit HMW1 model (Fig. 4.4) was equal to -1.35 and HMW2 model (Fig. 4.S5) was equal to -1.26. The proteins in the protein data base with the most similar structure to the predicted HMW1 was a heme binding protein of *E. coli* (39) and the protein most similar to HMW2 was the passenger domain of the *E. coli* autotransporter EspP

(28). HMW1 and HMW2 predicted models were structurally similar, each containing two globular domains in the amino-terminal region (Fig.4.4 and Fig. 4.S5). Interestingly, the HMW1 binding domain (amino acids 114 – 473 (11)) is restricted to the larger of the two globular regions in the predicted HMW1 model whereas the HMW2 binding domain (amino acids 112 – 475 (11)) encompasses both globular regions of the HMW2 model; this could, however, simply reflect inaccuracies in the model reconstruction. The HMW binding domain is predicted to sit on top of a stalk region composed of parallel β-strands that extend to the HMW carboxy-terminus. The amino acids predicted to be subject to episodic positive selection within Cluster 2-3 and Cluster 1, as determined by the MEME method (Table 4.3), were mapped onto the strain 12 HMW1 (Fig. 4.4) and HMW2 (Fig. 4.S5) models, respectively. The majority of positively selected amino acids identified by MEME in were located in regions predicted to form either loops or turns (Table 4.S5) and to be hydrophilic (Hopp-Woods score > 0) (Tables 4.S6).

The HMW adhesins are glycosylated (20, 21, 23). To explore the relationship between glycosylated and positively selected amino acids, the glycosylated amino acids within the NTHi strain 12 HMW1 (23) were highlighted in the HMW1 predicted model (Fig. 4.4). There are no empirical data related to HMW2 glycosylation, therefore glycosylated asparagines were predicted, and mapped, based upon the N*X*(S/T) consensus sequence (23) (Fig. 4.S5). The majority of glycosylated amino acids also mapped to the binding domain, but, in contrast to positively selected amino acids, glycosylated amino acids were distributed across both of the globular regions of the HMW predicted structure that encompasses the core binding domain. With a single exception in HMW2, none of the glycosylated amino acids were predicted to be under episodic positive selection.

## Discussion

In this study, we tested for evidence of positive selection acting on the NTHi high molecular weight (HMW) adhesins using a collection of publicly available *hmwA* nucleotide sequences. Both methods we employed, MEME and omegaMap, estimate the ratio of the rates of non-synonymous (*dN*) to synonymous (*dS*) substitutions, commonly defined as *omega* ($\omega = dN/dS$), where values of $\omega > 1$ are indicative of positive selection. First, we analyzed the HMW signal sequence region (the first 441 amino acids) of 12 *hmwA* sequences and found no evidence of positive selection. Next, we analyzed 15 sequences of the HMW-family that encompassed the entire HMW mature protein, excluding the signal sequence. To allow for differences in selection pressures among paralogs, we separated the HMW-family sequences into two sequence clusters, Cluster 1 and Cluster 2-3, and analyzed each cluster separately. We found evidence of positive selection acting within the HMW-family of mature proteins as well as within both of the phylogenetically defined *hmwA* sequence clusters, and in all analyses the majority of positively selected amino acids were localized to the HMW adhesin binding domain.

Evolutionary analyses of the 13 full-length *hmwA* sequences and two additional HMW mature protein sequences highlight *hmwA* sequence diversity, which involves both insertions and deletions as well as marked nucleotide diversity. NTHi are naturally transformable and known to undergo homologous recombination, this is evident at the whole genome level (9, 10, 34) and has been suggested to occur specifically within the *hmwA* genes (8, 19). Consistent with these findings, we detected evidence of recombination within the HMW-family as a whole and, also, within Cluster 1 and Cluster 2-3 sequences.

Recombination confounds phylogenetic reconstructions because no single tree can represent the evolutionary history of a recombinant gene. We therefore employed the Neighbor-Net method (7), which constructs networks that reflect inconsistencies in the phylogenetic signal, to visualize *hmwA* relationships. The network topology for the *hmwA* mature binding region (Fig. 4.2) revealed three *hmwA* sequence clusters. A previous phylogenetic analysis that focused on relationships among a subset of the HMW mature protein sequences included in the current study revealed two HMW clusters that reflected *in vitro* binding characteristics, one composed of HMW1-like proteins and the second of HMW2-like proteins (8). In the current study, these

sequences are members of HMW Clusters 1 and 2 (Fig 4.2). In this study, we identified a third cluster of phylogenetically distinct *hmwA* sequences, a finding consistent with a previous study of HMW core binding domains by Giufre *et al.*, (19) that suggested not all *hmwA* core binding sequences fell into one of two distinct clusters.

Interestingly, among the five strains with *hmwA* sequence data from both loci, each strain contains a *hmwA* from Cluster 1 while the second *hmwA* belongs to either Cluster 2 or 3 (Fig. 4.2). This result suggests that following the *hmwA* duplication event, the Cluster 1 *hmwA* may have retained its original binding affinity whereas the Cluster 2-3 *hmwA* evolved novel binding characteristics. *In vitro* studies have demonstrated that strain 12 HMW1, encoded by *hmw1A* of Cluster 2, mediates strong adherence, via interaction with α 2,3-sialic acid, to Chang, HaCaT, NCI-H292, and HEp-2 cell lines, whereas, HMW2, encoded by *hmw2A*, mediates adherence primarily to HaCaT and NCI-H292 cell lines via unidentified moieties (8, 11, 42, 46). The *in-vitro* binding characteristics have been determined for NTHi strains 12, 5 and 15 (8, 46). For these three strains, HMW1-like and HMW2-like amino acid sequences cluster with respect their *in vitro* adherence profiles and independently of their chromosomal location; *i.e.*, in both strain 5 and strain 15 the *hmwA* locus adjacent to HI1679 (*hmw1A* based on the NTHi strain 12 nomenclature) encodes a HMW that exhibits HMW2-like binding characteristics (8). The limited binding capacity of HMW2-like proteins coupled with the observation that all previously characterized HMW2-like proteins form a single cluster, suggests that HMW2-like binding (*i.e.*, sequences from Cluster 1) may be essential for NTHi whereas members of the HMW1-like group have evolved to take on expanded binding abilities. Thus, positive selection may be acting to shape HMW adhesin adherence capacity, allowing NTHi to expand its niche space within the respiratory tract. There are, however, two important limitations to this interpretation that must be kept in mind. First, the strains in this study are all clinical isolates (Table 4.1) and therefore represent only a subset of NTHi genetic diversity, much of which may reside in the population of strains colonizing healthy individuals. The second caveat is that the binding characteristics were all determined *in vitro* using a relatively limited number of cell types. Functional analyses are necessary to determine if the observed phylogenetic clustering reflects expanded adherence properties. More specifically, it will be interesting to determine if the Cluster 3 HMW adhesins display unique adherence properties.

There are number of possible explanations for the observed differences between the results of the omegaMap method and the MEME method (Table 4.2, 4.3, and Figs. 4.S2, 4.S4). First, the methods differ in their underlying models, as omegaMap takes a population genetics approach by employing an approximation to coalescent with recombination but does not explicitly model the genealogy (53). In contrast, the MEME method is a phylogenetic approach and is thus dependent on the estimated phylogenetic relationship among sequences (29). Secondly, the MEME method allows for variation from codon-to-codon as well as branch-to-branch and therefore is capable of detecting both pervasive and episodic diversifying selection (29). Analyses of empirical datasets demonstrate that episodic selection is common and that methods designed to detect only pervasive selection can underestimate the number of positively selected codons (37). While the absolute number of amino acids identified with each method was similar, omegaMap was more likely to identify a series of adjacent codons as positively selected whereas MEME identified singleton codons distributed throughout the binding region (Table 4.3, Figs. 4.S4). This may be a attributable to the block method employed by omegaMap in which a user-specified number of adjacent codons (defined as the block size) share a common parameter estimate, essentially parameters are averaged across the number of sites defined by the block size. The block size affects the amount of time required to complete an analysis, that is, analysis time increases as the block size is reduced, and, given the computational intensity of the model, computation time influences the choice of block size. In contrast, the MEME method considers each codon site individually; this is the equivalent to an omegaMap block size equal to one. Because the block method averages across blocks, a single site under strong selective pressure (large dN/dS) could influence the block value, leading multiple codons within that block showing evidence of positive selection. The converse could occur as well, failure to detect a single positively selected site could occur because of purifying selection acting on codons within the same block, dampening the average signal. It is important to note, however, that both methods generated qualitatively similar results: positively selected sites were located in hydrophobic regions of the binding domain predicted to form either loops or turns.

The exact nature of the selective pressure acting on the HMW adhesins cannot be elucidated from *in silico* studies alone. The fact that HMW adhesins stimulate a strong antibody mediated immune response (1), however, suggests that immune mediated selective pressure likely plays a role. Interestingly, Barenkamp and St. Geme III mapped B-cell epitopes in HMW1

and HMW2 targeted by two monoclonal antibodies (MAb) AD6 and 10C5 (5). The AD6 MAb was predicted to recognize a peptide in the last 75 amino acids of HMW1 and HMW2. The MEME method identified a single amino acid in the last 75 amino acids as being under episodic positive selection in the HMW-family (T1085, relative to strain 12 HMW1) as well as in both the Cluster 2-3 and Cluster 1 sequences which correspond to T1085 and R1025, relative to strain 12 HMW1 and HMW2, respectively. MAb 10C5 is specific to HMW1 and binds to an epitope located within the carboxy-terminal 155 amino acids. The MEME method predicted that amino acid L994 was under positive selection in the HMW-family and Cluster 2-3 sequence, but this site was not predicted to be under positive selection in Cluster 1 sequences. These findings are consistent with a recent study in *Streptococcus pneumoniae* that found regions containing antibody epitopes were more likely to contain positively selected codons than were non-epitope encoding regions (35). Taken together, these finding lend support to the potential role of immune mediated selection pressure in driving HMW diversity, but, the impact of these associations is difficult to assess since the MAbs were not produced nor tested against HMW adhesins in their native conformation. Further work is necessary to link the positively selected amino acids identified in this study with a functional role *in-vivo.*

The findings of this study contribute to our understanding of the mechanisms driving NTHi HMW evolution. The *hmwA* phylogenetic analysis reported here suggests that the majority of *hmwA* sequences fall into one of three distinct sequence clusters and that each strain encodes a Cluster 1-*hmwA* sequence. The molecular evolution of *hmwA* is driven by positive selection acting, primarily, on the region in and around the core-binding domain and may be a response to selective pressures applied by the host immune system. Gene prevalence studies suggest that HMW proteins are more prevalent in NTHi strains associated with AOM than in colonizing strains thus suggesting that HMW adhesins contribute to NTHi virulence (13, 48). Furthermore, HMWs stimulate a strong and lasting humoral response in colonized individuals (1, 46, 56). Given their potential role in virulence, coupled with the ability to stimulate an immune response, the HMW proteins have been proposed as candidates for inclusion in a multivalent NTHi vaccine (54-56). Therefore, understanding the genetic diversity of *hmwA* and the mechanisms driving that diversity not only advance our understanding of host-pathogen interactions but can also to help inform and guide NTHi vaccine development.

Table 4.1. Publicly available strains used in this study.

| Locus | Strain[a,b] | Gene | *hmwA* (# nts) | Source[c] | Reference (Comments): |
|---|---|---|---|---|---|
| HIU08876 | strain 12[a] | *hmw1A* | 4608 | AOM | (2) |
| HIU08875 | strain 12[a] | *hmw2A* | 4431 | AOM | (2) |
| CP000671 | PittEE[a] | *hmw1A* | 4743 | OME | (25)(*hmw1A* region: 745702-750447) |
| CP000671 | PittEE[a] | *hmw2A* | 4764 | OME | (25)(*hmw2A* region: 1118977-1123743) |
| AY497551 | strain 5[a] | *hmw1A* | 4794 | AOM | (8) |
| AY497552 | strain 5[a] | *hmw2A* | 4803 | AOM | (8) |
| AY497553 | strain 15[b] | *hmw2A* | 2997 | AOM | (8) |
| AY497554 | strain 15[b] | *hmw1A* | 3567 | AOM | (8) |
| YP_249393.1 | 86-028NP[a] | *hmw1A* | 4476 | NP of COM Patient | (24) |
| YP_248929.1 | 86-028NP[a] | *hmw2A* | 4626 | NP of COM Patient | (24) |
| AY601282 | G822[a] | *hmw1A* | 4446 | AOM | (12)(downstream of *HI1679*) |
| AY601283 | AAr105[a] | *hmw1A* | 4470 | AOM | (12)(downstream of *HI1679*) |
| AY601284 | AAr96[a] | *hmw2A* | 4542 | AOM | (12)(downstream of *HI1598*) |
| AF180944 | A950006[a] | *hmwA* | 4671 | COPD | (8, 52) |
| AJ937359 | strain 72[a] | *hmw1A* | 4407 | invasive | (19) |

[a] full-length *hmwA* sequences (signal sequence plus mature binding protein)

[b] lacks signal sequence (mature binding protein only)

[c] AOM = acute otitis media; OME = otitis media with effusion; NP = nasopharynx; COM = chronic otitis media; COPD = chronic obstructive pulmonary disease

Table 4.2. HMW-family positively selected amino acids.

| 15 *hmwA*[a] | PSS[b] | GLY[c] (site #) | KD[d] > 0 | HW[e] > 0 | PSS amino acid relative to HMW1 mature peptide[f] |
|---|---|---|---|---|---|
| MEME[g] | 55 | 3 (3, 471, 505) | 16 | 28 | 3, 29, 39, 44, 47, 48, 54, 65, 75, 112, 124, 153, 167, 178, 214, 225, 226, 243, 250, 252, 260, 271, 276, 297, 328, 372, 403, 421, 422, 431, 450, 461, 471, 485, 489, 496, 505, 521, 528, 537, 572, 580, 601, 613, 618, 622, 640, 649, 662, 668, 681, 686, 703, 994, 1085 |
| omegaMap | 51 | 0 | 7 | 30 | 16, 17, 19, 20, 21, 29, 48, 54, 55, 73, 124, 172, 182, 211, 212, 213, 214, 226, 230, 231, 241, 242, 243, 246, 250, 251, 252, 253, 254, 255, 261, 262, 263, 265, 266, 278, 280, 284, 342, 343, 344, 345, 346, 347, 348, 414, 455, 461, 462, 513, 514 |

[a] *hmwA* region encoding the HMW mature binding proteins, relative to NTHi strain 12 HMW1 amino acids 442 to 1536 (11)

[b] positively Selected Site (PSS), sites were considered PSS by MEME if $P < 0.05$ and PSS by omegaMap if the posterior probability > 0.95

[c] glycosylated amino acids (23) relative to NTHi strain 12 *hmw1A* mature peptide[e]

[d] Kyte-Doolittle hydrophilicity score (window size = 9) (33); predicted hydrophobic amino acid

[e] Hopp-Woods hydrophobicity score (window size = 9) (26); predicted hydrophilic amino acid

[f] amino acid numbering is relative to the mature binding protein(11), excluding the signal sequence; the first amino acid of the mature binding protein corresponds to amino acid number 442 (Proline) of the NTHi strain 12 preprotein

[g] GARD analysis identified 13 statistically significant ($P < 0.1$) recombination breakpoints

Table 4.3. Cluster 1 and Cluster 2-3 positively selected amino acids

| Cluster 2-3 sequences[a] | PSS[b] | GLY[c] (site#) | KD[d] > 0 | HW[e] > 0 | PSS amino acid relative to HMW1 mature peptide[f] |
|---|---|---|---|---|---|
| MEME[k] | 23 | 1 (444) | 8 | 11 | 3, 34, 39, 47, 75, 80, 91, 170, 185, 189, 216, 252, 271, 318, 369, 403, 421, 436, 572, 622, 650, 994, 1085 |
| omegaMap | 20 | 0 | 3 | 9 | 10, 29, 124, 209, 212, 213, 214, 215, 216, 226, 251, 252, 253, 254, 289, 290, 291, 319, 463, 521 |

.

| Cluster 1 sequences[g] | PSS[b] | GLY[h] (site#) | KD[d] > 0 | HW[e] > 0 | PSS amino acid relative to HMW2 mature peptide[i,j] |
|---|---|---|---|---|---|
| MEME[l] | 25 | 1 (363) | 5 | 15 | 24, 26, 30, 110, 140, 156, 243, 262, 298, 308, 323, 345, 363, 370, 378, 435[j], 491, 512, 574[j], 582, 620, 652, 671, 697, 1025 |
| omegaMap | 37 | 0 | 4 | 24 | 16, 17, 18, 19, 24, 25, 26, 31, 36, 38, 55, 56, 62, 63, 75, 137, 240, 241, 242, 243, 245, 246, 247, 248, 249, 250, 251, 252, 289, 290, 292, 372, 378, 379, 380, 413, 414 |

[a] *hmwA* region encoding the HMW mature binding proteins, NTHi strain 12 HMW1 amino acids 442 to 1536 (11)

[b] positively Selected Site (PSS), sites were considered PSS by MEME if $P < 0.05$ and PSS by omegaMap if the posterior probability > 0.95

[c] glycosylated amino acids (23) relative to NTHi strain 12 *hmw1A* mature peptide[e]

[d] Kyte-Doolittle hydrophobicity score (window size = 9) (33); predicted hydrophobic amino acids

[e] Hopp-Woods hydrophilicity score (window size = 9) (26); predicted hydrophilic amino acids

[f,i] positively selected amino acids; amino acid numbering is relative to the mature binding protein(11), excluding the signal sequence; the first amino acid of the mature binding protein corresponds to amino acid number 442 (Proline) of the NTHi strain 12 HMW1 or HMW2 preprotein

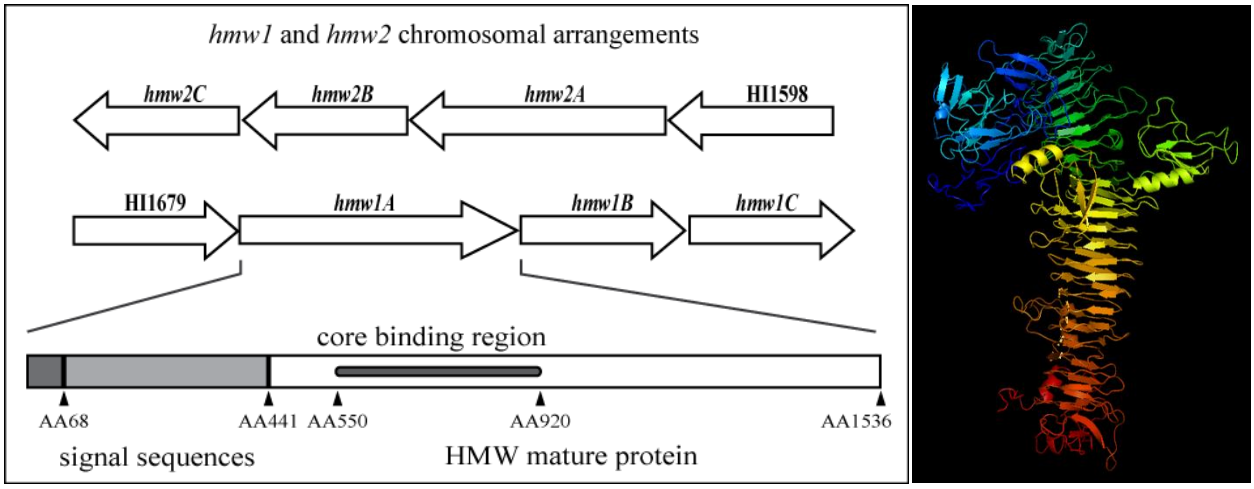
[g] *hmwA* region encoding the HMW mature binding proteins, NTHi strain 12 and HMW2 amino acids 442 to 1477 (11)

[h] amino acids predicted to be glycosylated based on amino acid sequence NX(S/T) (23)

[j] amino acids with KD > 0 and HW > 0, site HMW2 AA435 KD = 0.022, HW =0.567; AA574 KD = 0.078, HW = 0.356

[k] GARD analysis identified four statistically significant ($P < 0.1$) recombination breakpoints

[l] GARD analysis identified five statistically significant ($P < 0.1$) recombination breakpoints




(a) (b)

Figure 4.1. Chromosomal arrangements of the *hmw1* loci of NTHi strain 12. The *hmw* loci are located in conserved yet unlinked regions of the NTHi chromosome, in strain 12, the *hmw1* and *hmw2* loci are downstream of HI1679 and HI1598, respectively; chromosomal locations are named with respect to *H. influenzae* Rd. Each *hmw* locus consists of three genes, *hmwA, hmwB,* and *hmwC.* The functional HMW adhesin is encoded by *hmwA.* The strain 12 HMW1 adhesin preprotein contains a 441 amino acid signal sequence (grey boxes) that is cleaved during maturation. The core binding domain is localized to the amino terminus of the mature protein and is the region of the adhesin that interacts with the host epithelium. (b) Predicted three dimensional strain 12 HMW1 mature adhesin. The structure was predicted using the I-TASSER server (40, 60) and visualized with PyMOL (41). The color coding transitions from light blue at the amino terminus to red at the carboxy terminus.

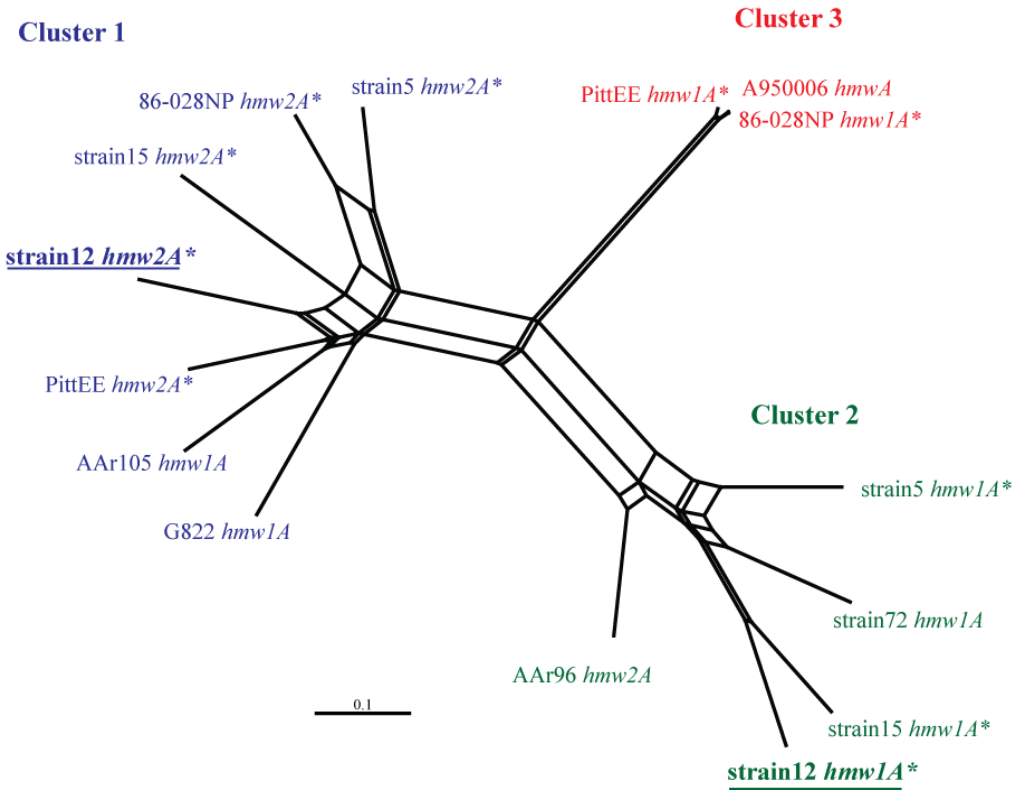
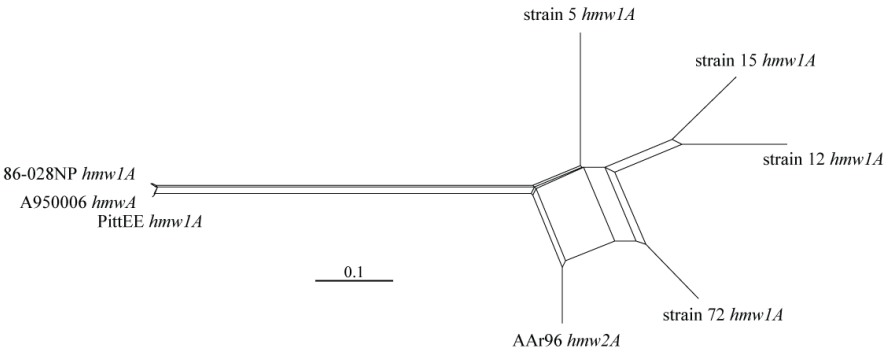
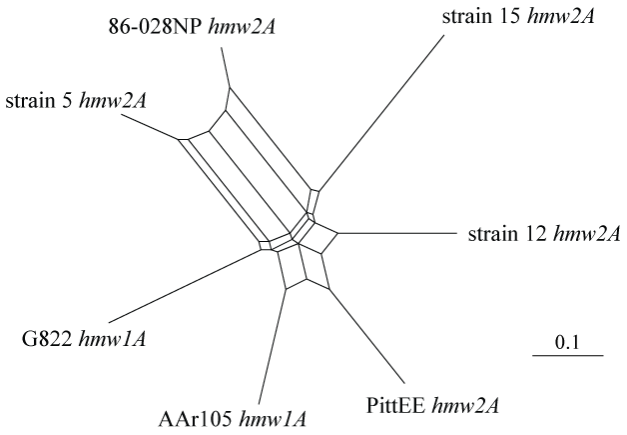



Figure 4.2. Neighbor-Net for the 15 *hmwA* mature binding protein regions. The *hmwA* loci arose from a duplication event; therefore, both *hmwA* genes from a specific strain represent paralogous genes. The 15 full length sequences were partitioned into one of three clusters, based on their placement in the Neighbor-Net. The network was estimated from maximum composite likelihood genetic distances (TableS3). Boxes in the Neighbor-Net reflect inconsistencies in the phylogenetic signal due to recombination. Sequence data were available for both copies of *hmw*, for four strains (indicated by an "*"). Each of the four strains encoded an *hmwA* from Cluster 1 (based on the sequences) but, the second *hmwA* from the four strains belonged to either Cluster 2 or Cluster 3 (based on the sequences). Branch lengths are drawn to scale and represent the number of pairwise base substitutions per site; LSfit index = 0.996.




(a) Cluster 2-3 sequences




(b) Cluster 1 sequences

Figure 4.3. Neighbor-Net for *hmwA* adhesin mature protein sequence Clusters. Boxes in the Neighbor-Net network reflect inconsistencies in the phylogenetic signal among *hmwA*. Branch lengths are drawn to scale and represent the number of pairwise base substitutions per site. (a) HMW1-group adhesins, LSfit index = 0.996. (b) HMW2-group adhesins, LSfit index = 0.998.

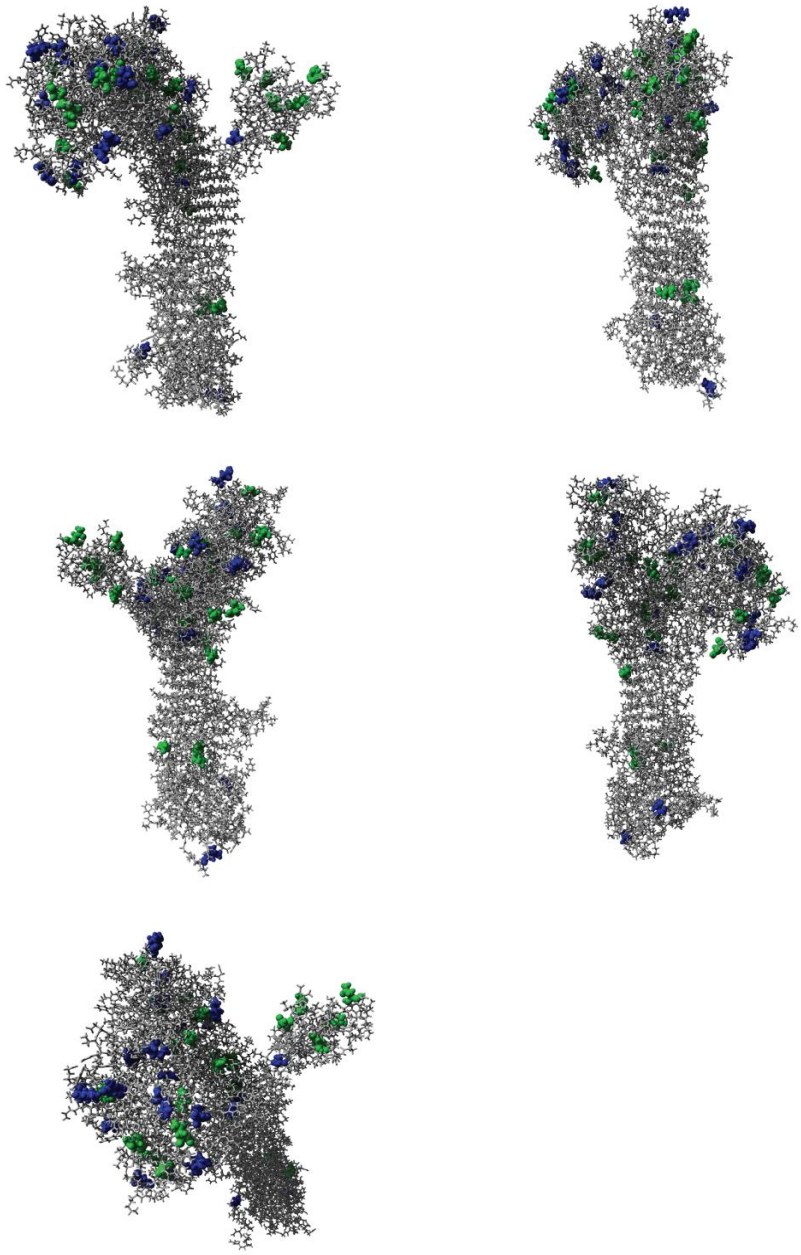


Figure 4.4. Theoretical structure for NTHi strain 12 HMW1 predicted by I-TASSER. The 23 amino acids predicted to be subjected to episodic positive selection, highlighted in blue (Table 4.3), are localized to the distal globular region of HMW1. Glycosylated amino acids are highlighted in green. The three structural model of NTHi strain 12 HMW1 was predicted using I-TASSER and visualized with YASARA. The model is rotated clockwise approximately 90-degrees between each panel and the last panel is top-down shot.

**Supplemental Tables and Figures.**

Table 4.S1. Full length *hmwA* maximum composite likelihood pairwise genetic distances (nucleotide substitutions per site).

| | **Strain *locus*** | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | strain AAr96 *hmw2A* | 0.000 | | | | | | | | | | | |
| 2 | strain 72 *hmw1A* | 0.267 | 0.000 | | | | | | | | | | |
| 3 | strain 12 *hmw1A* | 0.393 | 0.327 | 0.000 | | | | | | | | | |
| 4 | strain 5 *hmw1A* | 0.400 | 0.285 | 0.396 | 0.000 | | | | | | | | |
| 5 | PittEE *hmw1A* | 0.568 | 0.652 | 0.652 | 0.612 | 0.000 | | | | | | | |
| 6 | 86-028NP *hmw1A* | 0.620 | 0.620 | 0.684 | 0.553 | 0.045 | 0.000 | | | | | | |
| 7 | A950006 *hmwA* | 0.624 | 0.628 | 0.688 | 0.547 | 0.049 | 0.004 | 0.000 | | | | | |
| 8 | AAr105 *hmw1A* | 0.480 | 0.573 | 0.569 | 0.631 | 0.562 | 0.596 | 0.605 | 0.000 | | | | |
| 9 | PittEE *hmw2A* | 0.452 | 0.528 | 0.551 | 0.611 | 0.385 | 0.522 | 0.536 | 0.269 | 0.000 | | | |
| 10 | strain 12 *hmw2A* | 0.493 | 0.636 | 0.582 | 0.565 | 0.562 | 0.620 | 0.624 | 0.308 | 0.273 | 0.000 | | |
| 11 | 86-028NP *hmw2A* | 0.600 | 0.602 | 0.627 | 0.456 | 0.524 | 0.470 | 0.472 | 0.427 | 0.365 | 0.323 | 0.000 | |
| 12 | strain 5 *hmw2A* | 0.565 | 0.556 | 0.659 | 0.339 | 0.502 | 0.463 | 0.457 | 0.333 | 0.365 | 0.411 | 0.210 | 0.000 |
| 13 | G822 *hmw1A* | 0.589 | 0.636 | 0.710 | 0.594 | 0.475 | 0.431 | 0.433 | 0.342 | 0.397 | 0.363 | 0.388 | 0.319 | 0.000 |
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |

Table 4.S2. Statistically significant (P < 0.1) recombination break points identified by GARD analysis (30, 31).

| 13FL *hmwA*[a] | 15 *hmwA*[a] | | Cluster 1[a] | | Cluster 2-3[a] | |
|---|---|---|---|---|---|---|
| amino acid relative strain 12 full length HMW1[b,c] | amino acid relative to Strain 12 HMW1A MBP[d] | amino acid relative strain 12 full-length HMW1A[c] | amino acid relative to Strain 12 HMW1A MBP[d] | amino acid relative strain 12 full-length HMW1[c] | amino acid relative to Strain 12 HMW2 MBP[d] | amino acid relative to strain 12 full-length HMW2[c] |
| 148 | 74 | 515 | 547 | 988 | 91 | 532 |
| 180 | 130 | 571 | 793 | 1234 | 128 | 569 |
| 234 | 232 | 673 | 869 | 1310 | 360 | 801 |
| 319 | 362 | 803 | 934 | 1375 | 498 | 939 |
| 417 | 470 | 911 | | | 615 | 1056 |
| 561 | 522 | 963 | | | | |
| 637 | 547 | 988 | | | | |
| 686 | 617 | 1058 | | | | |
| 795 | 647 | 1088 | | | | |
| 911 | 722 | 1163 | | | | |
| 979 | 934 | 1375 | | | | |
| 1058 | 963 | 1404 | | | | |
| 1163 | 1071 | 1512 | | | | |
| 1406 | | | | | | |
| 1456 | | | | | | |

[a] *hmA* sequence alignments; 13 full-length sequences (13FL), 15 mature binding protein regions (15), eight sequences defined as Cluster 2 (**Figure_15hmwA neighbor net**), and seven sequences defined as Cluster 2-3 (**Figure_15hmwA neighbor net**)

[b] includes the 441amino acid signal sequence region

[c] numbering relative to strain 12 full-length HMW1A (including signal sequence)

[d] numbering relative to the strain 12 HMW1A or HMW2A mature binding protein (MBP) (excluding signal sequence)

Table 4.S3. *hmwA* mature protein coding region maximum composite likelihood pairwise genetic distances (nucleotide substitutions per site).

| | **Strain *locus*** | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Aar 96 *hmw2A* | 0.000 | | | | | | | | | | | | | |
| 2 | strain 72 *hmw1A* | 0.291 | 0.000 | | | | | | | | | | | | |
| 3 | strain 5 *hmw1A* | 0.409 | 0.338 | 0.000 | | | | | | | | | | | |
| 4 | strain 15 *hmw1A* | 0.504 | 0.403 | 0.455 | 0.000 | | | | | | | | | | |
| 5 | strain 12 *hmw1A* | 0.517 | 0.394 | 0.466 | 0.280 | 0.000 | | | | | | | | | |
| 6 | A950006 *hmwA* | 0.678 | 0.782 | 0.710 | 0.812 | 0.808 | 0.000 | | | | | | | | |
| 7 | 86-028NP *hmw1A* | 0.676 | 0.780 | 0.708 | 0.811 | 0.807 | 0.001 | 0.000 | | | | | | | |
| 8 | PittEE *hmw1A* | 0.658 | 0.768 | 0.717 | 0.825 | 0.824 | 0.024 | 0.024 | 0.000 | | | | | | |
| 9 | G822 *hmw1A* | 0.741 | 0.875 | 0.799 | 0.985 | 0.950 | 0.628 | 0.626 | 0.639 | 0.000 | | | | | |
| 10 | AAr105 *hmw1A* | 0.705 | 0.766 | 0.838 | 0.966 | 0.864 | 0.749 | 0.747 | 0.758 | 0.414 | 0.000 | | | | |
| 11 | PittEE *hmw2A* | 0.629 | 0.734 | 0.833 | 0.812 | 0.831 | 0.700 | 0.698 | 0.618 | 0.454 | 0.337 | 0.000 | | | |
| 12 | strain 12 *hmw2A* | 0.659 | 0.830 | 0.761 | 0.934 | 0.963 | 0.789 | 0.787 | 0.763 | 0.415 | 0.412 | 0.357 | 0.000 | | |
| 13 | strain 15 *hmw2A* | 0.760 | 0.849 | 0.813 | 0.892 | 0.787 | 0.720 | 0.719 | 0.721 | 0.518 | 0.441 | 0.512 | 0.420 | 0.000 | |
| 14 | 86-028NP *hmw2A* | 0.759 | 0.915 | 0.750 | 0.899 | 0.847 | 0.637 | 0.636 | 0.664 | 0.483 | 0.498 | 0.430 | 0.413 | 0.425 | 0.000 |
| 15 | strain 5 *hmw2A* | 0.726 | 0.826 | 0.599 | 0.962 | 0.893 | 0.632 | 0.631 | 0.639 | 0.421 | 0.409 | 0.442 | 0.509 | 0.537 | 0.238 |
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |

Table 4.S4. *hmwA* mature protein coding region maximum composite likelihood pairwise genetic distances (nucleotide substitutions per site) by group.

(a) Cluster 2-3 sequences

| | **Strain *locus*** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 1 | AAr96 *hmw2A* | 0.000 | | | | | | | |
| 2 | strain72 *hmw1A* | 0.290 | 0.000 | | | | | | |
| 3 | strain 5 *hmw1A* | 0.430 | 0.374 | 0.000 | | | | | |
| 4 | strain 15 *hmw1A* | 0.489 | 0.426 | 0.437 | 0.000 | | | | |
| 5 | strain 12 *hmw1A* | 0.529 | 0.433 | 0.477 | 0.268 | 0.000 | | | |
| 6 | A950006 *hmwA* | 0.686 | 0.835 | 0.739 | 0.800 | 0.829 | 0.000 | | |
| 7 | 86-028NP *hmw1A* | 0.685 | 0.831 | 0.736 | 0.800 | 0.829 | 0.001 | 0.000 | |
| 8 | PittEE *hmw1A* | 0.670 | 0.820 | 0.747 | 0.811 | 0.843 | 0.021 | 0.021 | 0.000 |
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |

(b) Cluster 1 sequences

| | **Strain *locus*** | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 1 | G822 *hmw1A* | 0.000 | | | | | | |
| 2 | AAr105 *hmw1A* | 0.505 | 0.000 | | | | | |
| 3 | PittEE *hmw2A* | 0.587 | 0.413 | 0.000 | | | | |
| 4 | strain 12 *hmw2A* | 0.504 | 0.500 | 0.432 | 0.000 | | | |
| 5 | strain 15_hmw2A | 0.669 | 0.578 | 0.677 | 0.540 | 0.000 | | |
| 6 | 86-028NP *hmw2A* | 0.614 | 0.637 | 0.554 | 0.505 | 0.541 | 0.000 | |
| 7 | strain 5 *hmw2A* | 0.527 | 0.518 | 0.581 | 0.629 | 0.693 | 0.263 | 0.000 |
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 |

Table 4.S5. Location of PSS in Cluster 2-3 and Cluster 1 sequences mapped to the predicted secondary structure of NTHi strain 12 HMW1 or HMW2, respectively.

(a) Cluster 2-3

| Cluster 2-3 | coil | turn | helix | sheet | total |
|---|---|---|---|---|---|
| MEME | 15 | 2 | 1 | 5 | 23 |
| oM | 15 | 3 | 2 | 0 | 20 |

(b) Cluster 1

| Cluster 1 | coil | turn | helix | sheet | total |
|---|---|---|---|---|---|
| MEME | 15 | 6 | 1 | 3 | 25 |
| oM | 31 | 2 | 3 | 1 | 37 |

Table 4.S6. HMW1- and HMW2- group predicted Kyte-Doolittle (KD) values and Hopp-Woods (HW) for NTHi amino acids with KD ≠ 0.

(a) HMW1 Kyte-Doolittle hydrophobicity

| | | PSSs (P < 0.05) | | |
|---|---|---|---|---|
| | | YES | NO | Total |
| KD | >0 | 8 | 311 | 319 |
| | <0 | 15 | 586 | 601 |
| | Total | 23 | 897 | 920 |

(b) HMW1 Hopp-Woods hydrophilicity

| | | PSSs (P < 0.05) | | |
|---|---|---|---|---|
| | | YES | NO | Total |
| HW | > 0 | 11 | 435 | 446 |
| | < 0 | 11 | 459 | 470 |
| | Total | 22 | 894 | 916 |

(c) HMW2 Kyte-Doolittle hydrophobicity

| | | PSSs (P < 0.05) | | |
|---|---|---|---|---|
| | | YES | NO | Total |
| KD | >0 | 5 | 255 | 260 |
| | <0 | 20 | 579 | 599 |
| | Total | 25 | 834 | 859 |

(d) HMW2 Hopp-Woods hydrophilicity

| | | PSSs (P < 0.05) | | |
|---|---|---|---|---|
| | | YES | NO | Total |
| HW | >0 | 15 | 443 | 458 |
| | <0 | 9 | 393 | 402 |
| | Total | 24 | 863 | 860 |

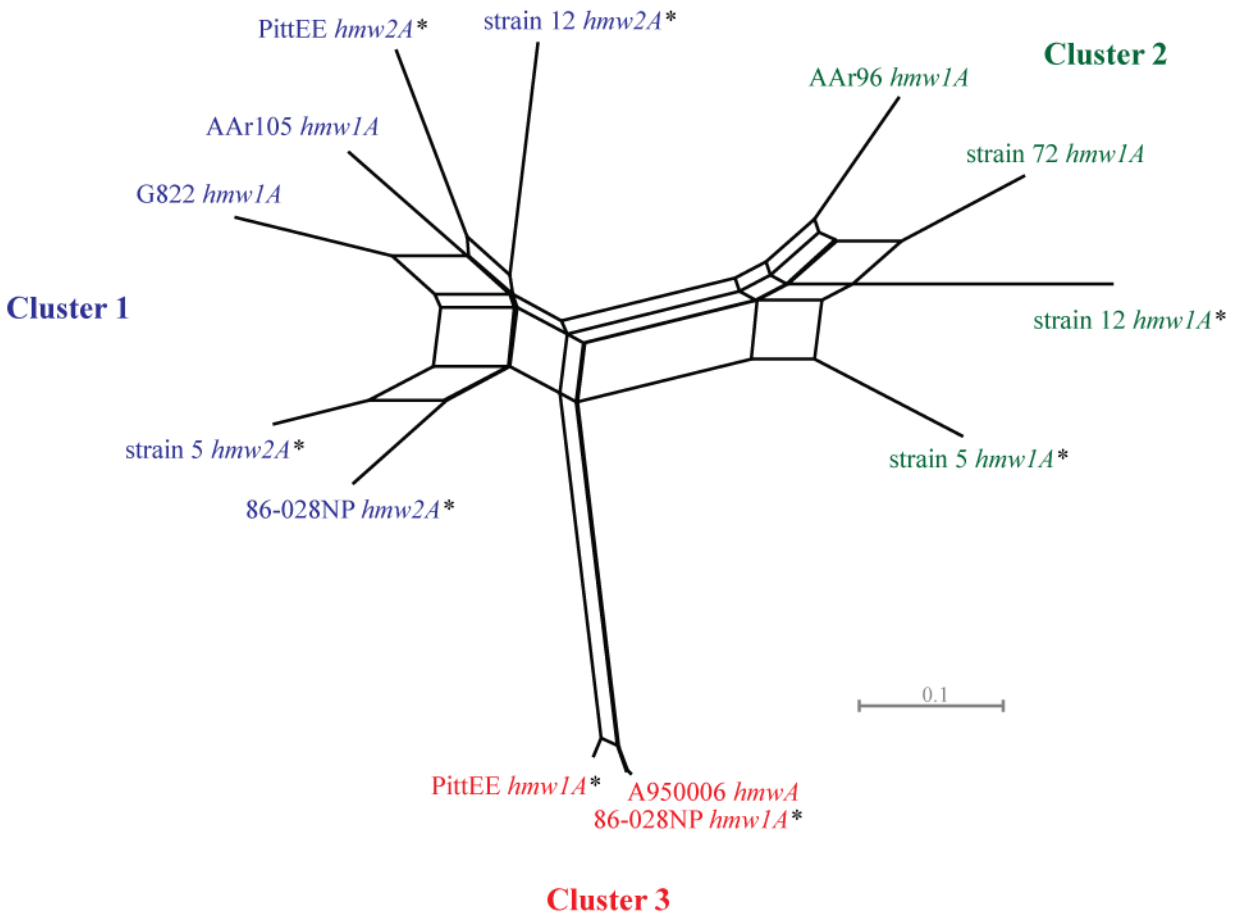



Figure 4.S1. Neighbor-Net for 13 full-length *hmwA* sequences estimated using maximum composite likelihood distances (Table S1). There is statistically significant evidence of recombination within the full-length *hmwA*, reflected by boxes in the network. *hmwA* sequences formed three distinct clusters. Sequence data was available for both *hmw1A* and *hmw2A* from four strains (indicated by an "*") and each of the four strains encodes an *hmwA* that belonged to Cluster 1, however, the second *hmwA* from each strain belonged to either Cluster 2 or Cluster 3. *hmw1A* and *hmw2A* from a strain represent paralogous genes, therefore, sequences were partitioned into one of two mutually exclusive groups, Cluster 1 or Cluster 2-3. Branch lengths are drawn to scale and represent the number of pairwise base substitutions per site; LSfit index = 0.9924.

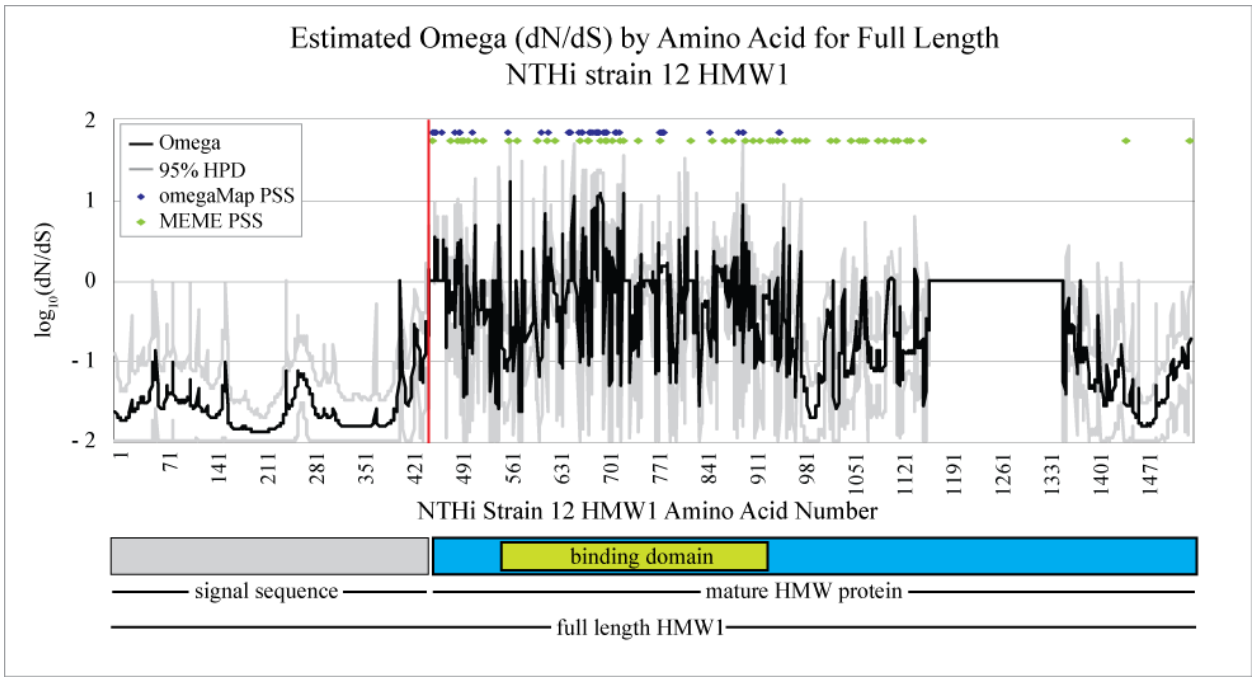



Figure 4.S2. Estimated omega (ω = dN/dS) by codon for the full-length HMW-family of adhesins. The mean (black line) and higher and lower 95% credible intervals (grey lines) for ω were estimated by omegaMap; all values have been $\log_{10}$ transformed. Log(ω) values < 0 indicate regions under purifying (negative) selection, = 0 neutral evolution, and > 0 positive selection. Amino acid positions are plotted with reference to NTHi strain 12 HMW1 full length preprotein, the signal sequence and binding domains regions (11) are indicated beneath the plot. The vertical red line marks the separation between the signal sequence and the mature HMW adhesin. Thirteen sequences were included in the analysis of the signal sequence and fifteen sequences were included in the mature HMW-adhesin analysis (Table 4.1), the output files of the two separate analyses were merged to generate the plot. The signal sequence, amino acids 1 – 441, is predicted to be under strong purifying selection (log(ω) < 0). The majority of sites predicted to be under positive selection (Table 4.2) were localized to the binding domain region of the mature HMW protein and are identified above the plot with either blue (omegaMap; n = 51) or green (MEME; n = 55) diamonds. All regions containing gaps in the multiple sequence alignment, *e.g.*, between amino acids 1150 – 1340, were coded as ω = 1.

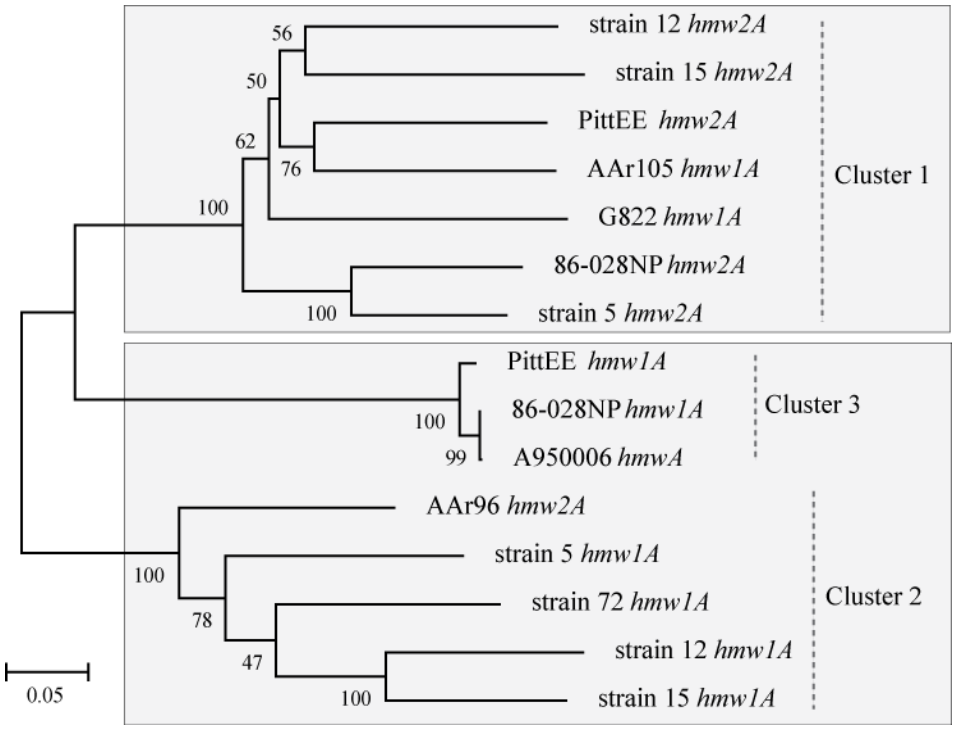
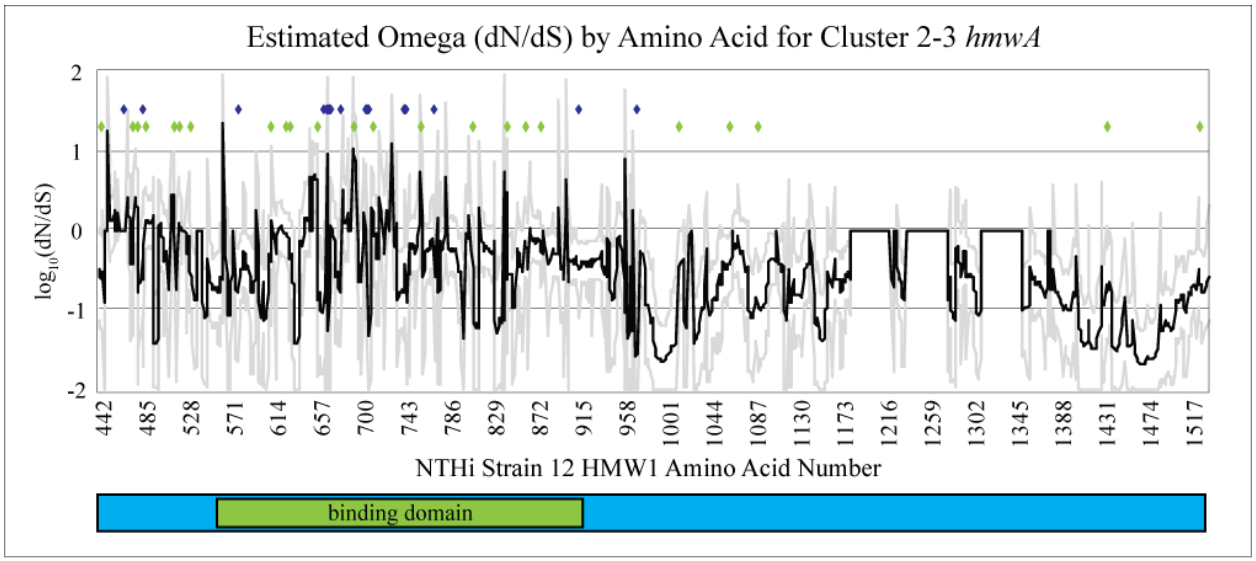
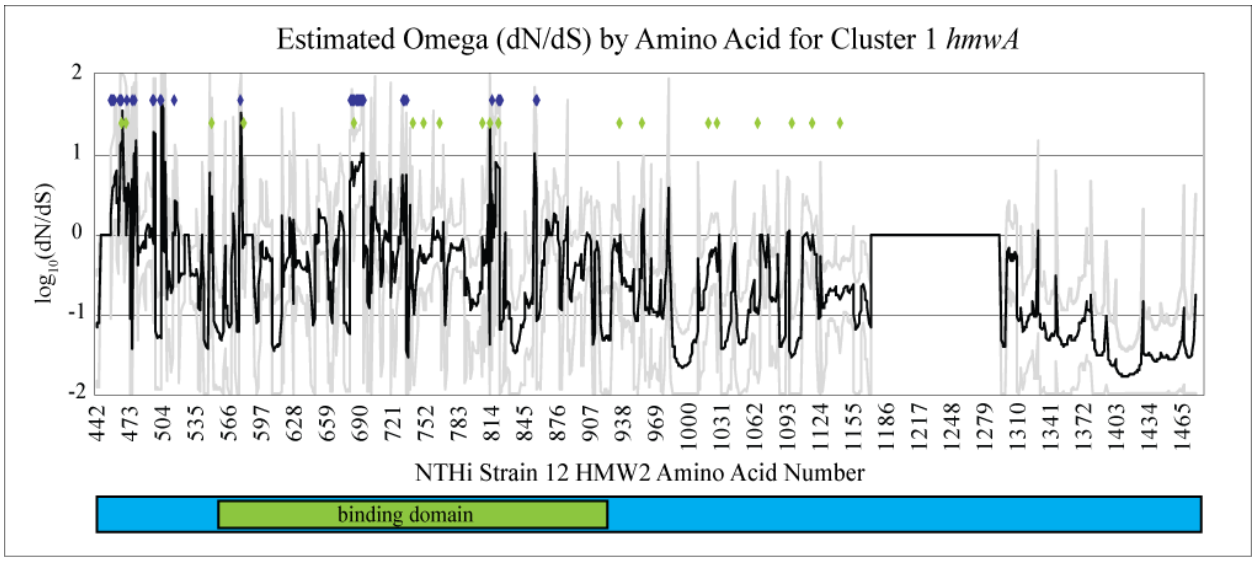



Figure 4.S3. Maximum likelihood phylogenetic analysis of the 15 *hmwA* mature binding region sequences. The bootstrap consensus tree was inferred from 1000 replicates and branches corresponding to partitions reproduced in less than 50% bootstrap replicates were collapsed (15). The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test are shown next to the branches. The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. All positions containing gaps and missing data were eliminated and there were a total of 2424 positions in the final dataset. Evolutionary analyses were conducted in MEGA5 (51).




(a) Cluster 2 - 3 *hmwA* sequences




(b) Cluster 1 *hmwA* sequences

Figure 4.S4. Estimated omega (ω = dN/dS) by codon for (a) Cluster 2-3 and (b) Cluster 1 *hmwA* mature protein coding region. The mean (black line) and higher and lower 95% credible intervals (grey lines) for ω were estimated by omegaMap; all values have been $\log_{10}$ transformed. Log(ω) values < 0 indicate regions under purifying (negative) selection, = 0 neutral evolution, and > 0 positive selection. Amino acid positions are plotted with reference to NTHi strain 12 (a) HMW1 or (b) HMW2 protein (11) are indicated beneath the plot. The majority of sites predicted to be under positive selection (Table 4.3) were localized to the binding domain region of the mature HMW protein and are identified above the plot with either blue (omegaMap) or green (MEME) diamonds. All regions containing gaps in the multiple sequence alignment, *e.g.*, between amino acids 1170 – 1290 of Cluster 1, were coded as ω = 1.

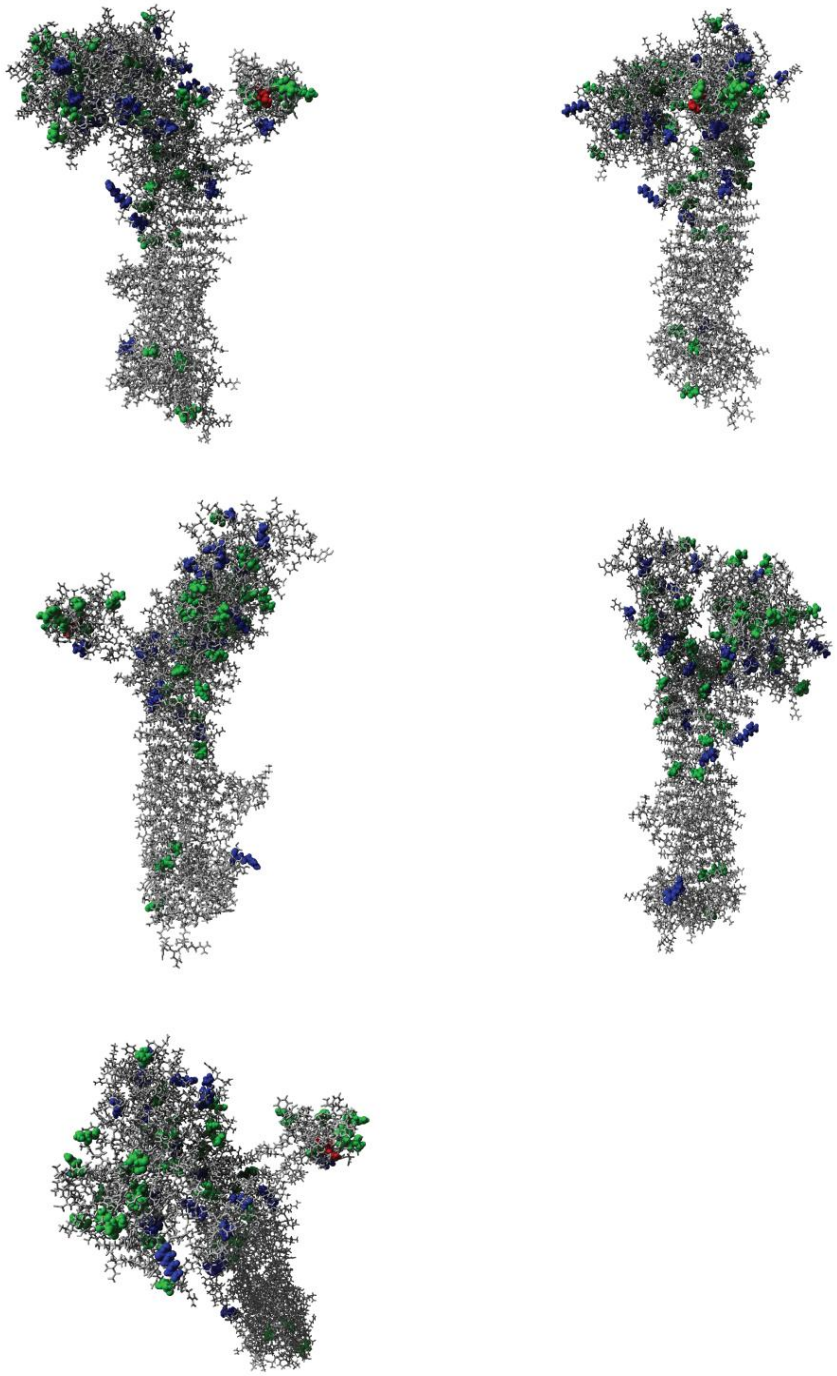


Figure 4.S5. HMW2 predicted three dimensional model. Amino acids predicted to be glycosylated based on the NX(S/T) consensus sequence (23) are highlighted in green and positively selected amino acids, as identified by using the MEME method, are highlighted in purple. There was a single glycosylated amino acid that was also predicted to be positively selected (highlighted in red).

## Literature Cited


1. Barenkamp, S. J., and F. F. Bodor. 1990. Development of serum bactericidal activity following nontypable *Haemophilus influenzae* acute otitis media. Pediatr Infect Dis J 9:333-9.
2. Barenkamp, S. J., and E. Leininger. 1992. Cloning, expression, and DNA sequence analysis of genes encoding nontypeable *Haemophilus influenzae* high-molecular-weight surface-exposed proteins related to filamentous hemagglutinin of *Bordetella pertussis*. Infect Immun 60:1302-13.
3. Barenkamp, S. J., and J. W. St Geme, 3rd. 1994. Genes encoding high-molecular-weight adhesion proteins of nontypeable *Haemophilus influenzae* are part of gene clusters. Infect Immun 62:3320-8.
4. Barenkamp, S. J., and J. W. St Geme, 3rd. 1996. Identification of a second family of high-molecular-weight adhesion proteins expressed by non-typable *Haemophilus influenzae*. Mol Microbiol 19:1215-23.
5. Barenkamp, S. J., and J. W. St Geme, 3rd. 1996. Identification of surface-exposed B-cell epitopes on high molecular-weight adhesion proteins of nontypeable *Haemophilus influenzae*. Infect Immun 64:3032-7.
6. Bruen, T. C., H. Philippe, and D. Bryant. 2006. A simple and robust statistical test for detecting the presence of recombination. Genetics 172:2665-81.
7. Bryant, D., and V. Moulton. 2004. Neighbor-net: an agglomerative method for the construction of phylogenetic networks. Mol Biol Evol 21:255-65.
8. Buscher, A. Z., K. Burmeister, S. J. Barenkamp, and J. W. St Geme, 3rd. 2004. Evolutionary and functional relationships among the nontypeable *Haemophilus influenzae* HMW family of adhesins. J Bacteriol 186:4209-17.
9. Cody, A. J., D. Field, E. J. Feil, S. Stringer, M. E. Deadman, A. G. Tsolaki, B. Gratz, V. Bouchet, R. Goldstein, D. W. Hood, and E. R. Moxon. 2003. High rates of recombination in otitis media isolates of non-typeable Haemophilus influenzae. Infect Genet Evol 3:57-66.
10. Connor, T. R., J. Corander, and W. P. Hanage. 2012. Population subdivision and the detection of recombination in non-typable Haemophilus influenzae. Microbiology 158:2958-64.
11. Dawid, S., S. Grass, and J. W. St Geme, 3rd. 2001. Mapping of binding domains of nontypeable *Haemophilus influenzae* HMW1 and HMW2 adhesins. Infect Immun 69:307-14.
12. Ecevit, I. Z., K. W. McCrea, C. F. Marrs, and J. R. Gilsdorf. 2005. Identification of new hmwA alleles from nontypeable *Haemophilus influenzae*. Infect Immun 73:1221-5.
13. Ecevit, I. Z., K. W. McCrea, M. M. Pettigrew, A. Sen, C. F. Marrs, and J. R. Gilsdorf. 2004. Prevalence of the *hifBC, hmw1A, hmw2A, hmwC*, and hia Genes in *Haemophilus influenzae* Isolates. J Clin Microbiol 42:3065-72.
14. Edgar, R. C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792-7.
15. Felsenstein, J. 1985. Confidence limits on phylogenies: An approach using the bootstrap. Evolution 39:783-791.

16. Fitch, W. M. 1986. An estimation of the number of invariable sites is necessary for the accurate estimation of the number of nucleotide substitutions since a common ancestor. Prog Clin Biol Res 218:149-59.
17. Fitch, W. M., and E. Margoliash. 1967. A method for estimating the number of invariant amino acid coding positions in a gene using cytochrome c as a model case. Biochem Genet 1:65-71.
18. Giufre, M., A. Carattoli, R. Cardines, P. Mastrantonio, and M. Cerquetti. 2008. Variation in expression of HMW1 and HMW2 adhesins in invasive nontypeable *Haemophilus influenzae* isolates. BMC Microbiol 8:83.
19. Giufre, M., M. Muscillo, P. Spigaglia, R. Cardines, P. Mastrantonio, and M. Cerquetti. 2006. Conservation and diversity of HMW1 and HMW2 adhesin binding domains among invasive nontypeable *Haemophilus influenzae* isolates. Infect Immun 74:1161-70.
20. Grass, S., A. Z. Buscher, W. E. Swords, M. A. Apicella, S. J. Barenkamp, N. Ozchlewski, and J. W. St Geme, 3rd. 2003. The *Haemophilus influenzae* HMW1 adhesin is glycosylated in a process that requires HMW1C and phosphoglucomutase, an enzyme involved in lipooligosaccharide biosynthesis. Mol Microbiol 48:737-51.
21. Grass, S., C. F. Lichti, R. R. Townsend, J. Gross, and J. W. St Geme, 3rd. 2010. The Haemophilus influenzae HMW1C protein is a glycosyltransferase that transfers hexose residues to asparagine sites in the HMW1 adhesin. PLoS Pathog 6:e1000919.
22. Grass, S., and J. W. St Geme, 3rd. 2000. Maturation and secretion of the non-typable *Haemophilus influenzae* HMW1 adhesin: roles of the N-terminal and C-terminal domains. Mol Microbiol 36:55-67.
23. Gross, J., S. Grass, A. E. Davis, P. Gilmore-Erdmann, R. R. Townsend, and J. W. St Geme, 3rd. 2008. The Haemophilus influenzae HMW1 adhesin is a glycoprotein with an unusual N-linked carbohydrate modification. J Biol Chem 283:26010-5.
24. Harrison, A., D. W. Dyer, A. Gillaspy, W. C. Ray, R. Mungur, M. B. Carson, H. Zhong, J. Gipson, M. Gipson, L. S. Johnson, L. Lewis, L. O. Bakaletz, and R. S. Munson, Jr. 2005. Genomic sequence of an otitis media isolate of nontypeable *Haemophilus influenzae*: comparative study with *H. influenzae* serotype d, strain KW20. J Bacteriol 187:4627-36.
25. Hogg, J. S., F. Z. Hu, B. Janto, R. Boissy, J. Hayes, R. Keefe, J. C. Post, and G. 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 Biol 8:R103.
26. Hopp, T. P., and K. R. Woods. 1981. Prediction of protein antigenic determinants from amino acid sequences. Proc Natl Acad Sci U S A 78:3824-8.
27. Huson, D. H., and D. Bryant. 2006. Application of phylogenetic networks in evolutionary studies. Mol Biol Evol 23:254-67.
28. Khan, S., H. S. Mian, L. E. Sandercock, N. Y. Chirgadze, and E. F. Pai. 2011. Crystal structure of the passenger domain of the Escherichia coli autotransporter EspP. J Mol Biol 413:985-1000.
29. Kosakovsky Pond, S. L., B. Murrell, M. Fourment, S. D. Frost, W. Delport, and K. Scheffler. 2011. A random effects branch-site model for detecting episodic diversifying selection. Mol Biol Evol 28:3033-43.

30. Kosakovsky Pond, S. L., D. Posada, M. B. Gravenor, C. H. Woelk, and S. D. Frost. 2006. Automated phylogenetic detection of recombination using a genetic algorithm. Mol Biol Evol 23:1891-901.
31. Kosakovsky Pond, S. L., D. Posada, M. B. Gravenor, C. H. Woelk, and S. D. Frost. 2006. GARD: a genetic algorithm for recombination detection. Bioinformatics 22:3096-8.
32. Krieger, E., K. Joo, J. Lee, S. Raman, J. Thompson, M. Tyka, D. Baker, and K. Karplus. 2009. Improving physical realism, stereochemistry, and side-chain accuracy in homology modeling: Four approaches that performed well in CASP8. Proteins 77 Suppl 9:114-22.
33. Kyte, J., and R. F. Doolittle. 1982. A simple method for displaying the hydropathic character of a protein. J Mol Biol 157:105-32.
34. Lacross, N. C., C. F. Marrs, and J. R. Gilsdorf. 2013. Population Structure in Nontypeable Haemophilus influenzae. Infect Genet Evol.
35. Li, Y., T. Gierahn, C. M. Thompson, K. Trzcinski, C. B. Ford, N. Croucher, P. Gouveia, J. B. Flechtner, R. Malley, and M. Lipsitch. 2012. Distinct Effects on Diversifying Selection by Two Mechanisms of Immunity against Streptococcus pneumoniae. PLoS Pathog 8:e1002989.
36. Loytynoja, A., and N. Goldman. 2010. webPRANK: a phylogeny-aware multiple sequence aligner with interactive alignment browser. BMC Bioinformatics 11:579.
37. Murrell, B., J. O. Wertheim, S. Moola, T. Weighill, K. Scheffler, and S. L. Kosakovsky Pond. 2012. Detecting individual sites subject to episodic diversifying selection. PLoS Genet 8:e1002764.
38. Nei, M., and S. Kumar. 2000. Molecular Evolution and Phylogenetics. Oxford University Press, New York.
39. Otto, B. R., R. Sijbrandi, J. Luirink, B. Oudega, J. G. Heddle, K. Mizutani, S. Y. Park, and J. R. Tame. 2005. Crystal structure of hemoglobin protease, a heme binding autotransporter protein from pathogenic Escherichia coli. J Biol Chem 280:17339-45.
40. Roy, A., A. Kucukural, and Y. Zhang. 2010. I-TASSER: a unified platform for automated protein structure and function prediction. Nat Protoc 5:725-38.
41. Schrodinger, L. 2010. The PyMOL Molecular Graphics System, version 1.3r1.
42. St Geme, J. W., 3rd. 1994. The HMW1 adhesin of nontypeable *Haemophilus influenzae* recognizes sialylated glycoprotein receptors on cultured human epithelial cells. Infect Immun 62:3881-9.
43. St Geme, J. W., 3rd, and D. Cutter. 1995. Evidence that surface fibrils expressed by Haemophilus influenzae type b promote attachment to human epithelial cells. Mol Microbiol 15:77-85.
44. St Geme, J. W., 3rd, and D. Cutter. 1996. Influence of pili, fibrils, and capsule on in vitro adherence by Haemophilus influenzae type b. Mol Microbiol 21:21-31.
45. St Geme, J. W., 3rd, M. L. de la Morena, and S. Falkow. 1994. A *Haemophilus influenzae* IgA protease-like protein promotes intimate interaction with human epithelial cells. Mol Microbiol 14:217-33.
46. St Geme, J. W., 3rd, S. Falkow, and S. J. Barenkamp. 1993. High-molecular-weight proteins of nontypable *Haemophilus influenzae* mediate attachment to human epithelial cells. Proc Natl Acad Sci U S A 90:2875-9.
47. St Geme, J. W., 3rd, and S. Grass. 1998. Secretion of the Haemophilus influenzae HMW1 and HMW2 adhesins involves a periplasmic intermediate and requires the HMWB and HMWC proteins. Mol Microbiol 27:617-30.

48. St Geme, J. W., 3rd, V. V. Kumar, D. Cutter, and S. J. Barenkamp. 1998. Prevalence and distribution of the *hmw* and *hia* genes and the HMW and Hia adhesins among genetically diverse strains of nontypeable *Haemophilus influenzae*. Infect Immun 66:364-8.
49. Swords, W. E., B. A. Buscher, K. Ver Steeg Ii, A. Preston, W. A. Nichols, J. N. Weiser, B. W. Gibson, and M. A. Apicella. 2000. Non-typeable Haemophilus influenzae adhere to and invade human bronchial epithelial cells via an interaction of lipooligosaccharide with the PAF receptor. Mol Microbiol 37:13-27.
50. Tamura, K., M. Nei, and S. Kumar. 2004. Prospects for inferring very large phylogenies by using the neighbor-joining method. Proc Natl Acad Sci U S A 101:11030-5.
51. Tamura, K., D. Peterson, N. Peterson, G. Stecher, M. Nei, and S. Kumar. 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol 28:2731-9.
52. van Schilfgaarde, M., P. van Ulsen, P. Eijk, M. Brand, M. Stam, J. Kouame, L. van Alphen, and J. Dankert. 2000. Characterization of adherence of nontypeable *Haemophilus influenzae* to human epithelial cells. Infect Immun 68:4658-65.
53. Wilson, D. J., and G. McVean. 2006. Estimating diversifying selection and functional constraint in the presence of recombination. Genetics 172:1411-25.
54. Winter, L. E., and S. J. Barenkamp. 2006. Antibodies specific for the high-molecular-weight adhesion proteins of nontypeable *Haemophilus influenzae* are opsonophagocytic for both homologous and heterologous strains. Clin Vaccine Immunol 13:1333-42.
55. Winter, L. E., and S. J. Barenkamp. 2010. Construction and immunogenicity of recombinant adenovirus vaccines expressing the HMW1/HMW2 or Hia adhesion proteins of nontypeable Haemophilus influenzae. Clin Vaccine Immunol.
56. Winter, L. E., and S. J. Barenkamp. 2003. Human antibodies specific for the high-molecular-weight adhesion proteins of nontypeable *Haemophilus influenzae* mediate opsonophagocytic activity. Infect Immun 71:6884-91.
57. Xia, X., and Z. Xie. 2001. DAMBE: software package for data analysis in molecular biology and evolution. J Hered 92:371-3.
58. Xia, X., Z. Xie, M. Salemi, L. Chen, and Y. Wang. 2003. An index of substitution saturation and its application. Mol Phylogenet Evol 26:1-7.
59. Yang, Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J Mol Evol 39:306-14.
60. Zhang, Y. 2008. I-TASSER server for protein 3D structure prediction. BMC Bioinformatics 9:40.

# Chapter 5

## Phase Variation and Host Immunity Against High Molecular Weight (HMW) Adhesins Shape Population Dynamics of Nontypeable *Haemophilus influenzae* Within Human Hosts [4]


### Abstract

Nontypeable *Haemophilus influenzae* (NTHi) is a gram-negative bacterium that commonly resides within the human pharynx. Because NTHi is human-restricted, its long term survival is dependent upon its ability to successfully colonize new hosts. Adherence to host epithelium, mediated by bacterial adhesins, is one of the first steps in NTHi colonization. NTHi express several adhesins, including the high molecular weight (HMW) adhesins that mediate attachment to the respiratory epithelium where they interact with the host immune system to elicit a strong humoral response. *hmwA*, which encodes HMW, undergoes phase variation mediated by 7-base pair tandem repeats located within its promoter region. Repeat number affects both *hmwA* transcription and HMW production such that as the number of tandem repeats increases HMW production decreases. Cells expressing large amounts of HMW may be critical for the establishment and maintenance of NTHi colonization, but they might also incur greater fitness costs when faced with an HMW-specific antibody mediated immune response. We hypothesized that the occurrence of large deletion events allows NTHi to maintain adherent cells in the presence of antibody-mediated immunity. To study this, we developed a deterministic mathematical model, incorporating *hmwA* phase variation and antibody mediated immunity, to explore the trade-off between bacterial adherence and immune evasion. We employed uncertainty and sensitivity analyses to calibrate the model and to identify mechanisms driving NTHi within-host dynamics. The model predicts that antibody levels and avidity, catastrophic loss rates, and population carrying capacity all significantly affected numbers of adherent NTHi


[4] This work has been prepared for submission to the Journal of Theoretical Biology with the following authors: Davis GS, Marino S, Marrs CF, Gilsdorf JR, Dawid S, and Kirschner DK.

cells within a host. Effects of these mechanisms vary in both time and magnitude. These results suggest that the occurrence of large, yet rare, deletion events allows for stable maintenance of a small population of adherent cells in spite of HMW-specific antibody induced immunity. These adherent subpopulations may be important for sustaining colonization and/or maintaining transmission.

## Introduction

*Haemophilus influenzae* is a gram-negative coccobacillus that commonly resides within the human pharynx as a commensal and as a potential pathogen. These bacteria are differentiated into typeable and nontypeable strains based on the presence or absence, respectively, of a polysaccharide capsule. Encapsulated strains are divided into six capsular serotypes (a–f), with serotype b (Hib) being commonly associated with invasive diseases such as meningitis and septicemia among non-immune children. In contrast, nonencapsulated strains, which are commonly referred to as nontypeable *H. influenzae* (NTHi), are generally associated with localized infections of the respiratory tract such as pneumonia, sinusitis, and acute otitis media (AOM). AOM is a common childhood disease and in the United States, approximately 83% of children have had at least one episode of AOM by the age of three and 45% have suffered three or more AOM episodes (65). In adults, NTHi strains are commonly associated with acute exacerbations in patients suffering from chronic obstructive pulmonary disease (COPD) (25, 50, 55).

NTHi diseases impose a significant burden on the health care system. According to 2008 WHO estimates, COPD was responsible for 3.28 million deaths worldwide, making it the fourth leading cause of death (49, accessed September 19, 2012). In the US alone, the total economic cost of COPD/asthma for 2008 was estimated to be $68 billion (31). AOM also imposes a significant burden on the health care system in the US with a total annual estimated cost of approximately $4.2 billion during 2009 (48). Furthermore, both acute exacerbations in COPD patients and AOM often result in antibiotic prescriptions (39, 52). Thus, reducing incidence of NTHi-associated diseases can reduce a significant burden on the healthcare system and has the potential to reduce antibiotic usage and associated concerns regarding emerging antibiotic resistance. To these ends, current efforts to reduce NTHi disease incidence are primarily focused on preventive measures, such as vaccines.

NTHi strains are spread from person to person via infected respiratory droplets where they then establish pharyngeal colonization. Among healthy children, colonization generally occurs early during childhood and colonization prevalence ranges between 25 – 84% (11, 23, 30). Pathogenic NTHi arise from the community of NTHi strains colonizing healthy individuals.

Thus, colonization is one of the first steps of NTHi pathogenesis and interventions that either reduce or prevent colonization would therefore likely decrease the burden of NTHi disease.

Adherence to the host epithelium is one of the first steps in bacterial colonization (43). Pharyngeal colonization requires that a newly transmitted strain successfully overcome the host's mucociliary clearance mechanisms and innate immunity. Adhesins are a class of cell surface structures used by bacteria, including NTHi, to attach to substrates such as the host epithelium. During initial stages of colonization adhesin-mediated adherence likely plays a critical role in establishment of colonization, allowing newly transmitted cells to avoid being removed from the pharynx by mucociliary clearance mechanisms. These adhesins, however, can also become a liability for the bacteria as they tend to be key antigens that attract the attention of the host adaptive immunity and stimulate an antibody mediated immune response. As demonstrated in numerous studies, NTHi colonization and disease stimulates an adaptive immune response resulting in the presence of serum antibodies such as IgG, IgM, and IgA, many of which specifically target NTHi outer membrane proteins (4, 5, 28, 34, 51). Thus, colonization lasting for more than a few days requires that the bacterial population overcome the additional selective pressures applied by antibody-mediated immunity.

NTHi adherence to the host epithelium is mediated by a number of surface exposed pilin and non-pilin adhesins, including the high molecular weight (HMW) adhesins (57-60, 63). HMW adhesins are present in approximately 40-75% of all NTHi strains (6, 20-22, 61, 67) and when present, they are the immunodominant outer membrane proteins (5). Functional HMW adhesins are encoded by *hmwA*, which resides within the *hmw* locus and is present in two copies in conserved but unlinked regions of the NTHi chromosome (6, 12, 16, 27). This diversity serves two roles, first it helps define the tissue tropism of a particular strain, for example, the two copies of HMW encoded by a single strain generally confer different *in vitro* binding characteristics (12, 60). Furthermore, anti-HMW adhesin antibodies display varying degrees of cross reactivity against the HMW adhesins from heterologous NTHi strains (67). Amino acid diversity therefore confers antigenic diversity, and, this may be critical for ensuring availability of immunologically naive hosts given that HMW adhesins are highly immunogenic (5).

The HMW adhesins are phase variable. HMW adhesin phase variation is mediated by simple sequence repeats (SSRs) of seven base-pairs (bp) located within the *hmwA* promoter

region (6, 15). Reversible changes in repeat number affect HMW adhesin levels such that as repeat number increases *hmwA* transcript levels, and HMW protein levels, decrease in a graded fashion (15). Since HMW adhesins are highly immunogenic (5), phase variable reductions in HMW adhesin expression provides a mechanism by which NTHi populations are able to limit, or possibly even evade, antibody mediated immunity. Because phase variable changes in repeat numbers is a stochastic process, that occurs independently of any selective pressure, an immune evasive NTHi population retains the potential for regaining an adherent phenotype.

Differential expression of HMW adhesins is associated with NTHi pathogenicity during both AOM (15) and COPD (14). In a study of paired isolates, those collected from the middle ears of children with AOM had a higher number of SSRs, and lower levels of HMW adhesin, than did their isogenic counterparts isolated at the same time from the nasopharynx (15). Serial isolates from the sputum of individuals with COPD revealed that over time the SSR number increased, which was often associated with decreased HMW adhesin levels and, in some instances, significantly decreased adherence to human respiratory epithelial cells in vitro (14). Further support for phase variation-mediated immune evasion comes from the results of an HMW-targeted NTHi vaccine study in which isogenic NTHi challenge strains with increased *hmwA* repeat number, and thereby decreased HMW levels, were able to evade HMW antibody-mediated killing *in vivo* (2, 15).

The fitness of a NTHi strain is dependent upon its ability to transmit to and colonize, at least transiently, new hosts. With respect to the HMW ahesins, a NTHi population can display a spectrum of phenotypes ranging from adherent (few repeats and high HMW adhesin levels) to immune evasive (increased repeat numbers and relatively low HMW adhesin levels; Fig. 5.1a). At one end of this spectrum, high HMW adhesin levels might be important for transmission or for maintaining colonization. At the other end, NTHi population displaying low HMW adhesin levels and a non-adherent, immune evasive, phenotype may be favored during colonization. The distribution of a NTHi population along the spectrum of HMW adhesin levels likely reflects an evolutionary tradeoff between the population's ability to transmit between hosts versus its ability to maintain colonization in spite of antibody mediated immunity. Thus, an open question is: how does a colonizing NTHi population retain HMW-mediated host adherence, when within-host selective pressures favor a non-adherent, immune evasive, NTHi population? One solution to this

challenge is for a NTHi lineage to be rapidly transmitted between hosts (*i.e.*, serially transmitted), before antibody-mediated immune pressure drives HMW adhesins to levels that might preclude efficient adherence. Empirical evidence, however, does not provide strong support for serial transmissions. If strains were rapidly transmitted among individuals one might expect that at any given time there would be a high degree of sharing among close contacts. While there is some evidence for sharing of strains among adults, within families, and among children attending the same day care, this seems to be the exception rather than the rule (18, 23, 29, 44, 47, 54, 56, 62, 66). In fact, NTHi colonizing strains stand out for their degree of genetic diversity within and between hosts. An alternative strategy to serial transmission is that phenotypically immune evasive NTHi populations, displaying low levels of HMW, have a mechanism for regaining, or maintaining, an adherent phenotype during colonization.

It is the second hypothesis that we explore in this study. We hypothesize that the occurrence of large, yet rare, deletions within the repetitive DNA region that controls *hmwA* transcription allows for the maintenance of a population of adherent cells even when the population is faced with a robust antibody-mediated host response. We explore this hypothesis by developing a simple mathematical model consisting of a system of ordinary differential equations that describe a within-host NTHi population and the antibody mediated immune response of a naïve host. Our conclusions suggest that this hypothesis can indeed explain the spectrum of phenotypes and their role in survival within the human host.

## The Model

Our goal is to represent, using a mathematical model, NTHi within-host population dynamics. The NTHi population will be stratified with respect to the number of repeats in their *hmwA* promoter region (Fig. 5.1b), which we assume is a proxy for the amount of surface expressed HMW adhesin and is proportional to a bacterial cell's ability to adhere to host epithelium. We assume that each bacterial cell possesses a single *hmw* locus and we divide the total NTHi population into subpopulations based upon the exact number of 7-bp SSRs in the *hmwA* promoter. Our model consists of 17 NTHi subpopulations $REP_i(t)$ (i = repeat number) and the host antibody immune response IR(t). Interactions between populations and their rates of change are described by a system of 18 ordinary differential equations (ODEs). As the equations are similar for each NTHi subpopulations, we describe a representative equation below and refer the reader to the Appendix for the full list of equations. Below we describe characteristics of the model and biological support for the assumptions.

(Eq 5.1)

**NTHi population.** The total NTHi population size within a single host is constrained by the carrying capacity, K, of the pharynx. The 17 subpopulations NTHi subpopulations are defined by the exact number of 7-bp SSRs, ranging from 12 to 28, within the *hmwA* promoter. The number of NTHi subpopulations was chosen to capture the majority of the range in repeat numbers observed in population-level studies (3, 15, 19, 26, 27, 67). During DNA replication, the number or repeats in the *hmwA* promoter region can remain unchanged or a mutational event can occur in which one or more repeats are gained or lost via slipped strand mispairing (15). If repeats are gained or lost, the resultant daughter cell transitions into a new subpopulation defined by the cell's SSR number. Since HMW adhesin levels are associated with a bacterial cell's

ability to adhere to host epithelial cells (14, 15), NTHi subpopulations can be characterized phenotypically as adherent (*i.e.*, encoding few repeats and producing high levels of HMW), or immune evasive, (*i.e.*, encoding several repeats and producing low levels of HMW). Based on these data, we define "adherent" bacterial cells as those in subpopulations 1 – 3 (encoding 12 – 14 repeats, respectively) and "evasive" bacterial cells as those in subpopulation 15 – 17 (encoding 26 – 28 repeats, respectively).

Phase variation rates for *hmwA*-associated 7-bp repeats have not been determined. There is, however, empirical data for phase variation rates for the NTHi *mod* gene which, in a manner similar to *hmwA*, possesses variable numbers of tetranucleotide SSRs (17). Both *hmwA* and *mod* SSRs are gained or lost by the same mechanism, namely slipped stranded mispairing (15, 17). Analysis of *mod* phase variation suggests that: (1) the most common type of event is the gain or loss of a single repeat, (2) repeats are more likely lost than gained, and (3) that for any type of event, as the number of the SSRs increases, so too does the mutation rate (17). Interestingly, De Bolle *et al.* (17) documented two large deletion events in strains with large SSR numbers. In a study of dinucleotide repeats, Morel *et al.* (45) also documented large yet rare deletion events in $(AC)_{51}$ SSR tracts inserted onto the *E. coli* chromosome. Since the number of SSRs in the *hmwA* promoter affects HMW adhesin expression levels, the simultaneous deletion of several SSRs provides a mechanism by which a parent-cell with a long tract of SSRs (an immune evasive phenotype) can produce a daughter cell with a drastically reduced number of SSRs (an adherent phenotype). We refer to the simultaneous loss of several repeats as a "catastrophic loss" event.

**Model parameterization**. Based on the study by De Bolle *et al*., we defined five different phase variation events in our ordinary differential equation model (Appendix A Table 5.A1 and Table 5.A2) (17). First, bacterial cells can gain or lose a single repeat at rates $\alpha_i$ and $\beta_k$, respectively, where *i* designates subpopulations 1 to 16 and *k* designates subpopulations 1 to 15. Bacterial cells can also gain or lose two repeats in a single event with rates $\delta_m$ and $\chi_w$, respectively, where *m* designates subpopulations 2 to 17 and *w* designates subpopulations 3 to 17. As reported by De Bolle with respect to tetranucleotide repeats (17), for each event type, the rate at which repeats are lost is assumed to be greater than the rate at which they are gained and the mutation rate increases as the number of SSRs increases. The length of the repeat tract therefore determines the level of HMW adhesin expression as well as the rate at which

expression levels phase vary. Finally, bacterial cells with the largest number of SSRs (*i.e.*, 26, 27, or 28 repeats), can lose several repeats in a single event at rate $\varphi_j$, where $j$ = subpopulations 15, 16, or 17. These large deletion events (catastrophic losses) result in daughter cells transitioning from a subpopulation with an immune evasive phenotype to one with an adherent phenotype in a single step. Catastrophic losses are the rarest event type. Consistent with the results of De Bolle *et al.* (17), for a given number of repeats, we assumed $\beta_k > \alpha_i > \chi_i > \delta_m > \varphi_j$.

NTHi effective growth rate, $\tau$, is a function of bacterial cell replication rates, death rates, and clearance rates from the pharynx (*e.g.*, by host mucociliary clearance mechanisms). We assumed that NTHi replication rates are the same for each NTHi subpopulation meaning that repeat number does not affect replication rate. Bacterial cell death rates are divided into subpopulation-independent mechanisms and subpopulation-specific mechanisms; for example, HMW adhesin-specific antibody-mediated adaptive immunity is subpopulation specific. We incorporated mucociliary clearance into our model by specifying subpopulation specific clearance rates, $cr_{max}$, that were greatest for non-adherent cells and negligible for adherent cells. Thus, subpopulation specific effective growth rates, $\tau_n$, are a function of replication rates, subpopulation independent cell death rates, and subpopulation specific clearance rates where $n$ = subpopulations 1 -17 and $\tau_{n < } \tau_{n-1}$.

HMW adhesins stimulate a host mediated bactericidal antibody response (5) and this likely imposes a fitness cost on HMW adhesin-producing cells. Cells displaying high levels of surface exposed HMW adhesin present more antigens to the host immune system than do cells with lower HMW adhesin levels and as a consequence they likely experience higher HMW adhesin-specific antibody-mediated death rates. Both HMW adhesin levels, and *hmwA* transcription, decrease as the number of 7-bp repeats increase (14, 15, 26). To capture this dynamic, we assumed that the rate of antibody-mediated killing is proportional to *hmwA* transcription. We defined the term $\mu_n$, where $n$ = subpopulation 1 – 17, that describes the relationship between *hmwA* repeat number and surface exposed HMW adhesin level. The host immune response, described below, directly interacts with the NTHi subpopulation via $\mu$.

Once the model was developed, baseline rates for the various parameters were estimated. Rates were estimated from published literature when available and in the absence of published data, we employed uncertainty analysis (42) to define baseline parameter estimates that gave rise

to biologically reasonable outputs. All parameter values are summarized in Table 5.A1 and Table 5.A2 of the Appendix. The details of how each parameter was estimated are given below.

**Sensitivity analyses.** There is an intrinsic variability in many of the parameter values of our mathematical model, due to extensive variability in the data, as well as uncertainty regarding their in vivo values, which often are incomplete. We address this uncertainty by coupling an extensive and efficient sampling of the parameter space with a generalized correlation methodology to assess effects of uncertainties in our parameter estimation on model outcomes.

Sensitivity analysis (SA) is a method for quantifying uncertainty in any type of complex model. The objective of SA is to identify critical inputs (parameters and initial conditions) of a model and to quantify how input uncertainty affects model outcome(s). We use Latin Hypercube Sampling (LHS) as a sampling scheme and Partial Rank Correlation Coefficient (PRCC) as a sensitivity index. Uncertainty and sensitivity analysis methodologies are described in detail in (42) and briefly below.

LHS is a so-called *stratified sampling without replacement* technique, where the random parameter distributions are divided into *N* equal probability intervals, which are then independently sampled. *N* represents the sample size, which here is set to 5000. A matrix of 5000 rows and k number of columns (k is equal to the number of parameters varied in the analysis) is generated. Each row of the matrix serves as input for a single model simulation. The outputs of N model simulations are then saved and used for calculating sensitivity indexes. Since our model is a dynamical system, 20-day time course outputs are generated. PRCC was then calculated for each parameter and statistical significance was assessed (42). Our outcome of interest for sensitivity analyses was the total number of adherent cells (Table 5.A1), defined as the sum of the number NTHi cells in subpopulations one to three.

**Computer simulations**. Once the parameters were estimated, we solved the system of ordinary differential equations (ODEs) to track NTHi population dynamics over time. All simulations were performed using MATLAB's (ver 7.10, R2010a, Copyright 1984-2010, The MathWorks, Inc.) ode113 solver for non-stiff differential equations.

There are no data on the number of NTHi cells required to successfully colonize a host, but, both NTHi and host factors likely play a role in determining the infectious dose. We chose

to study colonization dynamics under the assumption that 120 adherent NTHi cells, distributed across the first three NTHi subpopulations, enter the pharynx; our model is flexible and stable to variations in initial conditions. We consider the total number of adherent bacterial cells (*i.e.*, sum of cell numbers in subpopulations 1 – 3) as a key readout for the system.

## Results

In the absence of antibody-mediated immunity, the distribution of NTHi cells among subpopulations was solely a function of phase variation rates. Under baseline conditions, the NTHi total population distribution gradually shifted so that at the end of the 20 day simulation subpopulations 1 through 8 contained nonzero levels of bacterial cells (Fig. 5.2a and Fig. 5.A3). Throughout the simulation, the majority (99.6%) of the total population remained within the first three subpopulations (*i.e.*, the adherent phenotype) (Fig. 5.2a and Fig. 5.A3). At equilibrium, all seventeen NTHi subpopulations contained a positive number of bacterial cells and the total population distribution was slightly skewed; however, with the first three subpopulations comprised 80.2% of the total population (data not shown). The distribution at equilibrium reflected the overall mutation rates which favored the loss of repeats over their gain (Fig. 5.A1). Because the mutation rate for the subpopulation containing 12 repeats was nonzero, there were always NTHi subpopulations with longer repeat tracts, even though deletions were more likely than insertions. In the absence of host-mediated immunity, catastrophic loss events did not impact the overall NTHi population distribution during the 20-day simulation.

Next we explored the effect of an immune response (Fig. 5.3) on the NTHi population distribution in the presence and absence of catastrophic loss events (Fig. 5.2b-c, Fig. 5.A4, and Fig. 5.A5). Baseline simulations were initiated with an immune response of 0.001, which resulted in 0.001% of the total population being killed on day 1. When compared to the effect of mutation rates alone, however, the immune response had a negligible effect until day 10 (Fig. 5.2 and Figs. 5.A3-5.A5). The immune response was approximately 50% of its maximal level on day 10 (Fig. 5.3) and the number of adherent cells in the population dropped from approximately 1E5 on day 11 to less than a single cell on day 12. As the immune response continued to develop, immune-mediated selective pressure forced the overall NTHi population distribution toward subpopulations with increased repeat number and decreased HMW levels (Fig.5. 2 and Figs. 5.A3-5.A5). In the absence of catastrophic losses, adherent cells were eliminated from the population after day 12 and were never regained during the remainder of the 20 day simulation (Fig. 5.2b, Fig. 5.A4). Inclusion of catastrophic loss events had little impact on the NTHi population distribution through day 16, but by day 17 a small population of adherent cells was

established and it increased each day until the end of the 20-day simulation (Fig. 5.2c, Fig. 5.4, and Fig. 5.A5).

The adherent NTHi population at day 20 under baseline model conditions consisted of approximately 92 NTHi cells (Fig. 5.4). This raised concern given that the final population was lower than the initial inoculum used for our baseline simulations; that is, if there are only 92 adherent cells total how could colonization be initiated with an initial inoculum of 120 cells? To determine if our model results were sensitive to changes in the initial conditions, we performed simulations with varying numbers of input bacterial cells. Even with an inoculum size of one cell, the NTHi population approaches the baseline carrying capacity of 1E09 by the end of the first day, long before the immune response influenced the population distribution (data not shown). We therefore concluded that our results are insensitive to the initial inoculum size.

**Sensitivity analyses.** To identify parameters that significantly impacted our outcome variable of interest, sensitivity analyses were conducted on the model with parameter values from Table A1 and ranges given therein (see Methods for details). Maximum antibody levels (*Ab_max*), population carrying capacity (*K*), and antibody avidity (*s*) were each significantly correlated with the numbers of adherent cells ($P < 0.01$) (Fig. 5.5). The impact of their effects, however, varied in both time and magnitude.

Pharyngeal carrying capacity, *K*, was strongly correlated with the number of adherent cells (Fig. 5.5**)**. At all time points, there was a statistically significant ($P < 0.001$) positive correlation between carrying capacity and the number of adherent cells. During the early stages of colonization, prior to development of the adaptive immune response, the NTHi population distribution was simply a function of mutation rates. Therefore, during the early stages of the simulation most NTHi cells were located within the first three subpopulations (*e.g.*, on day 10, adherent cells make up over 99% of the total NTHi population) and changes in the number of adherent cells were directly proportional to carrying capacity, *K*. The intense selective pressure applied by the immune response rapidly eliminated the adherent subpopulations such that from days 12 to 15, there was less than one adherent cell in the population (Fig. 5.4). The positive correlation between adherent cells and carrying capacity during this period when there was less than a single cell in the model resulted from the assumption that the population acts like a continuous function. During the later stages of the simulation, the number of adherent cells was a

function of both carrying capacity and the catastrophic loss rate (*intercept_cl*). By day 20 under baseline conditions, there were approximately 4.9E7 immune evasive cells in subpopulations 15 – 17; that is, the subpopulations capable of undergoing catastrophic losses to produce adherent daughter cells. For a given catastrophic loss rate, increasing the pharyngeal carrying capacity (*K*), increased the number of adherent cells.

The catastrophic loss rate (*intercept_cl*) also significantly impacted numbers of adherent cells. For the sensitivity analysis, the catastrophic loss rates were varied independently of all other mutational events by varying the intercept of the line describing the relationship between repeat number and phase variation rate (Fig. 5.A1). The rates at which one or two repeats were gained or lost were varied by varying the scaling factor *q* (Table 5.A1), in this way the relationships between mutation rates (*i.e.,* $\beta_k > \alpha_i > \chi_i > \delta_m$) were maintained during the sensitivity analyses. The rate of catastrophic events was significantly positively correlated ($P < 0.001$) with the number of adherent cells from days 12 to 20 (Fig. 5.5), confirming our hypothesis that catastrophic loss provide a mechanism that allows for maintenance of adherent cells in the face of antibody mediated immunity. The PRCCs between the catastrophic loss rates and the numbers of adherent cells increased over time as the immune pressure forced the NTHi population to a predominantly immune evasive phenotype (Fig. 5.2b and Fig. 5.A5); the PRCC reached its maximal value of 0.26 on day 20 (Fig. 5.5).

The sensitivity analysis also found that *Ab_max* (the maximum level of anti-HMW antibodies produced in response to NTHi colonization) and the numbers of adherent cells were strongly correlated (Fig. 5.3). *Ab_max* was negatively correlated ($P < 0.001$) with the number of adherent cells from days 7 to 14 (Fig.5.5). This time period corresponds to rapidly increasing antibody levels (Fig. 5.3), immune-mediated elimination of adherent cells, and a shift in the total population distribution that favored subpopulations with increased repeat number (Fig. 5.2c and Fig. 5.A5). Beginning on day 12, subpopulations 15 to 17 became non-zero and on day 16 catastrophic loss events slowly repopulated the adherent subpopulations (Fig. 5.A5). By day 15, the relationship between *Ab_max* and the number of adherent cells became significantly positively correlated ($P < 0.001$) and continued to increase through the end of the simulation at day 20 (Fig. 5.5).

Finally, our analysis predicts that *s* (which is analogous to antibody avidity) and the number of adherent cells were also strongly correlated (Fig. 5.5). Therefore, increasing *s* increases the per unit killing rate of bacterial cells by the immune response. The effect of *s* was evident by day 5 of the simulation at which time the *s* and the number of adherent cells were significantly negatively correlated ($P < 0.001$) (Fig. 5.5). In general, the PRCCs for *s* tracked closely with those of *Ab_max* (Fig. 5.5). This was not surprising since both parameters reflect the relationship between the immune response and per unit NTHi killing.

## Discussion

In this paper we have presented a mathematical model that explores how phase variation and antibody-mediated immunity can shape the population structure of nontypeable *Haemophilus influenzae* within a host. Specifically, we focused on phase variable expression of the HMW adhesin, a surface exposed immunogenic protein that mediates attachment of NTHi to the respiratory epithelial cells of its host (5, 60). HMW adhesin phase variation is mediated by simple sequence repeats of seven base pairs, located within the promoter region of the adhesin gene, *hmwA.* Phase variable down-regulation of HMW adhesin production, resulting from the accumulation of SSRs, presumably allows NTHi to avoid antibody-mediated killing. We predict that the occurrence of large deletion events, even though rare, could allow for the presence and maintenance of small but stable subpopulations of adherent cells, with high HMW adhesin production, in the presence of antibody-mediated immunity. We speculate that these adherent subpopulations may be important for maintaining colonization and/or sustaining NTHi transmission between hosts (Fig. 5.6).

NTHi populations are faced with several challenges during colonization, some of these include intra- and interspecific competition with members of the pharyngeal microbiome (13, 41, 68), surviving assaults by the host immune system, and resisting host mucociliary clearance mechanisms. Following transmission, adherence of newly transmitted bacterial cells to the host epithelium is one of the first steps of the colonization process. Physical attachment to the host epithelium, mediated by surface exposed bacterial structures such as the HMW adhesins, helps bacterial cells resist mucociliary clearance mechanisms. HMW adhesins are, however, highly immunogenic and antibodies directed against these proteins are bactericidal (5). Therefore, in the absence of a mechanism to evade antibody mediated immunity, an adherent, HMW adhesin-expressing NTHi population risks being eliminated from its host by antibody mediated immunity.

Bacterial phase variation, reversible genetic changes that affect gene expression, offers a survival strategy for overcoming host immunity (7, 64) and this mechanism is the basis of the hypothesis tested herein. NTHi contains several phase variable loci, many of which are mediated by tetranucleotide repeat tracks (53). HMW adhesins, however, undergo phase variation by gaining or losing heptanucleotide repeats within the *hmwA* promoter region; stochastic changes

in repeat numbers are mediated by slipped stranded mispairing during DNA replication (15) (Fig. 5.1a). Changes in repeat number affect HMW adhesin levels in a graded fashion, as repeat numbers increase HMW adhesin protein expression decreases (15). Phase variation, therefore, generates a bacterial population that is phenotypically diverse with regards to HMW adhesin levels and this diversity potentially confers a survival advantage to the NTHi population. For example, population level HMW adhesin diversity increases the probability that at least some members of the population can survive rapid environmental changes, such as those that might occur during a transmission event. Incremental changes in *hmwA* expression, however, are difficult to track *in vivo*. Mathematical modeling therefore provides an essential tool for exploring the implications of *hmwA* phase variation during colonization.

NTHi possesses numerous SSRs, but tetranucleotide repeats tracts are the best characterized of the SSR systems. Contraction of tetranucleotide SSR tracts is favored over their expansion, a trend that may reflect selection for reduced genome size (*i.e.*, increased replication rate) (17, 38). Therefore, in the absence of selection, mutation rates alone would favor a population dominated by adherent NTHi that have relatively few SSRs and high HMW adhesin levels. The maintenance of variable length *hmwA* SSR tracts (14, 15, 26) suggests that phase variable HMW adhesin expression confers a selective advantage that outweighs the potential fitness cost associated with increased genome size. Since the length of a SSR tract affects both phase variation rate and HMW adhesin levels, SSR tract length can be tuned to maximize the fitness of a bacterial population within its specific host.

The role of HMW adhesins in mediating adherence to epithelial cells has been established *in vitro* (14, 60, 67). Since adherence is one of the first steps in colonization, we assumed that the HMW adhesins play a critical role during NTHi transmission and colonization. We therefore initiated model simulations with a small population of NTHi cells expressing high levels of HMW adhesin, that is, adherent cells with few SSRs within the *hmwA* promoter. Under this condition, inability to decrease HMW adhesin expression by phase variation resulted in the complete elimination of the adherent NTHi population by the end of the 20-day simulation (data not shown). In or model, HMW adhesin phase variation allowed the NTHi population to adapt to increasing antibody levels. As the concentration HMW adhesin-specific antibodies increased during the simulation, cells expressing high HMW adhesin levels (adherent cells) were

preferentially killed and the number of cells with low HMW adhesin levels (evasive cells) increased (Fig. 5.4). These results suggest that immune driven selective pressure has the potential to drive the NTHi population to a non-adherent state. Tauseef and Bayliss (64) recently demonstrated that phase variable expression of the outer membrane protein PorA, which also results in graded changes in protein expression, mediates *Neisseria meningitidis* escape from bactericidal antibodies *in vitro*. If HMW adhesin mediated adherence does, in fact, play a critical role in NTHi transmission, loss of the adherent phenotype could potentially drive the population to an evolutionary dead-end.

In a study of NTHi *mod* phase variation, De Bolle *et al.* (17) documented two large SSR deletion events in strains possessing relatively long tetranucleotide SSR tracts. They posited that large deletions may serve to prevent the accumulation of large SSR tracts. Within the context of *hmwA* phase variation, the simultaneous deletion of several repeats would therefore provide a mechanism by which a parent NTHi cell with a relatively long SSR tract can produce a daughter cell with a drastically reduced SSR tract (Fig. 5.1**)**. Phenotypically, this is a mechanism by which a non-adherent cell, producing low HMW adhesin levels, can generate adherent progeny, producing high levels of HMW adhesins. Importantly, this could provide a mechanism for maintaining both adherent and immune evasive populations present at all times in the host. We incorporated rare large deletion events (*i.e.*, catastrophic losses) into our model and demonstrated that they provide a mechanism capable of producing and maintaining a small subpopulation of adherent NTHi cells in spite of a strong antibody-mediated immune response.

How can we use these results to make relevant predictions regarding NTHi colonization and, potentially, pathogenesis? Sensitivity analyses identified four parameters, *Ab_max* (maximal antibody level), *s* (antibody avidity), *intercept_cl* (the rate of catastrophic loss events), and *K* (carrying capacity), that significantly impacted NTHi population dynamics and were strongly correlated with the total number of adherent NTHi cells (Fig. 5.5). The effect of *Ab_max* and *s*, varied both in time and magnitude (Fig. 5.5). Both of these parameters directly affect the immune mediated killing rate, increasing *Ab_max* increases the amount of antibody present at any given time during the simulation whereas increasing *s* increases the per unit killing rate for a given antibody level. An effective vaccine could increase both of these parameters simultaneously. In a study of infants immunized with polysaccharide conjugate vaccines, higher

affinity antibodies (which is analogous to increasing *s* in our model) were associated with higher levels of IgG (increased *Ab_max*) and, independently, with increased serum bactericidal activity (32) suggesting that these immune parameters may act synergistically.

Our model results suggest, however, that the impact of increasing *Ab_max* (maximal antibody level) on the number of adherent cells is complex. On the one hand, increasing *Ab_max* increases the rate at which adherent NTHi cells are eliminated from the host during the early days of colonization (Fig. 5.A6). On the other hand, bacterial populations subjected to increased immune pressure regained adherent cells earlier during the course of our simulation than did populations subjected to less intense immune selection (*i.e.*, lower *Ab_max*) (Fig. 5.A6). The earlier appearance of adherent cells resulted from increased selection pressure rapidly driving the population to the immune evasive phenotype from which adherent progeny were generated by catastrophic losses. This finding suggests that individuals with a more robust immune response may be more likely to harbor adherent NTHi cells during the first 20 days of colonization. This result warrants further exploration, but suggests, counter intuitively, that vaccination could potentially increase NTHi transmission.

Finally, our model demonstrated that the pharyngeal carrying capacity and the number of adherent cells are positively correlated (Fig. 5.5). In general, NTHi population size and *hmwA* phase variation rates determine the number of phase variants in a population. More specifically, this is especially important for the immune evasive population that undergoes catastrophic losses to generate adherent progeny. Assuming HMW-mediated adherence plays a role during transmission, this result could have important implications for the transmission of disease causing strains. Acute otitis media is not generally assumed to be associated with increased NTHi transmission of the disease causing strain. However, during AOM the pharyngeal NTHi community, which in healthy individuals commonly supports multiple different NTHi strains, is dominated by the disease-causing isolate present in the middle ear (10, 35, 40). Decreased NTHi strain diversity means that a larger portion of the available pharyngeal capacity could be occupied by a disease causing strain, allowing it to achieve a larger overall population size and therefore harbor increased numbers of adherent cells. This observation suggests, therefore, that disease could potentially improve a particular strain's fitness, that is, its ability to transmit between hosts.

The positive correlation between carrying capacity and adherent cells suggests that therapeutic agents or preventive measures that reduce pharyngeal carrying capacity could eliminate adherent cells and potentially reduce NTHi transmission. We explored this result in more detail by conducting simulations with carrying capacities ranging from 1E04 to 1E06. Under these conditions, there were no adherent cells remaining at day 20 (data not shown) suggesting that decreased pharyngeal carrying capacity can potentially eliminate adherent cells which may, in turn, decrease NTHi transmission.

As with any experimental system, we made several important assumptions that could impact our results. Since *hmwA* phase variation rates are not known, we assumed that the rate of heptanucleotide SSR variation was equal to that of tetranucleotide mediated phase variation described by De Bolle *et al.* (17). This, however, may not be the case and, for example, differences in the rate of DNA repair for tetranucleotide and heptanucleotide repeats could bias our rate estimates. Phase variation rates of both tetranucleotide repeats and *hmwA* heptanucleotide repeats are *recA*-independent (9, 15). Furthermore, tetranucleotide repeat variation is not affected by mutations in genes encoding components of the mismatch repair system, but is increased by inactivation of the Klenow domain of *polI* and of RnaseHI (8, 9); the effects of these mutations on *hmwA* phase variation are not known. Ultimately, differences in DNA repair mechanisms would affect phase variation rates and therefore varying the scaling factor q (Table 5.A1) can serve as a proxy for any mechanistic differences in DNA repair mechanisms. Sensitivity analyses suggests that differences in mutation rates, within the range tested in our analysis (Table 5.A1), are not correlated with the number of adherent cells (Fig. 5.4) Thus, while it is possible that there are differences between tetranucleotide and heptanucleotide repair mechanisms, we believe that any such differences would have little impact our model outcome.

We described phase variation rates on a per-repeat basis, meaning that for a given number of repeats, the absolute number of nucleotides is greater for a heptanucleotide repeat tract than a tetranucleotide tract. Since phase variation rate increases with repeat tract length, if it is the absolute tract length that dictates mutation rates, then we may have underestimated the mutation rates and the correlations that we observed would likely be conservative. We also chose to model only a subset of the potential SSR mutation events; for example, we did not incorporate

mutational events involving three repeats. While inclusion of additional event types would likely affect the overall NTHi population distribution during a model simulation, it seems unlikely that it would change our primary conclusion that catastrophic losses allow for the maintenance of a subpopulation of adherent cells. The relationship between HMW adhesin levels and antibody-mediated killing is unknown; however, we assumed that HMW levels and antibody-mediated killing rates are proportional to *hmwA* transcription levels (Fig. 5.A2). It is possible that the relationship is linear or, alternatively, that there is a strict threshold effect. As with the inclusion of additional mutational events, the exact functional form of the relationship between HMW adhesin level and killing rates would likely impact the overall population distribution during the course of a simulation but have little influence on our overall conclusion that catastrophic losses allow for the maintenance of adherent cells.

This is the first study, to the best of our knowledge, to explore the impact of a graded phase variation system on a colonizing bacterial population within the context of evolving host immunity using mathematical modeling. We parameterized our model using empirically derived mutation rates (17), allowing phase variation rates to vary by both event type and repeat tract length. The antibody-mediated immune response of a naive host was modeled in a phenomenological fashion to capture the dynamics of a primary adaptive immune response. Even using this simple model we were able to predict important roles for phase variation and adaptive immunity in shaping within-host NTHi population structure during colonization. Further studies can now build a more fine-grained model to allow for specific and mechanistic predictions about the interactions between NTHi and its human host.

**Acknowledgements**

This work was supported with funding from the Interdisciplinary Program in Infectious Diseases (TA32A1049816), the Molecular Mechanisms of Microbial Pathogenesis Training Program (AI007528), and the University of Michigan, School of Public Health, Department of Epidemiology and NIH grants R33 HL092853 and R01 EB012579- awarded to DEK. All authors report no conflicts of interest.

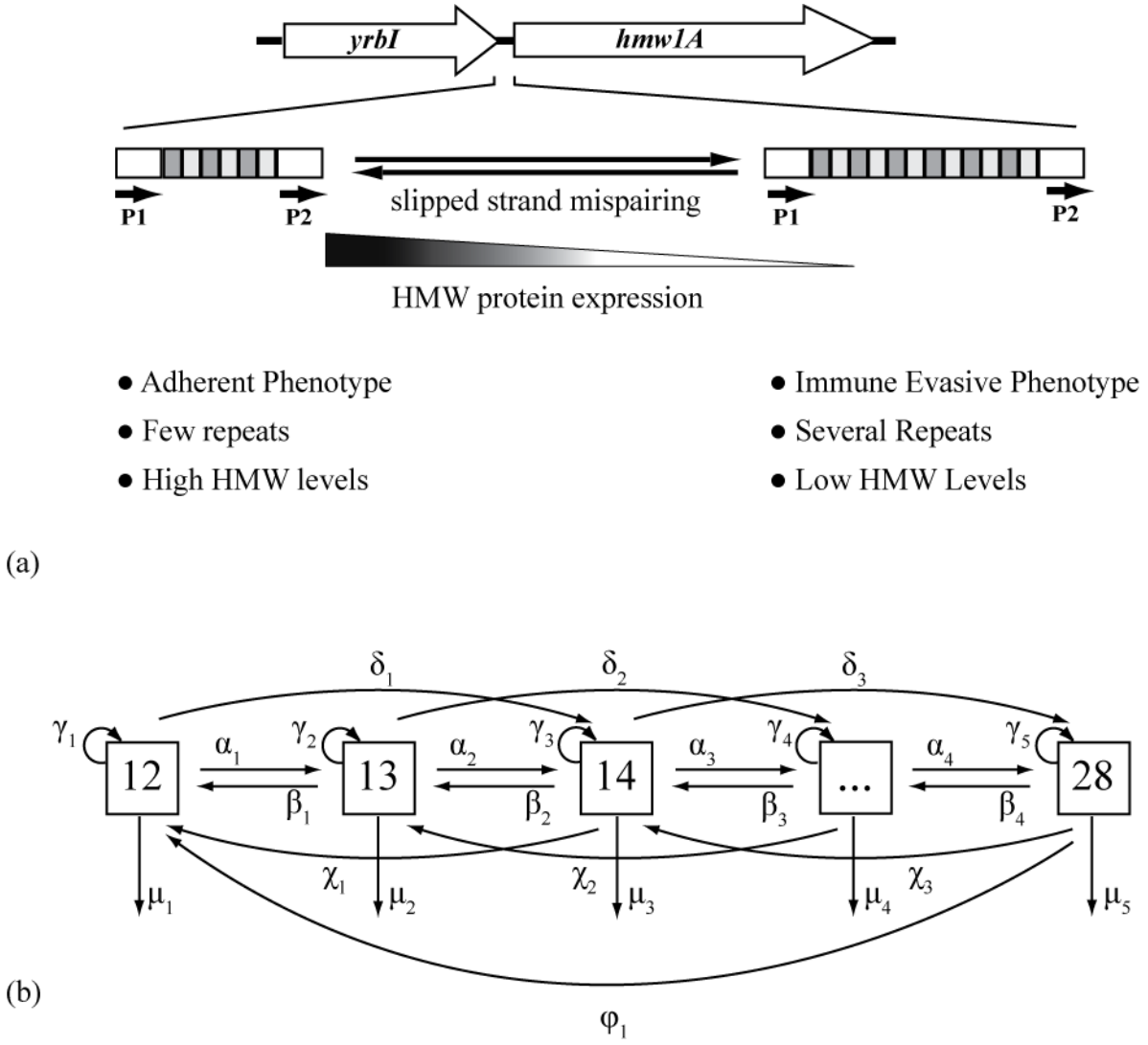



Figure 5.1. The HWM phase variation mechanism illustrated with NTHi 86028-NP *hmw1A.* (a) The *hmw1A* promoter is located within the intergenic region between *yrbI* and *hmw1A.* Phase variation is mediated by the gain and loss of heptanucleotide repeats (shaded rectangles), located within the *hmw1A* promoters (demarcated by P1 and P2), via slipped strand mispairing during DNA replication (15). Differences in repeat number affect *hmwA* transcription and HMW levels in a graded fashion such that fewer repeats are associated with increased transcription and higher HMW-levels (15). Phenotypically, cells expressing high levels of HMW can be characterized as adherent cells whereas cells expressing low HMW levels can be described as immune evasive cells. (b) Schematic representation of the within-host NTHi population model. Each compartment represents a subpopulation of NTHi cells with a specific number of repeats, ranging from 12 to 28, upstream of *hmwA.* Arrows represent the rates at which NTHi subpopulations change over time, parameter names and their baseline values are listed in Table 5.A1 of the Appendix. The system is described by 18 ordinary differential equations (Appendix).

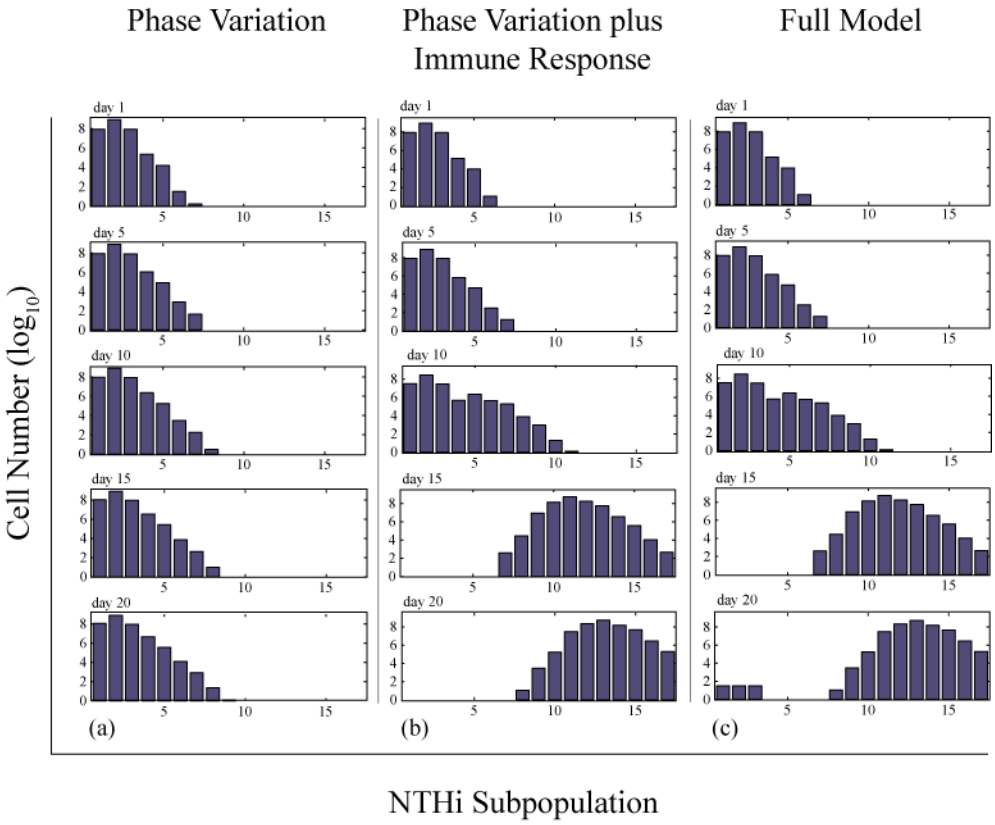



Figure 5.2. NTHi cell number for each subpopulation on days 5, 10, 15, and 20. Each bar represents the number of NTHi cells ($log_{10}$) in each of the 17 subpopulations. (a) Population distribution as a result of phase variation alone. (b) Effects of phase variation and immune selection on NTHi population. The immune response imposes a selective pressure that results in overall bacterial population distributions towards cells with increased repeat number. (c) Full model, including phase variation, immunity, and catastrophic loss events, under baseline conditions. Catastrophic loss events allow for the recovery of a small population of adherent cells by day 20.

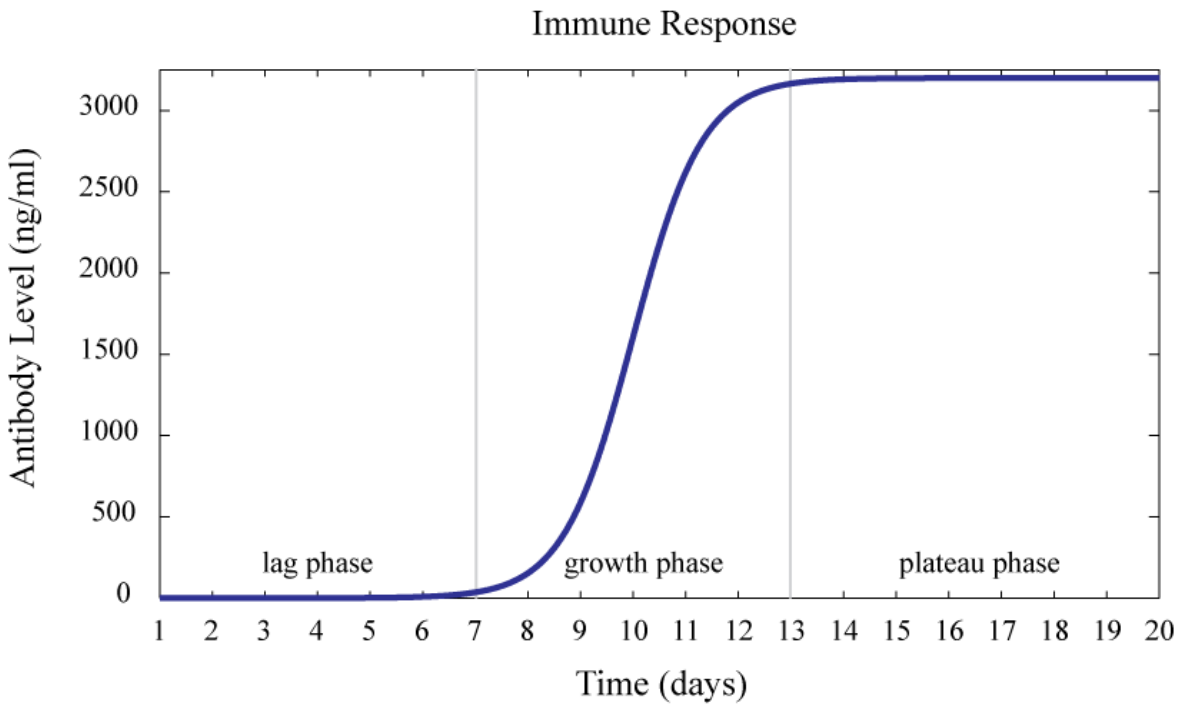



Figure 5.3. Baseline immune response over 20 days of simulation. The initial level of the immune response is 0.001 ng/ml of anti-HMW IgG antibody and the maximal level, defined as Ab_max, is 3200 ng/ml. We specify three phases of the immune response, demarcated by the grey vertical lines, as the lag phase, growth phase, and plateau phase. Between day 7 and day 13, the immune response increases from approximately 1% of *Ab_max* to 98.9% of *Ab_max*, respectively.

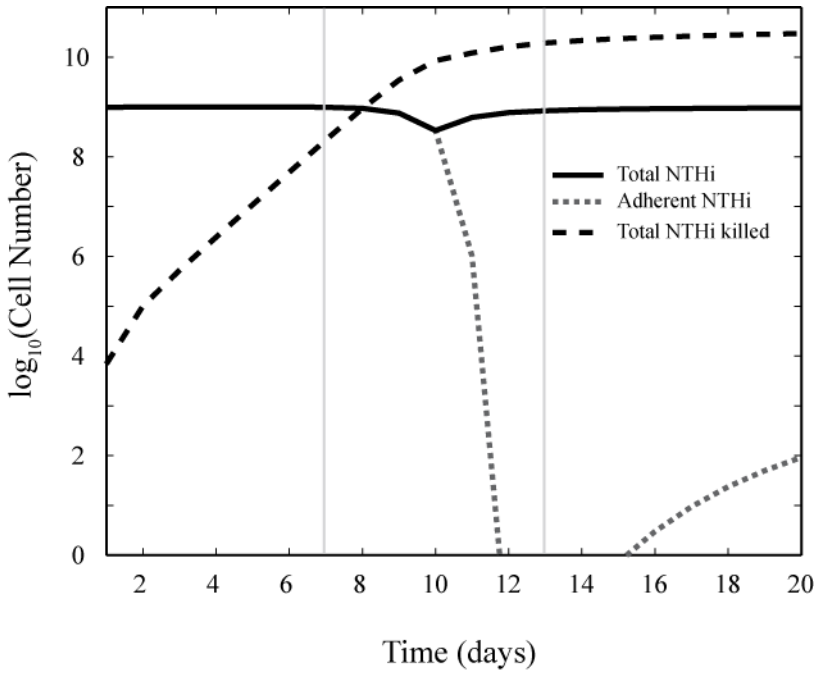



Figure 5.4. NTHi total population size the total number of NTHi killed per day under baseline conditions. The upper figure plots the total population (solid line; sum of subpopulations 1 to 17), the total number of adherent cells (dotted line; sum of subpopulations 1 to 3), and the total number of NTHi killed per day (dashed line) under baseline conditions. The bottom figure represents the baseline immune response. The effect of anti-HMW immunity becomes evident around day 9 as the total NTHi population begins to decline. While the total population is only slightly affected, the adherent population is completely eliminated from the host by day 12. However, as the immune pressure drives total bacterial population towards subpopulations with higher repeat numbers (Fig. 5.2) the adherent subpopulations become non-zero and continue to increase in number through simulation day 20. The three main phases of the immune response (Fig. 5.3) are demarcated with vertical grey lines as overlays to guide interpretation.

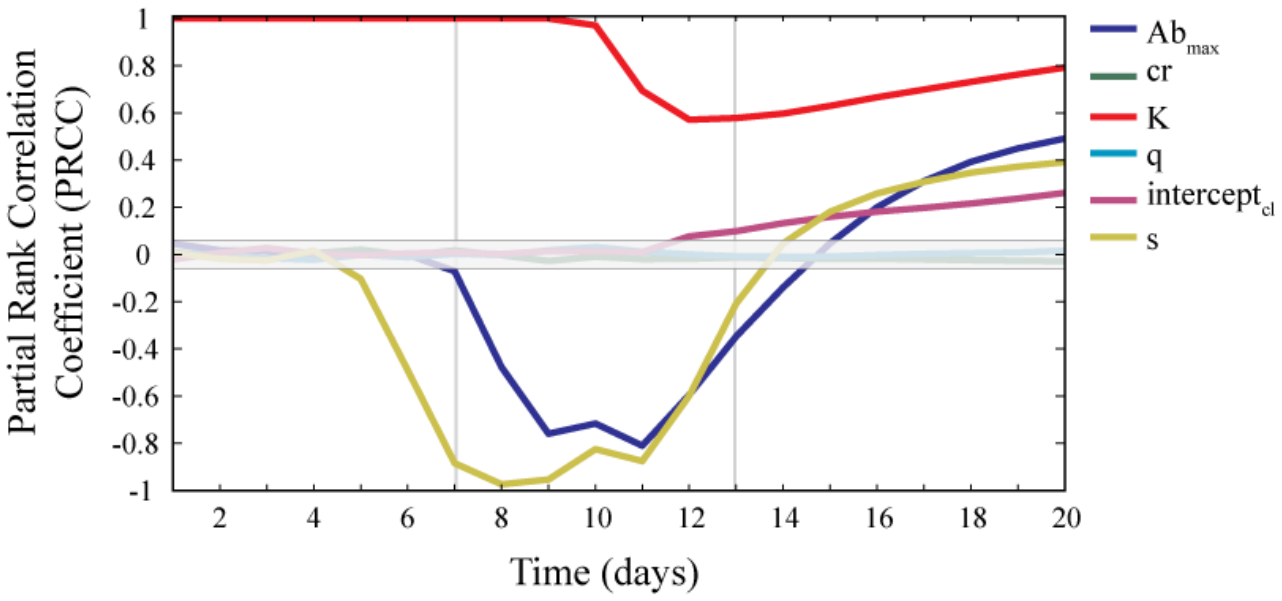



Figure 5.5. Partial rank correlation coefficients (PRCCs) of the baseline model uncertainty and sensitivity analyses plotted over time. Parameter estimates, and the ranges used for sensitivity analysis, are presented in Table 5.A1 of the Appendix. Maximum antibody levels (*Ab_max*), antibody avidity (s), rate of catastrophic losses ($intercept_{cl}$), and pharyngeal carrying capacity (*K*) are significantly correlated with the total number of adherent cells, however, the strength and direction or their effect varies over time. PRCC values outside of the central grey box, that is, PRCCs greater than approximately 0.5 and less than -0.5, are statistically significant (P-value $\leq$ 0.001) (equation 7 (42). The three main phases of the immune response (Fig. 5.3) are demarcated with vertical grey lines.

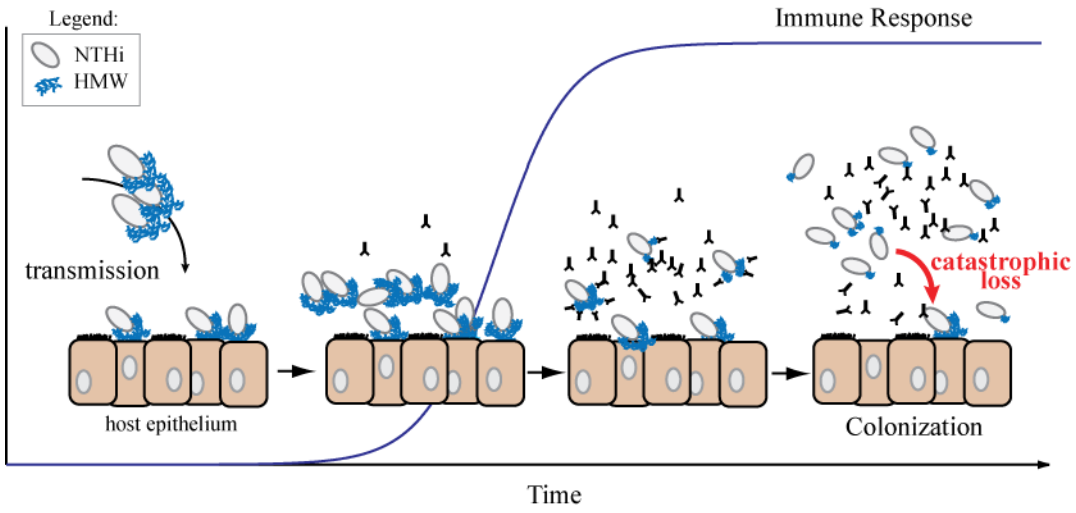



Figure 5.6. Conceptual model. NTHi is a human restricted bacterium and its long term survival is dependent upon its ability to colonize new hosts. Adherence to the host epithelium, mediated by bacterial adhesins such as HMW, is a first step in NTHi colonization. NTHi cells preferentially adhere to non-ciliated host epithelial cells and this attachment is one mechanism employed by NTHi to overcome cilia-mediated mucociliary clearance mechanisms of the host. As the NTHi cell population increases, HMWs stimulate an HMW-specific antibody mediated immune response in the host that preferentially targets NTHi cells expressing high HMW levels. Over time, the NTHi population shifts from a predominantly "adherent" phenotype to a predominantly "evasive" phenotype capable of colonizing a host despite a robust antibody mediated immune response. The simultaneous loss of several *hmwA* repeats, *i.e.*, catastrophic loss events, during DNA replication allows for the maintenance of a small yet stable population of adherent NTHi cells.

# Appendix A

## A.1. Parameter estimation.

**A.1.1. NTHi effective growth rate.** The baseline NTHi effective growth rate of 19.96 $day^{-1}$ was based upon an "effective mean generation time" of 50 minutes (46). Relative to NTHi subpopulations with few repeats, NTHi subpopulations with a high number of repeats and low HMW adhesin levels, were assumed to incur a fitness cost in the form of an increased mucociliary clearance rate. This assumption was based on two lines of reasoning. First, as demonstrated in tissue culture, cells with higher levels of HMW adhesin generally display higher levels of adherence (12, 14). This effect is a consequence of physical adherence to the epithelium. Second, studies of *H. influenzae* interactions with nasopharyngeal organ culture suggest that NTHi exert a toxin-mediate effect on ciliated cells that reduces their beat frequency and therefore would result in decreased clearance rates (1, 24). Adherent cells bind preferentially to non-ciliated cells (24), but proximity to the epithelial cell surface and ciliated cells may increase their toxin-mediated effects on ciliary beat frequency. This advantage is incorporated into the effective growth rate term which is a function of NTHi replication, natural death rates (*i.e.*, independent of antibody-mediated immunity), and clearance rates. The maximal clearance rate, $cr_{max}$, of 0.85 $day^{-1}$ was based upon estimates from a model of *Helicobacter pylori* dynamics (36). $cr_{max}$ affects the subpopulation-specific growth rate and increases linearly as repeat number increases, cells with a higher repeat number have a lower effective growth rate.

**A.1.2. Phase variation rates.** Estimated *hmwA* phase variation rates (Table 5.A1, Fig. 5.A1) were based on the on-to-off switching rate for the tetranucleotide phase variation system described by De Bolle *et al.* (17). The relationship between repeat number and phase variation rate, measured as mutations $repeat^{-1}day^{-1}$ is described by the linear function: mutation rate = (7E-6)*(repeat number) – (7E-5). To determine the rate of each event type, e.g., gain one repeat, lose one repeat, etc., we estimated the proportion of each event (from Table 2 (17)), multiplied that proportion by the repeat specific mutation rate, and transformed the rates to reflect the mutation rate $repeat^{-1}day^{-1}$ (Fig. A1**).**

**A.1.3. Relationship between HMW levels, *hmwA* 7-bp repeat number and antibody mediated killing.** HMW levels and *hmwA* transcription decrease as repeat number increases (14,

15, 26). The level of surface expressed HMW adhesin is an important parameter in our system since it directly interacts with the host immune system, both to initiate the immune response and to target NTHi cells for antibody-mediated killing. There is, however, no quantitative data relating *hmwA* transcript levels to the levels of surfaced expressed HMW adhesin, we therefore assume that the level of surface exposed HMW adhesin is proportional to *hmwA* transcript levels measured *in vitro*. The relationship between repeat number and HMW adhesin-level was estimated by fitting a curve to quantitative data for *hmwA* transcription for SSR tracts of 15 to 22 repeats (15). There were no data relating repeat number with *hmwA* transcription for cells containing 12 to 14 repeats, but 6 repeats resulted in the same amount of *hmwA* transcript as did a 15 repeats (Dawid, unpublished data). Based on this observation, we assumed that the relationship between repeat number and HMW adhesin level was the same for cells encoding 12 to 15 repeats. HMW adhesin levels for containing between 23 and 28 repeats were estimated from the curve fitted to the empirical data (Fig.5.A2).

**A.1.4. Scaling factor s.** The scaling factor, *s*, relates host antibody level to the subpopulation-specific killing rates. Uncertainty analysis was used to select the baseline parameter estimate for s = 0.00906 which was chosen so that approximately 1.0% of the total NTHi population was killed at day 5. For sensitivity analyses (Table 5.A1), we varied *s* by approximately 6.5-fold based upon studies of antibody avidity following Hib vaccination (32, 37). Our model is flexible to variations in this number.

**A.1.5. Immune response parameters $Ab_{max}$ and *r*.** The growth rate of the antibody response (Fig. 5.3) was chosen to capture the dynamics of a primary antibody response in a naive host (33). Uncertainty analysis was implement to select the baseline intrinsic growth rate of $r = 1.5$. Simulations were initiated with an immune response of 0.001. The baseline *Ab_max* value of 3200 ng IgG $ml^{-1}$ was based upon a study of serum IgG levels directed against NTHi outer membrane proteins during colonization (51). Under baseline conditions, the immune response developed gradually during the first seven days, increased rapidly between days 7 and 11, and reached its plateau at approximately day 13 (Fig. 5.3). Specifically, at day 5, 10, and 13, the immune response was approximately 0.06%, 50.6%, and 98.9% of its maximal level of 3200 ng/ml, respectively.

**A.1.6. Scaling factor q.** In order to maintain the relative relationships among mutation rates such that for a given number of repeats, $\beta_k > \alpha_i > \chi_i > \delta_m > \varphi_j$, where $k = i = m = j$, we introduce the scaling factor $q$. We multiply the intercept of the linear function describing each type of mutation rate by $q$; under baseline conditions $q = 1$. For sensitivity analyses, we vary only $q$ rather than independently varying the rate of each mutational event type. We chose to conduct analyses under two scenarios, one in which $q$ simultaneously varies all five mutation rates and a second in which $q$ affects the rates at which one or two repeats are gained/lost while the rate of catastrophic loss, $\varphi_j$, varies independently.

Table 5.A1. Parameters, baseline parameter values, and parameter ranges for sensitivity analyses.

| | **Parameters** | baseline values | sensitivity analysis (min, max) | Ref |
|---|---|---|---|---|
| α | rate at which a single repeat is gained (mutations per day per repeat) | 5.20E-04 – 1.12E-03 | | (17) |
| β | rate at which a single repeat is lost (mutations per day per repeat) | 1.12E-03 – 2.32E-03 | | (17) |
| δ | rate at which two repeats are gained in a single (mutations per day per repeat) | 1.04E-04 – 2.16E-04 | | (17) |
| χ | rate at which two repeats are lost in a single event (mutations per day per repeat) | 1.50E-04 – 2.90E-04 | | (17) |
| φ | "catastrophic" loss rate, several repeats in a single event (mutations per day per repeat) | 5.4E-05 – 5.8E-05 | | (17) |
| μ | *hmwA* repeat-specific antibody mediated killing rate ($d^{-1}$) | 0.007 – 1.0 | | (15) |
| τ | effective growth rate (generations per day) | 19.96 | | (46) |
| Ab_max | maximum ng/ml of HMW antibody (ng/ml) | 3200 | (320, 5775) | (51) |
| K | NTHi carrying capacity | $1 \times 10^9$ | ($10^5$, $10^{11}$) | |
| cr_max | rate NTHi is lost due to mucociliary clearance (per day) | 0.85 | (0.01, 8.5) | (36) |
| s | scaling factor for the relationship between repeat number and antibody mediated killing rate ; similar to antibody avidity | 0.00906 | (0.0023,0.0362) | this study |
| r | growth rate of immune response (ng per ml per day) | 1.5 | | this study |
| q | scaling factor that maintains the relationship between the rate at which repeats are gained and lost during replication | 1 | (0.01, 100) | this study |
| | **Derived Quantities** | | | |
| adherent cells | = sum of NTHi cells in subpopulations 1 - 3 | | | |
| proportion of adherent cells | = ( sum of NTHi cells in subpopulations 1 - 3)/(total number of NTHi cells in subpopulations 1 – 17) | | | |

[a] the minimum and maximum values were controlled by varying the scaling factor

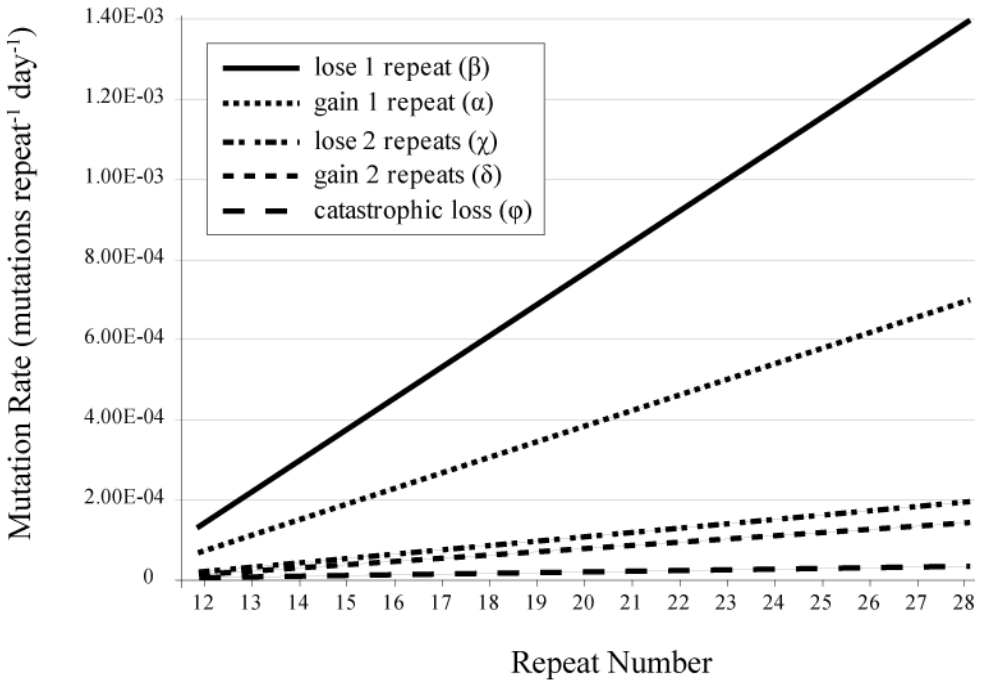



Figure 5.A1. Repeats are more likely to be lost than gained and for each event type, mutation rates increase with increasing repeat number**.** This figure illustrates the relationship between repeat number for each type of mutational event, both within each event type and between types, included in the model. Estimates were based on the tetranucleotide phase variation rates of NTHi *mod* reported by De Bolle, *et al.* (17). The linear relationships are as follows: (1) $\alpha_i$= 4E-05($REP_i$) + 4E-05, (2) $\beta_i$ = 8E-05($REP_i$) + 8E-05, (3) $\delta_i$ = 8E-06($REP_i$) + 8E-06, (4) $\chi_i$ = 1E-05($REP_i$) + 1E-05,and (5) $\varphi_i$ = 2E-06($REP_i$) + 2E-06.

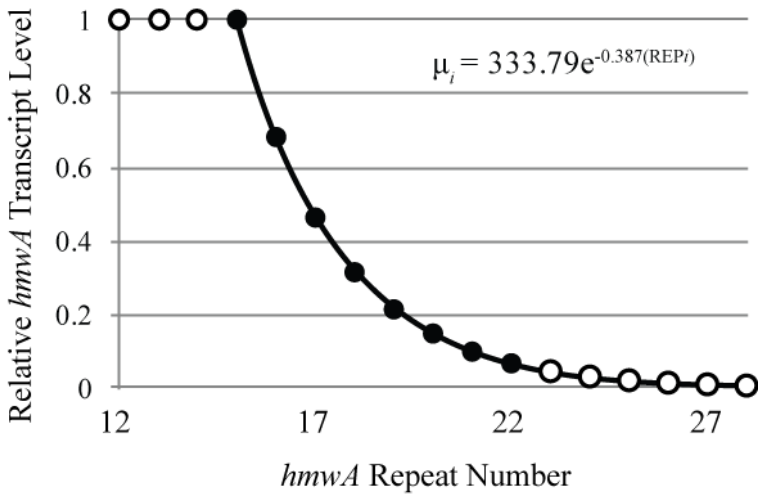



Figure 5.A2. *hmwA* transcription decreases with increasing repeat number. Closed circles (●) represent empirically determined values (15) and open circles (○) represent values estimated by fitting a curve to the empirical data. A recombinant strain encoding six repeats produced the same level of *hmwA* transcript as a strain encoding 15 repeats (Dawid, dissertation; data not shown), therefore, we assume strains encoding between 12 and 15 repeats produce the same amount of *hmwA* transcript. The relationship between *hmwA* transcript level, $\mu$, and repeat number (REP) is $\mu_i = 333.79e^{-0.387(REPi)}$.

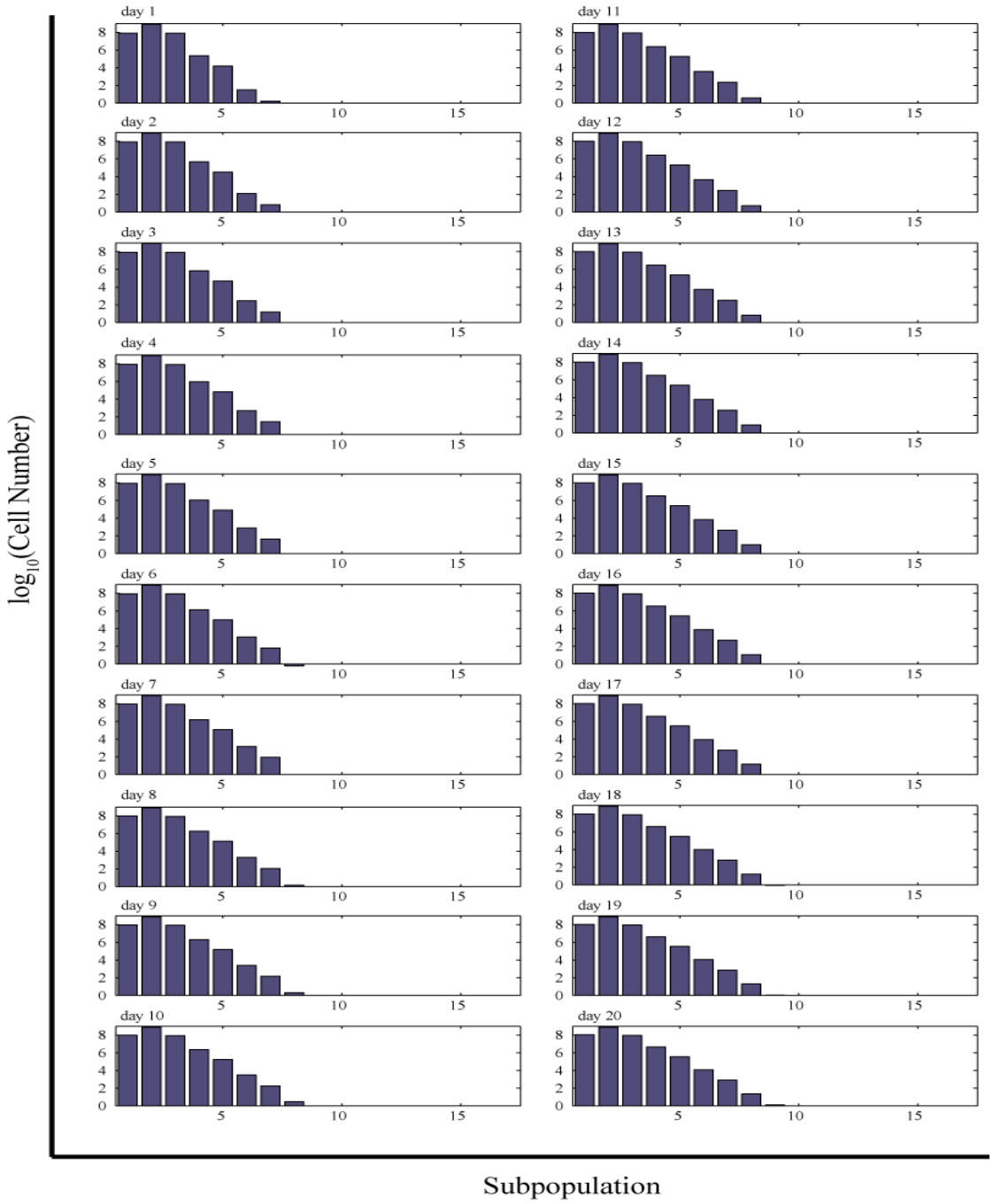



Figure 5.A3. NTHi subpopulation distributions, by day, as a function of phase variation alone. Each bar represents the number of NTHi cells in each of the 17 subpopulations defined by repeat number (Fig. 5.1b). The population distribution is solely a function of the rate at which one or two repeats are gained or lost and does not include catastrophic losses or an immune response; including catastrophic losses does not impact the population distribution (data not shown).

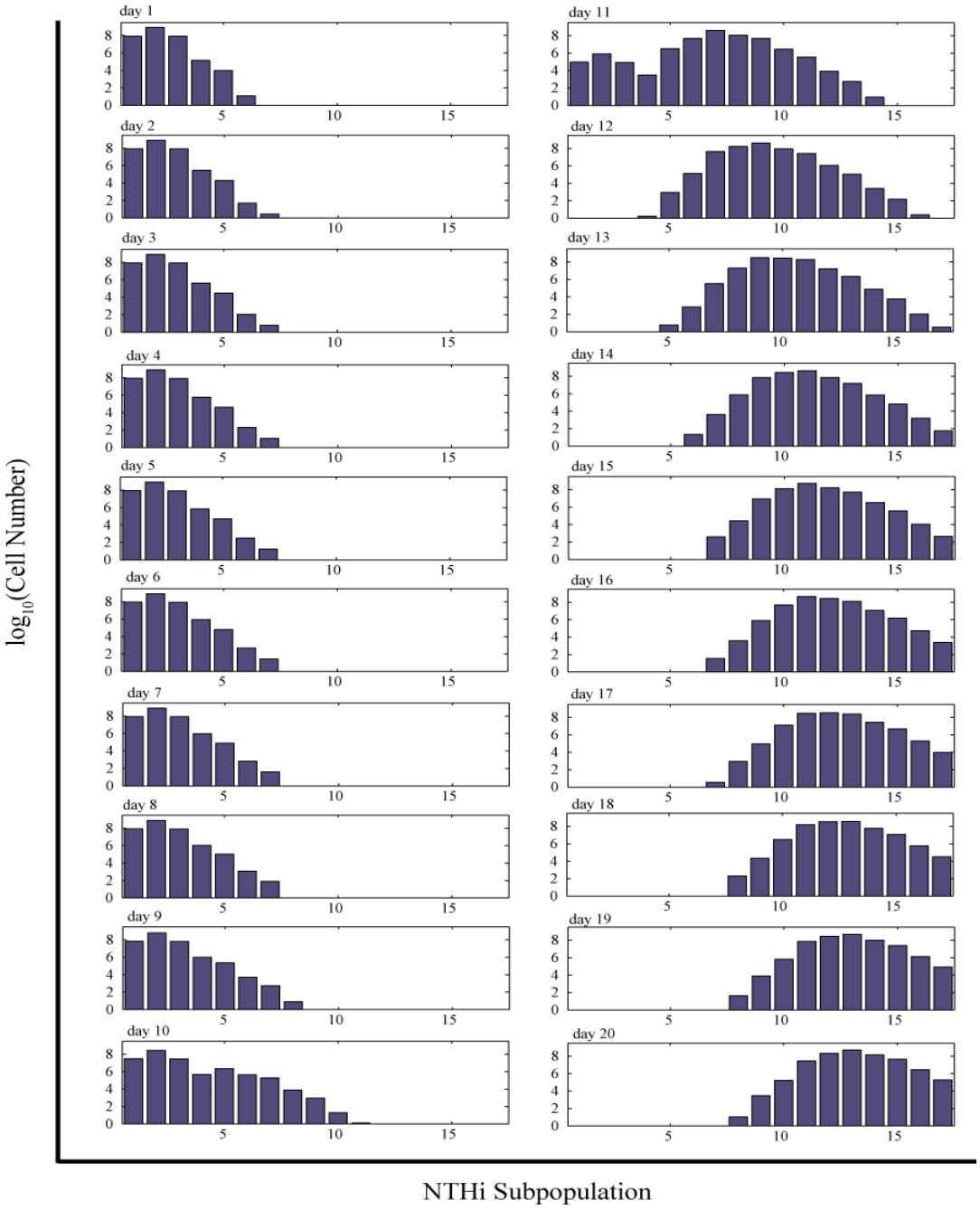



Figure 5.A4. NTHi subpopulation distributions, by day, as a function of phase variation and immunity. Each bar represents the number of NTHi cells in each of the 17 subpopulations defined by repeat number (Fig. 5.1b). This simulation includes phase variation and baseline immune response but does not allow for catastrophic losses. Relative to the simulation without the immune response, Fig. 5.2, the effects of the selective pressure applied by the immune response are apparent by day 10. By day 14 the adherent NTHi cells, subpopulations 1 to 3, have been completely eliminated.

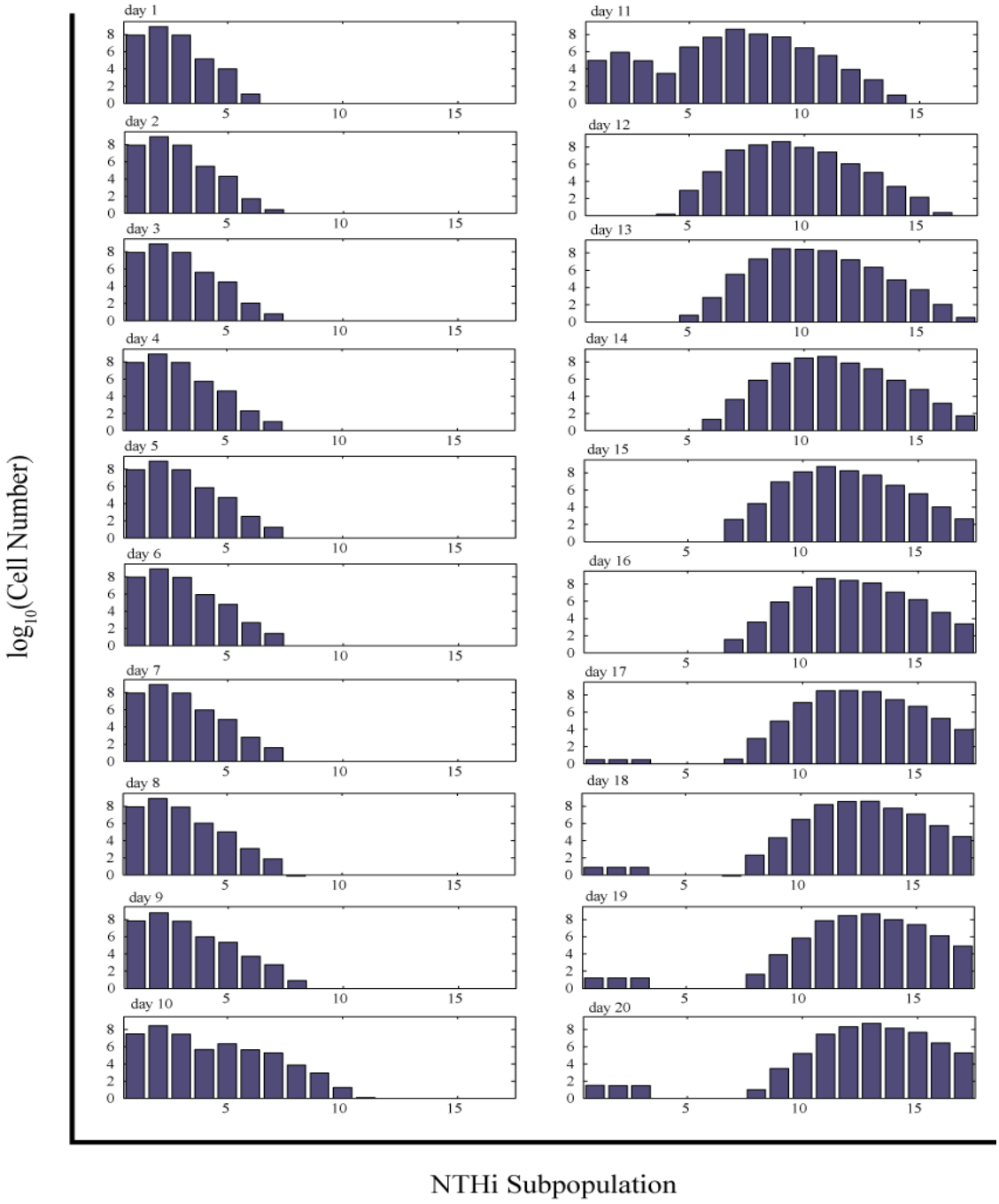



Figure 5.A5. NTHi subpopulation distributions, by day, under baseline conditions (Table 5.1 of Appendix A). Each bar represents the number of NTHi cells in each of the 17 subpopulations defined by repeat number (Fig. 5.1b). This simulation includes the baseline phase variation rates, baseline immune response and catastrophic losses. Similar to the simulation without catastrophic losses, Fig. 5.A4, the effects of the selective pressure applied by the immune response are apparent by day 10. By day 14 the adherent NTHi cells, subpopulations 1 to 3, have been completely eliminated. On day 17, however, catastrophic loss events produce a small population of adherent NTHi cells and the population gradually increases in cell number until day 20.

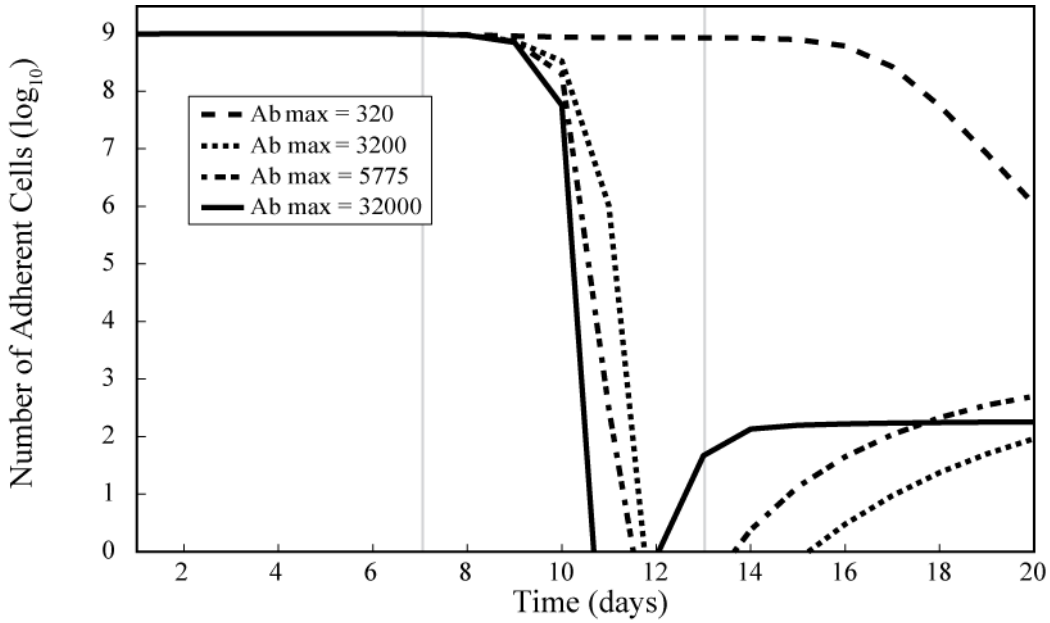



Figure 5.A6. Total numbers of adherent cells over time as a function of *Ab_max***.** For all *Ab_max* values, there is a precipitous decline in adherent cells beginning around day 10 under baseline conditions; at this time point, antibody levels are approximately 50% of *Ab_max*. Increasing *Ab_max* increases the rate at which the number of adherent cells decline with increasing immunity; however, it also decreases the total amount of time that there are a nonzero number of NTHi cells in the pharynx. The three main phases of the immune response (Fig. 5.3) are demarcated with vertical grey lines.

## Appendix B.

**Model equations.**

## Literature Cited


1. Bailey, K. L., T. D. Levan, D. A. Yanov, J. A. Pavlik, J. M. Devasure, J. H. Sisson, and T. A. Wyatt. 2012. Non-typeable Haemophilus influenzae decreases cilia beating via protein kinase C epsilon. Respir Res 13:49.
2. Barenkamp, S. J. 1996. Immunization with high-molecular-weight adhesion proteins of nontypeable *Haemophilus influenzae* modifies experimental otitis media in chinchillas. Infect Immun 64:1246-51.
3. Barenkamp, S. J. 1992. Outer membrane proteins and lipopolysaccharides of nontypeable *Haemophilus influenzae*. J Infect Dis 165 Suppl 1:S181-4.
4. Barenkamp, S. J. 1986. Protection by serum antibodies in experimental nontypable Haemophilus influenzae otitis media. Infect Immun 52:572-8.
5. Barenkamp, S. J., and F. F. Bodor. 1990. Development of serum bactericidal activity following nontypable *Haemophilus influenzae* acute otitis media. Pediatr Infect Dis J 9:333-9.
6. Barenkamp, S. J., and E. Leininger. 1992. Cloning, expression, and DNA sequence analysis of genes encoding nontypeable *Haemophilus influenzae* high-molecular-weight surface-exposed proteins related to filamentous hemagglutinin of *Bordetella pertussis*. Infect Immun 60:1302-13.
7. Bayliss, C. D., J. C. Hoe, K. Makepeace, P. Martin, D. W. Hood, and E. R. Moxon. 2008. *Neisseria meningitidis* escape from the bactericidal activity of a monoclonal antibody is mediated by phase variation of *lgtG* and enhanced by a mutator phenotype. Infect Immun 76:5038-48.
8. Bayliss, C. D., W. A. Sweetman, and E. R. Moxon. 2005. Destabilization of tetranucleotide repeats in *Haemophilus influenzae* mutants lacking RnaseHI or the Klenow domain of PolI. Nucleic Acids Res 33:400-8.
9. Bayliss, C. D., T. van de Ven, and E. R. Moxon. 2002. Mutations in polI but not *mutSLH* destabilize *Haemophilus influenzae* tetranucleotide repeats. EMBO J 21:1465-76.
10. Berrens, Z. J., C. F. Marrs, M. M. Pettigrew, S. A. Sandstedt, M. Patel, and J. R. Gilsdorf. 2007. Genetic diversity of paired middle-ear and pharyngeal nontypeable *Haemophilus influenzae* isolates from children with acute otitis media. J Clin Microbiol 45:3764-7.
11. Bou, R., A. Dominguez, D. Fontanals, I. Sanfeliu, I. Pons, J. Renau, V. Pineda, E. Lobera, C. Latorre, M. Majo, and L. Salleras. 2000. Prevalence of Haemophilus influenzae pharyngeal carriers in the school population of Catalonia. Working Group on invasive disease caused by Haemophilus influenzae. Eur J Epidemiol 16:521-6.
12. Buscher, A. Z., K. Burmeister, S. J. Barenkamp, and J. W. St Geme, 3rd. 2004. Evolutionary and functional relationships among the nontypeable *Haemophilus influenzae* HMW family of adhesins. J Bacteriol 186:4209-17.
13. Chien, Y. W., J. E. Vidal, C. G. Grijalva, C. Bozio, K. M. Edwards, J. V. Williams, M. R. Griffin, H. Verastegui, S. M. Hartinger, A. I. Gil, C. F. Lanata, and K. P. Klugman. 2013. Density Interactions between Streptococcus pneumoniae, Haemophilus influenzae and Staphylococcus aureus in the Nasopharynx of Young Peruvian Children. Pediatr Infect Dis J.
14. Cholon, D. M., D. Cutter, S. K. Richardson, S. Sethi, T. F. Murphy, D. C. Look, and J. W. St Geme, 3rd. 2008. Serial isolates of persistent *Haemophilus influenzae* in patients

with chronic obstructive pulmonary disease express diminishing quantities of the HMW1 and HMW2 adhesins. Infect Immun 76:4463-8.
15. Dawid, S., S. J. Barenkamp, and J. W. St Geme, 3rd. 1999. Variation in expression of the *Haemophilus influenzae* HMW adhesins: a prokaryotic system reminiscent of eukaryotes. Proc Natl Acad Sci U S A 96:1077-82.
16. Dawid, S., S. Grass, and J. W. St Geme, 3rd. 2001. Mapping of binding domains of nontypeable *Haemophilus influenzae* HMW1 and HMW2 adhesins. Infect Immun 69:307-14.
17. De Bolle, X., C. D. Bayliss, D. Field, T. van de Ven, N. J. Saunders, D. W. Hood, and E. R. Moxon. 2000. The length of a tetranucleotide repeat tract in *Haemophilus influenzae* determines the phase variation rate of a gene with homology to type III DNA methyltransferases. Mol Microbiol 35:211-22.
18. Dhooge, I., M. Vaneechoutte, G. Claeys, G. Verschraegen, and P. Van Cauwenberge. 2000. Turnover of *Haemophilus influenzae* isolates in otitis-prone children. Int J Pediatr Otorhinolaryngol 54:7-12.
19. Ecevit, I. Z., K. W. McCrea, C. F. Marrs, and J. R. Gilsdorf. 2005. Identification of new hmwA alleles from nontypeable *Haemophilus influenzae*. Infect Immun 73:1221-5.
20. Ecevit, I. Z., K. W. McCrea, M. M. Pettigrew, A. Sen, C. F. Marrs, and J. R. Gilsdorf. 2004. Prevalence of the *hifBC, hmw1A, hmw2A, hmwC*, and hia Genes in *Haemophilus influenzae* Isolates. J Clin Microbiol 42:3065-72.
21. Erwin, A. L., K. L. Nelson, T. Mhlanga-Mutangadura, P. J. Bonthuis, J. L. Geelhood, G. Morlin, W. C. Unrath, J. Campos, D. W. Crook, M. M. Farley, F. W. Henderson, R. F. Jacobs, K. Muhlemann, S. W. Satola, L. van Alphen, M. Golomb, and A. L. Smith. 2005. Characterization of genetic and phenotypic diversity of invasive nontypeable Haemophilus influenzae. Infect Immun 73:5853-63.
22. Erwin, A. L., S. A. Sandstedt, P. J. Bonthuis, J. L. Geelhood, K. L. Nelson, W. C. Unrath, M. A. Diggle, M. J. Theodore, C. R. Pleatman, E. A. Mothershed, C. T. Sacchi, L. W. Mayer, J. R. Gilsdorf, and A. L. Smith. 2008. Analysis of genetic relatedness of *Haemophilus influenzae* isolates by multilocus sequence typing. J Bacteriol 190:1473-83.
23. Faden, H., L. Duffy, A. Williams, D. A. Krystofik, and J. Wolf. 1995. Epidemiology of nasopharyngeal colonization with nontypeable *Haemophilus influenzae* in the first 2 years of life. J Infect Dis 172:132-5.
24. Farley, M. M., D. S. Stephens, M. H. Mulks, M. D. Cooper, J. V. Bricker, S. S. Mirra, and A. Wright. 1986. Pathogenesis of IgA1 protease-producing and -nonproducing Haemophilus influenzae in human nasopharyngeal organ cultures. J Infect Dis 154:752-9.
25. Garcha, D. S., S. J. Thurston, A. R. Patel, A. J. Mackay, J. J. Goldring, G. C. Donaldson, T. D. McHugh, and J. A. Wedzicha. 2012. Changes in prevalence and load of airway bacteria using quantitative PCR in stable and exacerbated COPD. Thorax 67:1075-80.
26. Giufre, M., A. Carattoli, R. Cardines, P. Mastrantonio, and M. Cerquetti. 2008. Variation in expression of HMW1 and HMW2 adhesins in invasive nontypeable *Haemophilus influenzae* isolates. BMC Microbiol 8:83.
27. Giufre, M., M. Muscillo, P. Spigaglia, R. Cardines, P. Mastrantonio, and M. Cerquetti. 2006. Conservation and diversity of HMW1 and HMW2 adhesin binding domains among invasive nontypeable *Haemophilus influenzae* isolates. Infect Immun 74:1161-70.

28. Gnehm, H. E., S. I. Pelton, S. Gulati, and P. A. Rice. 1985. Characterization of antigens from nontypable Haemophilus influenzae recognized by human bactericidal antibodies. Role of Haemophilus outer membrane proteins. J Clin Invest 75:1645-58.
29. Gratten, M., J. Montgomery, G. Gerega, H. Gratten, H. Siwi, A. Poli, and G. Koki. 1989. Multiple colonization of the upper respiratory tract of Papua New Guinea children with Haemophilus influenzae and Streptococcus pneumoniae. Southeast Asian J Trop Med Public Health 20:501-9.
30. Harabuchi, Y., H. Faden, N. Yamanaka, L. Duffy, J. Wolf, and D. Krystofik. 1994. Nasopharyngeal colonization with nontypeable *Haemophilus influenzae* and recurrent otitis media. Tonawanda/Williamsville Pediatrics. J Infect Dis 170:862-6.
31. Health, N. I. o. 2012. Morbidity and mortality: 2012 chart book on cardiovascular, lung, and blood diseases. *In* N. H. L. a. B. Institute. (ed.). National Institutes of Health.
32. Hetherington, S. V., and M. L. Lepow. 1992. Correlation between antibody affinity and serum bactericidal activity in infants. J Infect Dis 165:753-6.
33. Janeway, C. A., Travers, P., Walport, M., and Shlomchik, M. 2005. Immunobiology, 6th ed. Garland Science Publishing, New York.
34. Karasic, R. B., C. E. Trumpp, H. E. Gnehm, P. A. Rice, and S. I. Pelton. 1985. Modification of otitis media in chinchillas rechallenged with nontypable Haemophilus influenzae and serological response to outer membrane antigens. J Infect Dis 151:273-9.
35. Kaur, R., A. Chang, Q. Xu, J. R. Casey, and M. E. Pichichero. 2011. Phylogenetic relatedness and diversity of non-typable Haemophilus influenzae in the nasopharynx and middle ear fluid of children with acute otitis media. J Med Microbiol 60:1841-8.
36. Kirschner, D. E., and M. J. Blaser. 1995. The dynamics of Helicobacter pylori infection of the human stomach. J Theor Biol 176:281-90.
37. Lee, Y. C., D. F. Kelly, L. M. Yu, M. P. Slack, R. Booy, P. T. Heath, C. A. Siegrist, R. E. Moxon, and A. J. Pollard. 2008. Haemophilus influenzae type b vaccine failure in children is associated with inadequate production of high-quality antibody. Clin Infect Dis 46:186-92.
38. Levinson, G., and G. A. Gutman. 1987. Slipped-strand mispairing: a major mechanism for DNA sequence evolution. Mol Biol Evol 4:203-21.
39. Lindenauer, P. K., P. Pekow, S. Gao, A. S. Crawford, B. Gutierrez, and E. M. Benjamin. 2006. Quality of care for patients hospitalized for acute exacerbations of chronic obstructive pulmonary disease. Ann Intern Med 144:894-903.
40. Loos, B. G., J. M. Bernstein, D. M. Dryja, T. F. Murphy, and D. P. Dickinson. 1989. Determination of the epidemiology and transmission of nontypable Haemophilus influenzae in children with otitis media by comparison of total genomic DNA restriction fingerprints. Infect Immun 57:2751-7.
41. Margolis, E., A. Yates, and B. R. Levin. 2010. The ecology of nasal colonization of Streptococcus pneumoniae, Haemophilus influenzae and Staphylococcus aureus: the role of competition and interactions with host's immune response. BMC Microbiol 10:59.
42. Marino, S., I. B. Hogue, C. J. Ray, and D. E. Kirschner. 2008. A methodology for performing global uncertainty and sensitivity analysis in systems biology. J Theor Biol 254:178-96.
43. Mims, C. A., A. Nash, and J. Stephen. 2001. Mims' pathogenesis of infectious disease. Academic Press, San Diego, Calif. ; London.

44. Moller, L. V., A. G. Regelink, H. Grasselier, J. E. Dankert-Roelse, J. Dankert, and L. van Alphen. 1995. Multiple Haemophilus influenzae strains and strain variants coexist in the respiratory tract of patients with cystic fibrosis. J Infect Dis 172:1388-92.
45. Morel, P., C. Reverdy, B. Michel, S. D. Ehrlich, and E. Cassuto. 1998. The role of SOS and flap processing in microsatellite instability in Escherichia coli. Proc Natl Acad Sci U S A 95:10003-8.
46. Moxon, E. R. 1992. Molecular basis of invasive Haemophilus influenzae type b disease. J Infect Dis 165 Suppl 1:S77-81.
47. Murphy, T. F., S. Sethi, K. L. Klingman, A. B. Brueggemann, and G. V. Doern. 1999. Simultaneous respiratory tract colonization by multiple strains of nontypeable haemophilus influenzae in chronic obstructive pulmonary disease: implications for antibiotic therapy. J Infect Dis 180:404-9.
48. O'Brien, M. A., L. A. Prosser, J. L. Paradise, G. T. Ray, M. Kulldorff, M. Kurs-Lasky, V. L. Hinrichsen, J. Mehta, D. K. Colborn, and T. A. Lieu. 2009. New vaccines against otitis media: projected benefits and cost-effectiveness. Pediatrics 123:1452-63.
49. Organization, W. H. June 2011 2011, posting date. The 10 leading causes of death by broad income group (2008). World Health Organization. [Online.]
50. Perotin, J. M., S. Dury, F. Renois, G. Deslee, A. Wolak, V. Duval, C. De Champs, F. Lebargy, and L. Andreoletti. 2013. Detection of multiple viral and bacterial infections in acute exacerbation of chronic obstructive pulmonary disease: A pilot prospective study. J Med Virol 85:866-73.
51. Pichichero, M. E., R. Kaur, J. R. Casey, A. Sabirov, M. N. Khan, and A. Almudevar. 2010. Antibody response to Haemophilus influenzae outer membrane protein D, P6, and OMP26 after nasopharyngeal colonization and acute otitis media in children. Vaccine 28:7184-92.
52. Plasschaert, A. I., M. M. Rovers, A. G. Schilder, T. J. Verheij, and E. Hak. 2006. Trends in doctor consultations, antibiotic prescription, and specialist referrals for otitis media in children: 1995-2003. Pediatrics 117:1879-86.
53. Power, P. M., W. A. Sweetman, N. J. Gallacher, M. R. Woodhall, G. A. Kumar, E. R. Moxon, and D. W. Hood. 2009. Simple sequence repeats in *Haemophilus influenzae*. Infect Genet Evol 9:216-28.
54. Samuelson, A., A. Freijd, J. Jonasson, and A. A. Lindberg. 1995. Turnover of nonencapsulated *Haemophilus influenzae* in the nasopharynges of otitis-prone children. J Clin Microbiol 33:2027-31.
55. Sethi, S., N. Evans, B. J. Grant, and T. F. Murphy. 2002. New strains of bacteria and exacerbations of chronic obstructive pulmonary disease. N Engl J Med 347:465-71.
56. Smith-Vaughan, H. C., A. J. Leach, T. M. Shelby-James, K. Kemp, D. J. Kemp, and J. D. Mathews. 1996. Carriage of multiple ribotypes of non-encapsulated Haemophilus influenzae in aboriginal infants with otitis media. Epidemiol Infect 116:177-83.
57. St Geme, J. W., 3rd, and D. Cutter. 1995. Evidence that surface fibrils expressed by Haemophilus influenzae type b promote attachment to human epithelial cells. Mol Microbiol 15:77-85.
58. St Geme, J. W., 3rd, and D. Cutter. 1996. Influence of pili, fibrils, and capsule on in vitro adherence by Haemophilus influenzae type b. Mol Microbiol 21:21-31.

59. St Geme, J. W., 3rd, M. L. de la Morena, and S. Falkow. 1994. A *Haemophilus influenzae* IgA protease-like protein promotes intimate interaction with human epithelial cells. Mol Microbiol 14:217-33.
60. St Geme, J. W., 3rd, S. Falkow, and S. J. Barenkamp. 1993. High-molecular-weight proteins of nontypable *Haemophilus influenzae* mediate attachment to human epithelial cells. Proc Natl Acad Sci U S A 90:2875-9.
61. St Geme, J. W., 3rd, V. V. Kumar, D. Cutter, and S. J. Barenkamp. 1998. Prevalence and distribution of the *hmw* and *hia* genes and the HMW and Hia adhesins among genetically diverse strains of nontypeable *Haemophilus influenzae*. Infect Immun 66:364-8.
62. St Sauver, J., C. F. Marrs, B. Foxman, P. Somsel, R. Madera, and J. R. Gilsdorf. 2000. Risk factors for otitis media and carriage of multiple strains of *Haemophilus influenzae* and *Streptococcus pneumoniae*. Emerg Infect Dis 6:622-30.
63. Swords, W. E., B. A. Buscher, K. Ver Steeg Ii, A. Preston, W. A. Nichols, J. N. Weiser, B. W. Gibson, and M. A. Apicella. 2000. Non-typeable Haemophilus influenzae adhere to and invade human bronchial epithelial cells via an interaction of lipooligosaccharide with the PAF receptor. Mol Microbiol 37:13-27.
64. Tauseef, I., and C. D. Bayliss. 2013. Phase variation of PorA, a major outer membrane protein, mediates escape of bactericidal antibodies by Neisseria meningitidis. Infect Immun.
65. Teele, D. W., J. O. Klein, and B. Rosner. 1989. Epidemiology of otitis media during the first seven years of life in children in greater Boston: a prospective, cohort study. J Infect Dis 160:83-94.
66. Trottier, S., K. Stenberg, and C. Svanborg-Eden. 1989. Turnover of nontypable *Haemophilus influenzae* in the nasopharynges of healthy children. J Clin Microbiol 27:2175-9.
67. van Schilfgaarde, M., P. van Ulsen, P. Eijk, M. Brand, M. Stam, J. Kouame, L. van Alphen, and J. Dankert. 2000. Characterization of adherence of nontypeable *Haemophilus influenzae* to human epithelial cells. Infect Immun 68:4658-65.
68. Verhaegh, S. J., M. L. Snippe, F. Levy, H. A. Verbrugh, V. W. Jaddoe, A. Hofman, H. A. Moll, A. van Belkum, and J. P. Hays. 2011. Colonization of healthy children by Moraxella catarrhalis is characterized by genotype heterogeneity, virulence gene diversity and co-colonization with Haemophilus influenzae. Microbiology 157:169-78.

# Chapter 6

## Summary and Future Directions

### Summary

The work presented in this dissertation sought to: (1) develop a molecular typing method to differentiate typeable from true nontypeable *H. influenzae,* (2) characterize a geographically distributed collection of NTHi isolates with respect to *hmwA* prevalence and sequence diversity, (3) test for evidence of selective pressures operating on *hmwA,* and (4) develop a mathematical model that provides a theoretical framework for exploring how *hmwA* phase variation and host mediated immunity interact to shape NTHi within host population structure.

The major findings of this work are:

- The presence of *bexB,* as determined by PCR or DNA microarray, is a reliable marker of the capsule locus allowing for rapid and reliable differentiation between typeable and nontypeable *H. influenzae*.

- Among our collection of 170 NTHi commensal (n = 75) and OM (n = 95) strains collected between 1994 – 2002 from children seven years of age or under living in Finland, Israel, or the United States:
    - 72 % were *hmwA* positive
    - *hmwA*-positive strains were distributed throughout the MLST phylogeny
    - 93 % of the *hmwA*-positive strains tested contained two *hmw* loci
    - *hmwA* was 1.23 times more prevalent among OM strains than commensal strains

- Phylogenetic analysis of *hmwA* binding domain regions of 33 NTHi strains (16 commensal and 17 OM strains) from this collection revealed four distinct sequence clusters.
    - 94 % of *hmwA* positive strains possessed a Cluster 1 *hmwA*
    - 84 % of the *hmwA* sequences belonged to either Cluster 1 or 2
    - sequences did not cluster by locus, geographic location, or disease

- HMW adhesin molecular evolution is driven by positive (*e.g.*, diversifying) selection.
    - the majority of positively selected amino acids are localized to the *hmwA* binding domain region
    - the majority of positively selected amino acids are located in structural regions predicted to contain hydrophilic loops and turns

- Mathematical modeling of HMW adhesin phase variation demonstrated:
    - in the absence of antibody mediated immunity,
        - phase variation generates a nearly uniform within-host population distribution
        - including large deletion events has little effect on the population distribution
    - in the presence of antibody mediated immunity,
        - the within-host population is highly skewed and consists entirely of cells with low HMW-levels (immune evasive cells)
        - inclusion of large deletion events (catastrophic losses) allows for the maintenance of a small population of adherent cells that may be important for sustaining colonization and/or maintaining transmission

- The HMW adhesin phase variation model results suggest that antibody levels, antibody avidity, catastrophic loss rates, and population carrying capacity significantly affected numbers of adherent NTHi cells within a host.

## Future Directions

The work presented in Chapters 3, 4, and 5 addressed three different aspects of NTHi HMW adhesins. Each chapter marks only the beginning of a journey that, I believe, has provided new and interesting insights regarding the evolution of HMW adhesins, the role of HMW adhesin phase variation during colonization, and potential implications of phase variation for NTHi transmission. Below I outline several future directions aimed at advancing our understanding of HMW adhesins and their role in colonization and disease. I have made an effort to offer suggestions that maintain the interdisciplinary spirit of this work and thus include projects that range from the bench top, to the population-level, and to the theoretical.

The *hmwA* prevalence and sequencing project presented in Chapter 3 can take a number of future directions. The first thing that must be done is to verify *hmwA* prevalence and copy number results with a probe based approach to confirm the PCR based approach used in the data presented. In a small number of strains, I experienced considerable difficulty amplifying one or both *hmwA* loci; I attribute the difficulty to sequence diversity within the *hmwA* locus specific PCR primer annealing regions. Probe based methods are less sensitive to nucleotide diversity than are PCR-based methods since probes generally anneal to a much larger region of the target gene than do PCR primers. The *hmwA*-leader PCR primers used to screen our strain collection can be used to generate an *hmwA* probe that will hybridize to the signal sequence region of both $hmwA_{1679}$ and $hmwA_{1598}$. I suggest a two-step approach to screening the strain collection. First, genomic DNA of each strain will be interrogated in a dot blot assay using the *hmwA* probe; this will identify *hmwA*-positive strains but yields no information on the number of loci present. Next, the *hmwA*-positive strains identified with the dot blot, will be interrogated with a Southern blot analysis which will determine the number of *hmwA* loci per strain. Unfortunately, determining whether the loci are in conserved chromosomal locations, as previous studies suggest, would require a PCR-based approach.

Previous studies suggest that the HMW adhesins are associated with various other NTHi characteristics, such as, the absence of the *hia* adhesin and certain biochemical traits. Using the current strain collection, these would be interesting avenues to explore in further understanding the relationships among NTHi microbial factors. The *hia* adhesin is a homolog of the *H. influenzae* type b *hsf* adhesin (1). With few exceptions, NTHi strains encode either two

HMW adhesins or an Hia adhesin (1, 7, 8, 14). Similar to *hmw*, *hia* displays marked nucleotide diversity and thus would likely prove challenging to screen with a PCR-based approach. Thus a probe based approach, as described above, would be preferable. There are also several biochemical characteristics associated with strains that possess *hmw*. For example, *hmwA*-containing strains are typically urease-positive but tend to show negative test results for both ornithine decarboxylase and lysine decarboxylase activity (7). The urease operon, like the *hmw* loci, are more prevalent among disease (AOM and COPD) isolates than throat isolates (17). These biochemical properties can easily be screened with standard biochemical assays. Combining biochemical data with *hmwA* prevalence data will deepen our understanding of potential NTHi virulence associated phenotypes, and, more importantly will more precisely define the role, or lack thereof, of HMW adhesins in AOM pathogenesis. For example, an alternative hypothesis to HMW adhesins being important for pathogenesis is that urease activity is important for AOM pathogenesis and *hmwA* prevalence among disease isolates reflects linkage between *hmwA* and the urease operon, not *hmwA*'s contribution to pathogenesis. This might explain why HMW-positive strains are more prevalent among OM isolates but HMW protein production is reduced in OM strains relative to commensal strains. This hypothesis could be tested using a population-level epidemiologic approach and/or an experimental approach using an animal model of otitis media pathogenesis. The epidemiologic approach would rely upon statistical analyses to determine if there are significant associations between phenotype, *e.g.*, urease activity, and disease; epidemiologic studies could be conducted with phenotypic data, genotypic data, or a combination of both. Experimentally, this hypothesis could also be tested by comparing a wild-type urease-HMW-positive strain to isogenic single *ure* and *hmwA* mutants and a *ure-hmwA*-double mutant in an animal model (*e.g.*, chinchilla) of otitis media. Ideally, both approaches would be employed.

HMW adhesins are phase variable and previous studies have demonstrated that strains isolated from ears of children with acute otitis media encode a higher number of SSRs than do isogenic strains collected from the throat (5). Increases in repeat number have also been documented in serial isolates collected over time from the sputum of individuals suffering from COPD (2). The strain collection used in this study affords an opportunity to determine if, on average, OM isolates possess a higher number of *hmwA* SSRs than commensal isolates. Thus, another interesting follow-up study to Chapter 2 would be to determine the number of repeats

located within *hmwA* promoter region of each locus. These data, especially those collected from commensal isolates, could be used in combination to test the predictions of the mathematical model presented in Chapter 5. There are, however, some limitations to using this strain collection to compare SSR numbers in the throat versus the ear. First, the strains in this collection are not matched isolates; that is, they were not collected at the same time from the same individual. Comparisons between OM versus commensal isolates may, therefore, be confounded if the relationship between repeat number and HMW production varies from strain-to-strain depending on the total genetic content of a specific strain. Secondly, it is possible that sample collection methods could influence such a comparison, as OM isolates are cultured from fluid collected from the middle ear space whereas commensal isolates are collected from the throat using a swab. For example, fluid collected from the middle ear may be enriched for non-adherent NTHi (*i.e.*, those with a large number of SSRs and low HMW adhesin production) whereas collecting samples with a swab may increase the probability of collecting adherent cells (*i.e.*, those with fewer SSRs and higher HMW adhesin production).

The presence of *hmwA* only tells part of the story in NTHi pathogenesis. A more biologically relevant question is whether *hmwA*-positive strains actually produce functional HMW adhesins. To make the link between the presence of *hmwA* and production of HMW adhesins, Western blotting could be performed on the *hmwA*-positive strains. There are several methods by which antibodies targeting HMW could be acquired. First, pooled adult human serum would be a good source of antibodies for such an experiment, but pooled serum would contain antibodies directed against multiple different NTHi antigens. It would be necessary therefore to reduce the non-HMW antibodies by first absorbing the sera against NTHi strain 11, which does not encode HMW adhesins, or possibly against a genetically engineered NTHi strain 12 lacking both HMW adhesins. An alternative strategy would be to generate monoclonal antibodies against a conserved region of the HMW adhesins, for example, antibodies directed against epitopes localized to the HMW adhesin stalk region (Fig 4.1b). Quantitative data on HMW adhesin production would be especially interesting when considered in conjunction with data on *hmwA* repeat number; previous studies have demonstrated an inverse relationship between repeat number and protein production (4, 5, 9). Furthermore, comparing quantitative data could improve our understanding of role of HMW adhesins in colonization and disease. For example, it would be possible to determine if OM strains produce more or less HMW adhesin

than commensal strains. Finally, quantitative data would overcome some of the limitations inherent in simply relying on SSR number as a proxy for HMW protein production when making comparisons between strains as it should control for potential differences in strain-specific relationships between repeat number and protein production.

Phylogenetic analysis of the *hmwA* core binding domain sequences identified four distinct sequence clusters. Previous studies have defined two different HMW adherence profiles, HMW1-like and HMW2-like. The number of NTHi strains for which adherence profiles have been reported is limited; in fact, adherence profiles are only available for three of the NTHi strains (2, 12, 13) included in the analyses of Chapter 3 and 4, combined. van Schilfgarrde *et al.*, tested several NTHi strains against two specific tissue culture cell lines (Chang and NCI-H292), with or without dextran sulfate, and defined four distinct adherence patterns among 22 *hmw*-positive strains (16); 17/22 exhibited either HMW1-like or HMW2-like adherence. Based on the phylogenetic analyses presented in Chapters 2 and 3, all three of the strains encode one HMW adhesin belonging to Cluster 1 and a second belonging to Cluster 2 (Figure 3.2 and 4.2). Thus, it would be interesting to determine the adherence profiles of representative strains from *hmwA* sequence Clusters 3 and 4. Since current approaches to defining *in vitro* NTHi adherence profiles rely on a limited number of tissue cell lines, using a more sensitive technique, such as glycan microarrays (3, 10) or a larger sample of relevant human cells could provide more discriminatory power for defining HMW-specific adherence characteristics.

We observed a small number of strains in this collection with nearly identical *hmwA* sequences at both chromosomal loci, and this pattern was consistent with a recent gene conversion event. It would be interesting to actually test for gene conversion events, and estimate their frequency using a flux analysis similar to that employed in a recent study of *H. pylori* gene conversion (15). If gene conversion events are documented, then a next step would be to determine if the donor DNA came from within the cell or was taken up from the environment. This can be determined using bacterial cells deficient in DNA uptake mechanisms.

Several areas of HMW phase variation can be explored with the mathematical model presented in Chapter 5. HMW adhesins are usually present in two copies per chromosome. In the current model, however, only a single adhesin is considered. A logical extension of this model would be to include a second adhesin. The impact of a second adhesin would likely vary as a

function of several factors. The effect of including a second adhesin on the within host population structure would likely be most impacted by the degree of amino acid similarity between the two adhesins. The *hmwA* sequencing component of this work highlighted the degree of variation between *hmwA* loci within a cell, ranging from nearly identical to only 54% amino acid identity. For nearly identical HMW adhesins, the bacteria may derive little benefit by having a second adhesin since the primary driver of within host dynamics is host immunity. There may, however, be an advantage in terms of increased adherence for a cell expressing two nearly identical adhesins. In contrast, cells that display genetically diverse adhesins may experience very different within host dynamics driven by two factors: the extent to which expression of the adhesins is coordinated and the degree of cross-reactivity between the two *hmwA* loci. There is no direct evidence of coordinated regulation of HMW expression, although to the best of my knowledge this has not been formally tested, nor is there reason to suspect any form of coordinated phase variation at the two *hmwA* loci. The mechanism of SSR variation, slipped stranded mispairing, is a stochastic event that occurs during DNA synthesis and the numbers of tandem repeats at each locus within an individual NTHi isolate vary. Thus, if little antigenic cross-reactivity exists among HMWs and the *hmw* loci vary independently of one another, then incorporation of a second locus would be expected to have little impact on within-host population dynamics. On the other hand, if little cross protection exists and only a single adhesin is expressed at one time, then having two loci could extend the HMW-mediated adhesive activity and colonization potential of the NTHi population. This could impact both the number of adherent cells present at any given time and the duration of time that adherent cells are present. Both of these could have important implications for NTHi transmission.

Another focus of model modification could be the immune response. We observed that increasing the strength of the immune response (varying $Ab_{max}$) had an unexpected effect on NTHi within-host dynamics. Specifically, as $Ab_{max}$ increased the amount of time that adherent cells were maintained in the population was also increased. This could have important implications for assessing vaccine effectiveness. Thus, another logical extension of the model is to represent the immune response in the model in a more mechanistic way. This could be informed, for example, by the expectations of previous exposure to a specific HMW or to effective vaccination against a specific HMW. In the simplest case, this can be achieved by varying the rate and intensity of the immune response, for example, a secondary response might

be expected to be faster and to reach a higher titer ($Ab_{max}$) than a primary response as currently implemented.

In the current model, only the first 20 days of a colonization event are simulated. At the end of the simulation, the majority of the NTHi population is "immune evasive" and there is a small but stable population of adherent cells. Incorporating more complex immune dynamics, for instance, including waning of the primary response to the HMW adhesin may illuminate more complex within-host population dynamics. For example, as the antibody concentration begins to decrease, the number of adherent NTHi cells maintained within a host may increase; maintenance of a larger population of adherent cells would have implications for NTHi transmission.

Sensitivity analyses of the *hmwA* repeat region mutation rates demonstrated that the critical parameter determining within-host population structure is the rate of catastrophic loss events, not the overall mutation rates (*e.g.,* the gain/loss of one or two repeats). These results suggest that spending a great deal of time, and money, gathering precise estimates of *hmwA* SSR mutation rates with an *in vitro* experimental system may not be an efficient use of resources. This conclusion, however, is predicated on the assumption, which may not be true, that *hmwA* SSR, which are located within the promoter, are under the same constraints as the *mod* tetranucleotide SSRs, which are located within the coding frame. I would argue that characterizing the mutation rates is important, not necessarily for the impact it will have on the within host population dynamics, but for gaining a basic understanding of the mechanisms that regulate the mutability of SSRs located within the *hmwA* promoter. For example, specifically interrogating *hmwA* SSRs could verify whether or not *hmwA* SSR mutation rates increase with tract length as demonstrated with *mod* (6) and as assumed for our model parameterization. Additionally, an *in vitro* system could be used to verify the catastrophic loss mechanism that we proposed in our model. This would, however, be a daunting task give the potential rarity of such events, making our theoretical approach all the more necessary in generating hypotheses about this complex system.

In conclusion, I would like to finish by addressing the question that initially piqued my interest in the questions addressed in this work: Are the HMW adhesins good vaccine candidates? On the one hand, the HMW adhesins are surface exposed, highly immunogenic, and

more prevalent among AOM-causing strains than commensal strains, arguing for their suitability as vaccine antigens. The sequence diversity and the fact that the adhesins are phase variable, while posing challenges for vaccine development, do not necessarily preclude the possibility that they could be effective vaccine targets. For example, if AOM-causing strains formed phylogenetic clusters distinct from commensal isolates it might be possible to target specific HMW alleles, potentially reducing the complexity of the problem. The same can be said for HMW adhesin phase variation: if HMW adhesin production were required for virulence and/or transmission, an effective vaccine could potentially prevent disease given colonization or reduce transmission, respectively. Chapters 3, 4, and 5 address topics that are relevant to assessing the utility of HMW adhesins a vaccine targets. The fact that *hmwA* sequences in this study did not cluster by disease suggests that, as suspected, HMW adhesin amino acid diversity would present a significant challenge. Given the immunogenicity of HMW adhesins, the amino acids predicted to be subjected to positive selection are likely important for immune evasion. This information could potentially be used to inform antigen selection; for example, specifically avoiding the regions that are subject to positive selection might increase the probability of identifying antigens that stimulate a more broadly protective immune response. An alternative strategy would be to target the more conserved carboxyl terminus of the HMW adhesins; this is somewhat analogous to the approach being implemented in the pursuit of a universal influenza vaccine (11). The mathematical modeling results, however, suggest that a strong immune response directed against HMW, as might result from exposure to an antigen following successful vaccination, actually could lead to adherent cells appearing more rapidly during colonization. Assuming the adherence phenotype is important for transmission, this result suggests that vaccination may potentially increase NTHi transmission and this effect would be independent of the region of the mature HMW adhesin targeted, the core binding region or the conserved stalk region. Taken together, these findings argue against HMW adhesins as viable vaccine targets. I reach this conclusion, not because of the results presented in any one chapter, but instead by synthesizing all three chapters which I believe highlights the value of the interdisciplinary nature of this work.

## Literature Cited


1. Barenkamp, S. J., and J. W. St Geme, 3rd. 1996. Identification of a second family of high-molecular-weight adhesion proteins expressed by non-typable *Haemophilus influenzae*. Mol Microbiol 19:1215-23.
2. Buscher, A. Z., K. Burmeister, S. J. Barenkamp, and J. W. St Geme, 3rd. 2004. Evolutionary and functional relationships among the nontypeable *Haemophilus influenzae* HMW family of adhesins. J Bacteriol 186:4209-17.
3. Chen, L. M., P. Rivailler, J. Hossain, P. Carney, A. Balish, I. Perry, C. T. Davis, R. Garten, B. Shu, X. Xu, A. Klimov, J. C. Paulson, N. J. Cox, S. Swenson, J. Stevens, A. Vincent, M. Gramer, and R. O. Donis. 2011. Receptor specificity of subtype H1 influenza A viruses isolated from swine and humans in the United States. Virology 412:401-10.
4. Cholon, D. M., D. Cutter, S. K. Richardson, S. Sethi, T. F. Murphy, D. C. Look, and J. W. St Geme, 3rd. 2008. Serial isolates of persistent *Haemophilus influenzae* in patients with chronic obstructive pulmonary disease express diminishing quantities of the HMW1 and HMW2 adhesins. Infect Immun 76:4463-8.
5. Dawid, S., S. J. Barenkamp, and J. W. St Geme, 3rd. 1999. Variation in expression of the *Haemophilus influenzae* HMW adhesins: a prokaryotic system reminiscent of eukaryotes. Proc Natl Acad Sci U S A 96:1077-82.
6. De Bolle, X., C. D. Bayliss, D. Field, T. van de Ven, N. J. Saunders, D. W. Hood, and E. R. Moxon. 2000. The length of a tetranucleotide repeat tract in *Haemophilus influenzae* determines the phase variation rate of a gene with homology to type III DNA methyltransferases. Mol Microbiol 35:211-22.
7. Erwin, A. L., K. L. Nelson, T. Mhlanga-Mutangadura, P. J. Bonthuis, J. L. Geelhood, G. Morlin, W. C. Unrath, J. Campos, D. W. Crook, M. M. Farley, F. W. Henderson, R. F. Jacobs, K. Muhlemann, S. W. Satola, L. van Alphen, M. Golomb, and A. L. Smith. 2005. Characterization of genetic and phenotypic diversity of invasive nontypeable Haemophilus influenzae. Infect Immun 73:5853-63.
8. Erwin, A. L., S. A. Sandstedt, P. J. Bonthuis, J. L. Geelhood, K. L. Nelson, W. C. Unrath, M. A. Diggle, M. J. Theodore, C. R. Pleatman, E. A. Mothershed, C. T. Sacchi, L. W. Mayer, J. R. Gilsdorf, and A. L. Smith. 2008. Analysis of genetic relatedness of *Haemophilus influenzae* isolates by multilocus sequence typing. J Bacteriol 190:1473-83.
9. Giufre, M., M. Muscillo, P. Spigaglia, R. Cardines, P. Mastrantonio, and M. Cerquetti. 2006. Conservation and diversity of HMW1 and HMW2 adhesin binding domains among invasive nontypeable *Haemophilus influenzae* isolates. Infect Immun 74:1161-70.
10. Heimburg-Molinaro, J., M. Tappert, X. Song, Y. Lasanajak, G. Air, D. F. Smith, and R. D. Cummings. 2012. Probing virus-glycan interactions using glycan microarrays. Methods Mol Biol 808:251-67.
11. Pica, N., and P. Palese. 2013. Toward a universal influenza virus vaccine: prospects and challenges. Annu Rev Med 64:189-202.
12. St Geme, J. W., 3rd. 1994. The HMW1 adhesin of nontypeable *Haemophilus influenzae* recognizes sialylated glycoprotein receptors on cultured human epithelial cells. Infect Immun 62:3881-9.
13. St Geme, J. W., 3rd, S. Falkow, and S. J. Barenkamp. 1993. High-molecular-weight proteins of nontypable *Haemophilus influenzae* mediate attachment to human epithelial cells. Proc Natl Acad Sci U S A 90:2875-9.

14. St Geme, J. W., 3rd, V. V. Kumar, D. Cutter, and S. J. Barenkamp. 1998. Prevalence and distribution of the *hmw* and *hia* genes and the HMW and Hia adhesins among genetically diverse strains of nontypeable *Haemophilus influenzae*. Infect Immun 66:364-8.
15. Talarico, S., S. E. Whitefield, J. Fero, R. Haas, and N. R. Salama. 2012. Regulation of Helicobacter pylori adherence by gene conversion. Mol Microbiol 84:1050-61.
16. van Schilfgaarde, M., P. van Ulsen, P. Eijk, M. Brand, M. Stam, J. Kouame, L. van Alphen, and J. Dankert. 2000. Characterization of adherence of nontypeable *Haemophilus influenzae* to human epithelial cells. Infect Immun 68:4658-65.
17. Zhang, L., M. Patel, J. Xie, G. S. Davis, C. F. Marrs, and J. R. Gilsdorf. 2013. Urease Operon and Urease Activity in Commensal and Disease-Causing Nontypeable Haemophilus influenzae. J Clin Microbiol.