HOST-MICROBE EVOLUTIONARY CONFLICT: INVESTIGATING THE HOST SPECIFICITY OF *HELICOBACTER PYLORI* ADHESIN HOPQ VIA ITS ENGAGEMENT WITH PRIMATE CEACAM1

by

EDEN BRUSH

A THESIS

Presented to the Department of Biology
and the Robert D. Clark Honors College
in partial fulfillment of the requirements for the degree of
Bachelor of Science

June 2019

An Abstract of the Thesis of

Eden Brush for the degree of Bachelor of Science
in the Department of Biology to be taken June 2019

Title: Host-microbe evolutionary conflict: investigating the host specificity of *Helicobacter pylori* adhesin HopQ via its engagement with primate CEACAM1

Approved: _____

Dr. Matt Barber, Primary Adviser

How animals and microbes interact with each other can mean the difference between harmonious coexistence and deadly infection. These interactions create the potential for evolutionary conflict which can contribute to the antagonistic evolution of host and microbial genomes. Specific adhesion to host cells is often a necessary first step in bacterial pathogenesis; "adhesins" are proteins on bacterial surfaces that mediate host cell adhesion and subsequently, invasion and infection. The N domain of human carcinoembryonic antigen-related cell adhesion molecule 1 (CEACAM1), a host protein that modulates cell adhesion and other cell processes, is targeted and exploited by various human-associated bacterial adhesins.

The Barber Lab at the University of Oregon has recently discovered that primate CEACAM1 proteins are rapidly evolving, suggesting an evolutionary 'arms race' with the bacterial adhesins that target them. One such adhesin is Helicobacter outer protein Q (HopQ) of *Helicobacter pylori*. *H. pylori* is a human-specific bacterium that colonizes the stomach of approximately half of the human population worldwide and is the major causative agent for stomach ulcers and gastric cancer. The HopQ gene has two major

variants that are associated with both virulence and geographical location. It remains unclear how genetic diversity among adhesins such as HopQ impacts host specificity.

We tested our hypothesis that HopQ will bind differentially to various primate CEACAM1 proteins by performing biochemical binding experiments with *H. pylori* and recombinant, GFP-tagged, CEACAM1 N domains from a panel of primates. Interestingly, we found that HopQ binds to the N domains of human, chimpanzee, and gorilla CEACAM1. We also found multiple signatures of positive selection on sites of HopQ that contact rapidly evolving sites on the N domain of CEACAM1 lending support to a potential evolutionary "arms race" between the two. These findings are directly applicable to human health, as understanding the determinants of host specificity in human-associated pathogens could reveal new avenues for the treatment and prevention of infectious diseases as well as provide valuable information about which species are more susceptible to reverse zoonoses, the transfer of a disease-causing agent from humans to non-human animals.

Acknowledgements

First and foremost, I would like to thank my research principal investigator, Dr. Matt Barber, for exposing me to the field of molecular evolution of host-microbe interactions and for guiding and supporting my involvement in this research over the past two and a half years. My sincerest gratitude to him as well as my mentor, Dr. Kristin Kohler, for helping me investigate the host specificity of *H. pylori* adhesin HopQ. They have both taught me innumerable lessons and given me invaluable insights on the workings of academic research. Without their assistance and dedicated involvement in every step throughout the process, this thesis would have never been accomplished.

I would also like to thank the rest of the Barber Lab for their encouragement and willingness to help throughout this process. Additionally, I'd like to extend my gratitude to the rest of my thesis committee, Dr. Melissa Graboyes with the Clark Honors College and Dr. Peter Wetherwax with the Department of Biology. I have enjoyed taking classes and learning from both and appreciate their involvement in my academic and research endeavors.

Most importantly, I would like to acknowledge my parents for their emotional and financial support over the past four years; their unconditional love and unwavering belief in me has been monumental in the completion of this thesis and my degree.

Table of Contents

List of Figures

Introduction

The molecular evolution of host-microbe interactions is a particularly exciting area of research because host-microbe systems are highly relevant in human health and disease. Understanding the genetic and molecular basis of why human-specific pathogens cannot be transmitted to other hosts could provide insights into how to better target these pathogens for the prevention and control of infectious disease. Moreover, understanding the host specificity of a pathogen could help determine which species are more susceptible to reverse zoonoses, the transmission of a pathogen from a human to a non-human animal. Therefore, a microbe's host specificity provides information that could be vital in controlling the spread of infectious disease. As we see a global increase in industrial food animal production, the rapid movement of humans and non-human animals, and the habitats of humans and non-human wild animals intertwining with great complexity, the future promises more opportunities for human-associated pathogens to jump species and cause reverse zoonoses. Reverse zoonoses are important to investigate because bacteria and viruses could potentially mutate into a more easily transferrable, or more dangerous, form when they go from humans to non-human animals. Scientific research must be conducted in this area to provide a richer understanding of emerging and reemerging disease threats.

Bacteria encode classes of proteins called "adhesins" that allow them to adhere to host cells and grow. Pathogenic bacteria require adhesins to invade host cells and cause infection. Adhesins are located on the surfaces of bacteria making them attractive vaccine candidates because of their accessibility to antibodies. Many bacterial adhesins target and utilize carcinoembryonic antigen-related cell adhesion molecule 1

(CEACAM1) as a host receptor (Kuespert et al., 2006). CEACAM1 is part of a family of widely distributed immunoglobin superfamily-related glycoproteins that modulate diverse cellular functions including cell adhesion, differentiation, proliferation, and survival (Gray-Owen & Blumberg, 2006). There are seven well-described CEACAM family members that are all characterized by a membrane distal N domain and a variable number of additional extracellular domains (Gray-Owen & Blumberg, 2006) (Figure 1).

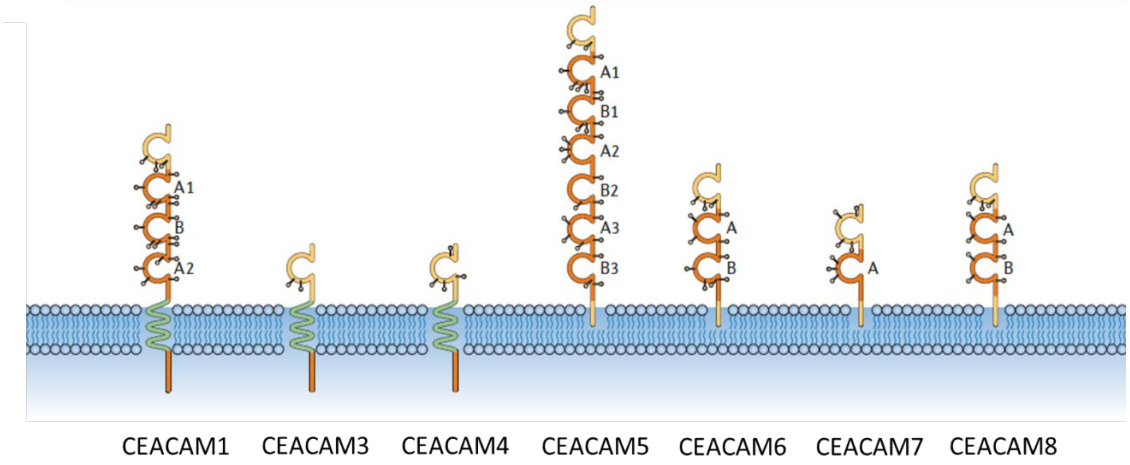
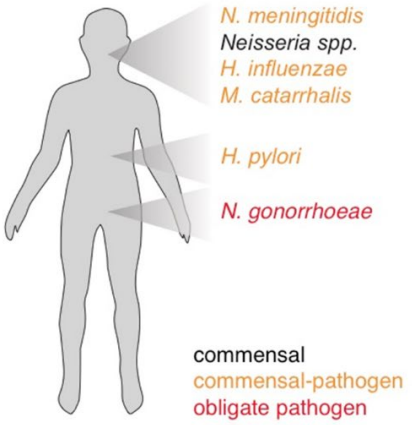


Figure 1. CEACAM family members (adapted from Gray-Owen & Blumberg, 2006)

All extracellular domains of the CEACAM family of proteins have a membrane distal N domain (yellow) and some have additional extracellular domains (orange).

CEACAM proteins are often expressed in epithelial cells where they can be in contact with bacteria that colonize epithelial tissues. Epithelial tissues line the outer surfaces of organs and blood vessels throughout the body, as well as the inner surfaces of cavities in many internal organs. The N domain of CEACAM1 is the main target of several bacteria specialized to colonize the human mucosa (Voges et al., 2010).

Bacterial attachment to CEACAM proteins results in a tight association between the bacterium and the cell surface; this can either stimulate the internalization of the bacterium or the translocation of bacterial particles which can trigger CEACAM-initiated gene expression events, thus interfering with host function (Moonens et al., 2018 and Tchoupa et al., 2014). Some of these bacteria are obligate pathogens, meaning they must cause disease in order to survive, while others are commensal pathogens—bacteria that are normally harmless to the host, but can cause disease. The variety of bacteria that bind CEACAM1 makes it an interesting host protein to study since bacterial attachment to a host cell is a critical step in bacterial infection. The bacterial adhesins that target the N domain of CEACAM1 are genetically highly diverse (Figure 2); however, these adhesins have evolved to share the same function of adhering to the N domain of CEACAM1 to mediate invasion and colonization (Voges et al., 2010).



Tissue	Bacterium	Adhesin
Respiratory epithelium	*N. meningitidis*	Opas
Respiratory epithelium	*H. influenzae*	OmpP1
Respiratory epithelium	*M. catarrhalis*	UspA1
Gastric epithelium	*H. pylori*	HopQ
Germinal epithelium	*N. gonorrhoeae*	Opas

Figure 2. Bacterial adhesins that target the N domain of CEACAM1 in epithelial tissues (adapted from Matt Barber) (Voges et al., 2010)

Multiple phylogenetically diverse adhesins target the N domain of CEACAM1 in various epithelial surfaces.

It has been speculated that CEACAM1 recognition may be a specific adaptation by bacteria to facilitate the colonization of the human mucosa (Kuespert et al., 2006). It

is possible, then, that CEACAM1 diversification is a pathogen-driven process. Infectious diseases are a powerful driver of natural selection: the host adapts because there is pressure to avoid being antagonized by the microbe, whereas the microbe adapts because it has pressure to maintain recognition of the host; this is referred to as a host-microbe evolutionary "arms race" (Figure 3).

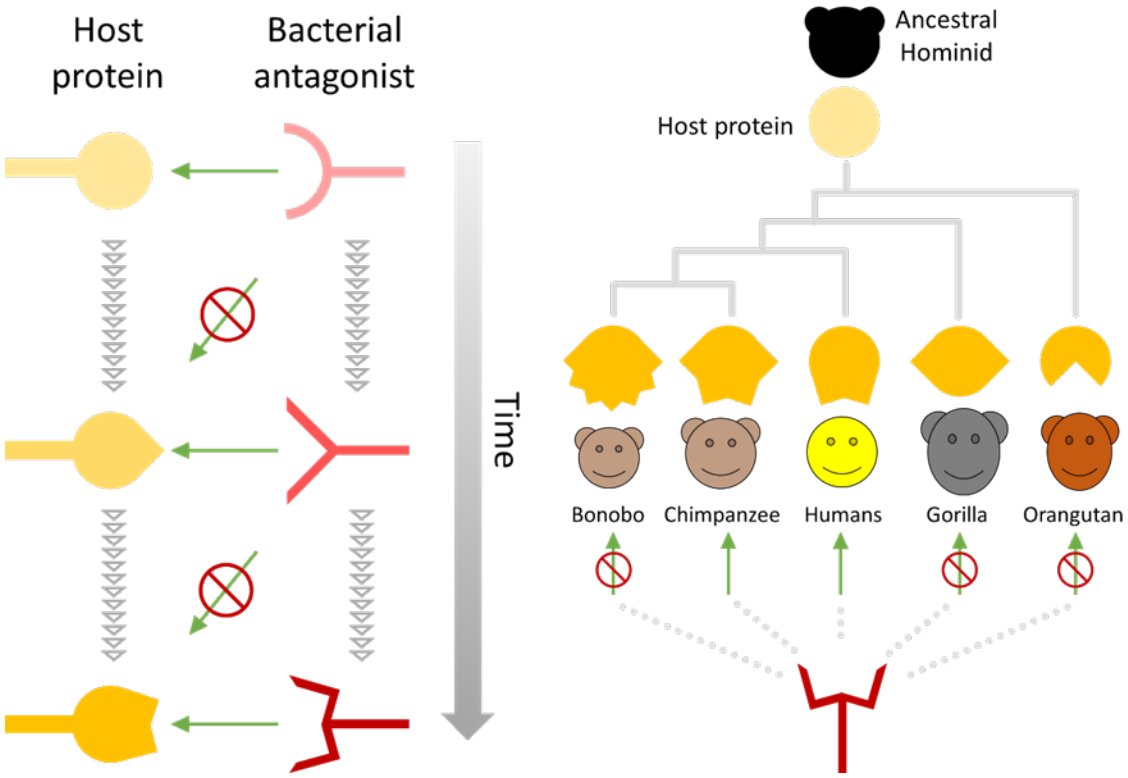


Figure 3. Host-microbe evolutionary "arms race" (adapted from Emily Baker)

Host protein variation that prevents bacterial attachment is counteracted by adaptation among bacterial adhesins. This sets up repeated episodes of counter-adaptation by both host and microbial populations. Only the bacterial adhesins that have maintained recognition of the host protein will be able to bind.

These interactions are an example of a classical "Red Queen" genetic conflict (Van Valen, 1973). The Red Queen hypothesis states that organisms must constantly evolve and proliferate not only to reproduce, but also to survive and out-compete opposing

organisms that are also ever-evolving in a constantly changing environment (Daugherty & Malik, 2012). The recurrence of positive selection can be an indicator that a host-pathogen conflict is underway. Positive selection is a type of natural selection that accelerates the fixation of an advantageous mutation in a population. Phylogenetic analyses can determine if a protein is rapidly evolving by locating signatures of positive selection. By looking at groups of closely related species, such as humans and other primates, we can better understand the consequences of rapid evolution with more resolution and statistical power. Since these species are closely related, we can identify sites on the protein of interest where evolutionary events have occurred due to natural selection and not random genetic drift.

The Barber Lab at the University of Oregon has recently discovered that the N domains of primate CEACAM1 proteins are rapidly evolving at sites that contact bacterial adhesins (unpublished) (Figure 4).

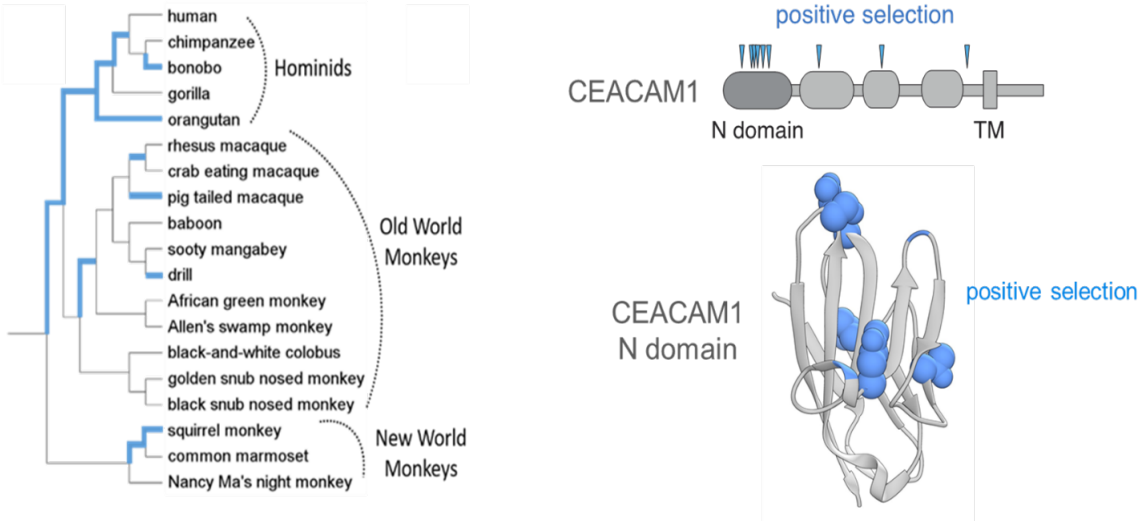


Figure 4. Signatures of positive selection in the N domain of primate CEACAM1 (Barber Lab)

Sites with signatures of positive selection (blue) are primarily on the N domain of CEACAM1. These sites are rapidly evolving in Hominids, Old World Monkeys, and New World Monkeys.

These signatures of rapid evolution suggest that CEACAM1 and the bacterial adhesins that target them may be engaged in evolutionary "arms races;" CEACAM1 variation that prevents bacterial attachment may then be counteracted by adaptation among bacterial adhesins, such that repeated episodes of counter-adaptation by both host and microbial populations occur.

Since the N domain of human CEACAM1 is a target of multiple pathogens, it is difficult to elucidate which pathogen(s), if any, might be driving its evolution. *Helicobacter pylori* is one such bacteria that specifically binds to the N domain of CEACAM1 via adhesin Helicobacter outer membrane protein Q (HopQ) in the digestive tract (Javaheri et al., 2016) (Figure 5).

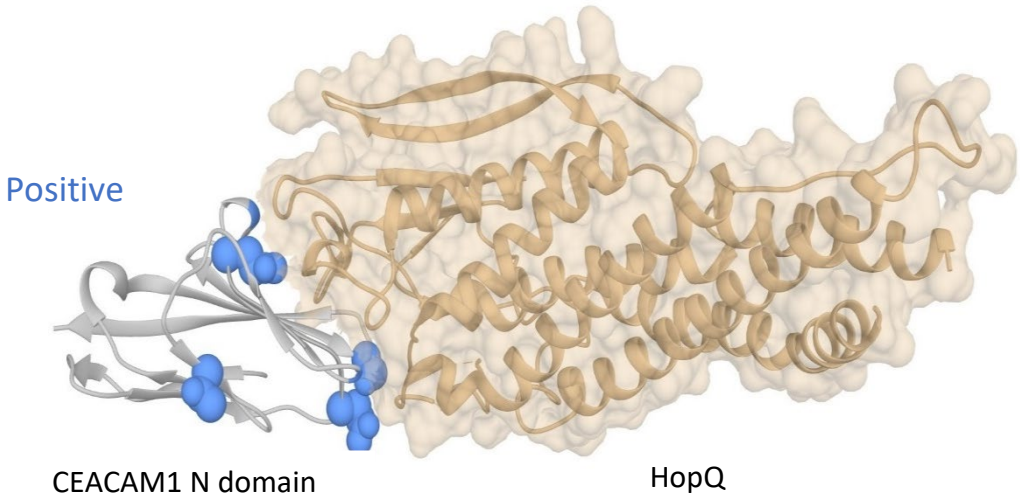


Figure 5. Rapidly evolving sites on the N domain of CEACAM1 contact *H. pylori* adhesin HopQ (Barber Lab)

Sites with signatures of positive selection on the N domain of CEACAM1 are in contact with HopQ suggesting a potential host-microbe evolutionary "arms race"

The adherence of HopQ to CEACAM1 mediates the delivery of the virulence factor CagA into host cells and stimulates the release of pro-inflammatory molecules, both of which are critical steps in an *H. pylori* infection (Bonsor et al., 2018 and Parsonnet et al., 1997).

H. pylori has coexisted with humans for tens of thousands of years, with genetic studies indicating that humans have been colonized with *H. pylori* for at least 58,000 years (Linz et al., 2007). *H. pylori* is a commensal-pathogen; though its presence in the gastric microbial community is critical for human health, it is naturally transformable, very recombinogenic, and has a high mutation rate making it vulnerable to gaining pathogenicity (Whalen & Massidda, 2015). *H. pylori* specifically colonizes the gastric epithelium of approximately half of the human population worldwide, making it one of the most common infections (Rahman et al., 2014). Though the majority of infections

are asymptomatic, some infected individuals experience chronic gastritis and stomach ulcers (Dunne et al, 2014). Significantly, *H. pylori* is also the major causative agent for gastric cancer (Javaheri et al., 2016). 1%-3% of infected individuals develop gastric cancer resulting in almost one million cases of gastric cancer being diagnosed each year (Wroblewski et al., 2010). Because of this, *H. pylori* is the second leading cause of cancer-related deaths and is classified as a group I carcinogen by the World Health Organization (Wroblewski et al., 2010). *H. pylori* infections disproportionately devastate developing countries, especially in Eastern Asia, due to a combination of untreated water, crowded conditions, and poor hygiene (Rahman et al., 2014). The global and regional burden of *H. pylori* infections is enormous; thus it is crucial that we continue to gain an understanding of the host and microbial factors that increase the risk of developing more severe clinical outcomes.

The HopQ gene exhibits diversity that represents two allelic variants, called type-I and type-II (Cao & Cover, 2002) (Figure 6).

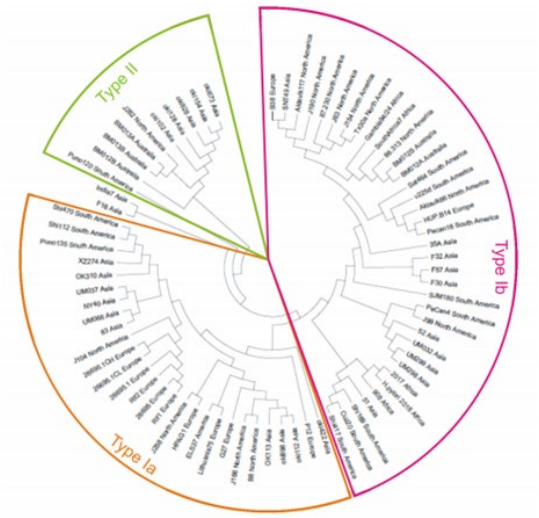


Figure 6. Diversity of HopQ gene (Javaheri et al., 2016)

HopQ genes can be grouped into two major allelic variants (type-I and type-II). The type-I HopQ genes are more diverse and can be further divided into the two subgroupings type Ia (orange shaded) and Ib (pink shaded). The type II HopQ genes are highlighted in green.

The alleles are roughly 70% identical at the amino acid level (Cao & Cover, 2002) and therefore bind to human CEACAM1 in slightly different ways (Moonens et al., 2018) (Figure 7).

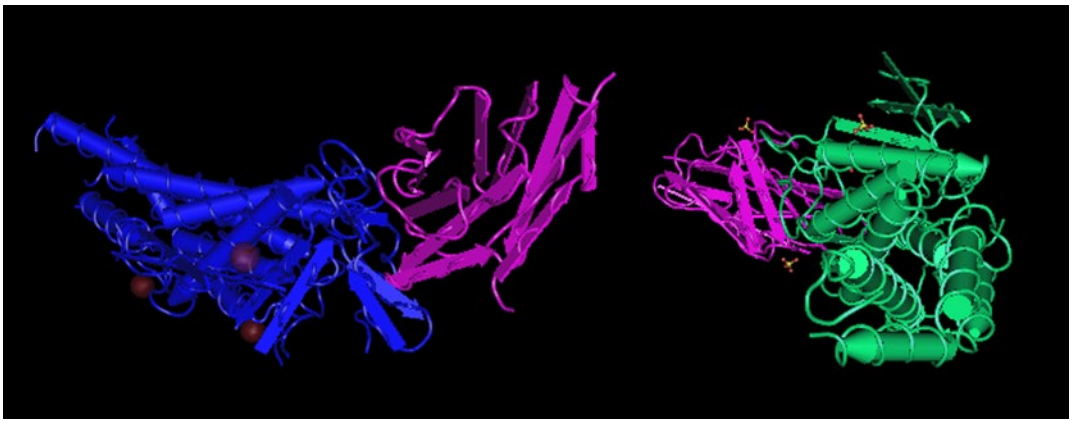


Figure 7. *H. pylori* binding to CEACAM1 differs depending on HopQ allele type (Moonens et al., 2018)

The binding of *H. pylori* to the N domain of CEACAM1 depends on the HopQ allele type as seen in the 3D reconstruction of type-I HopQ (blue) interaction with the N domain of CEACAM1 (purple) and type-II HopQ (green) interaction with N domain of CEACAM1 (purple).

H. pylori strains with the different allelic variants of HopQ differ in their virulence; strains with type-I HopQ are more virulent than strains with type-II HopQ and therefore cause worse symptoms including higher inflammation, gastric atrophy, and greater risk of developing gastric cancer (Blaser et al., 1995 and Cao & Cover, 2002). The type-I HopQ allele is associated with higher virulence because it is more frequently found in CagA-positive *H. pylori* strains (Javaheri et al., 2016 and Parsonnet et al., 1997). CagA is a virulence factor that is delivered into gastric epithelial cells via *H. pylori* secretion where it manipulates intracellular signaling of the host to promote neoplastic transformation (Hatakeyama, 2017). The interaction of type-I HopQ with the N domain of CEACAM1 is necessary for CagA transduction into the gastric epithelium and is therefore essential for *H. pylori* virulence (Javaheri et al., 2016). CagA was the first identified bacterial protein involved in human cancer; because of this, chronic infection with CagA-positive *H. pylori* strains is the strongest risk factor of gastric cancer (Blaser

et al., 1995 and Hatakeyama, 2017). *H. pylori* strains with the different allelic variants of HopQ also differ in their geographic distribution; strains with type-I HopQ appear to be more prevalent in Asian countries while strains with type-II HopQ seem to be more common in Western countries (Ohno et al., 2009). This may explain why Asian populations are at greater risk for stomach cancer (Rahman et al., 2014).

The *H. pylori* adhesin HopQ selectively binds to human CEACAM1, but not murine, bovine, or canine CEACAM1 orthologs (Javaheri et al., 2016). It is unknown whether HopQ can also recognize CEACAM1 orthologs from non-human primates; this is important to investigate since findings could give insight into evolutionary "arms races" between CEACAM1 and *H. pylori* as well as shed light on the impact of this rapid evolution on infectious disease. Since the HopQ gene exhibits such great diversity, our primary interest is to understand how this diversity might impact the host specificity of *H. pylori*. This information will increase our understanding of the genetics and evolutionary conflicts that occur at the host-pathogen protein interface in an *H. pylori* infection. Specifically, this thesis will investigate if *H. pylori* strains with type-I HopQ are rapidly evolving at sites that contact rapidly evolving sites in the N domain of CEACAM1 as well as examine the host specificity of *H. pylori* strains with type-I HopQ.

In addition to *H. pylori*, this thesis will also examine the binding interactions of *Haemophilus influenzae*, another species of bacteria that exploits the N domain of CEACAM1 (Tchoupa et al., 2015). *H. influenzae* asymptomatically colonizes the upper respiratory tract mucosa in healthy individuals (Mukundan et al., 2007). However, if *H. influenzae* makes the shift from commensal to pathogen, the *H. influenzae* adhesin outer

membrane protein P1 (OMP P1) specifically binds to CEACAM1 in the nasopharynx which leads to bacterial uptake by host cells and eventually infection (Tchoupa et al., 2015). Though OMP P1 and HopQ both exploit the N domain of CEACAM1 to cause infection, these adhesins are genetically unrelated (Voges et al., 2010). Like HopQ, the adhesin OMP P1 of *H. influenzae* selectively recognizes human CEACAM1, but not orthologs from other mammals (Tchoupa et al., 2015). To explore how the diversity of non-homologous bacterial adhesins impacts host specificity, we wanted to compare the binding interactions of two phylogenetically diverse adhesins. Thus, we also tested the binding of *H. influenzae* via OMP P1 to the N domains of various primate CEACAM1 proteins.

Though OMP P1 does not exhibit as much diversity as HopQ, *H. influenzae* strains differ based on the presence or absence of a capsule, a polysaccharide layer that lies outside the cell envelope (Falla et al., 1994). Encapsulated strains of *H. influenzae* are called typable whereas unencapsulated strains are called non-typable. There are six encapsulated serotypes, designated A through F, that each have distinct capsular polysaccharides (LaClaire et al., 2003). We investigated EAGAN, a type B *H. influenzae* strain, as well as two non-typeable strains of *H. influenzae* (strain 5 and strain 11) to determine if the presence or absence of a capsule impacts the binding of *H. influenzae* to the N domains of various primate CEACAM1 proteins.

Methods

In order to determine if sites on type-I HopQ are rapidly evolving, a phylogenetic tree was made using a diverse subset of *H. pylori* strains with type-I HopQ (Figure 8).

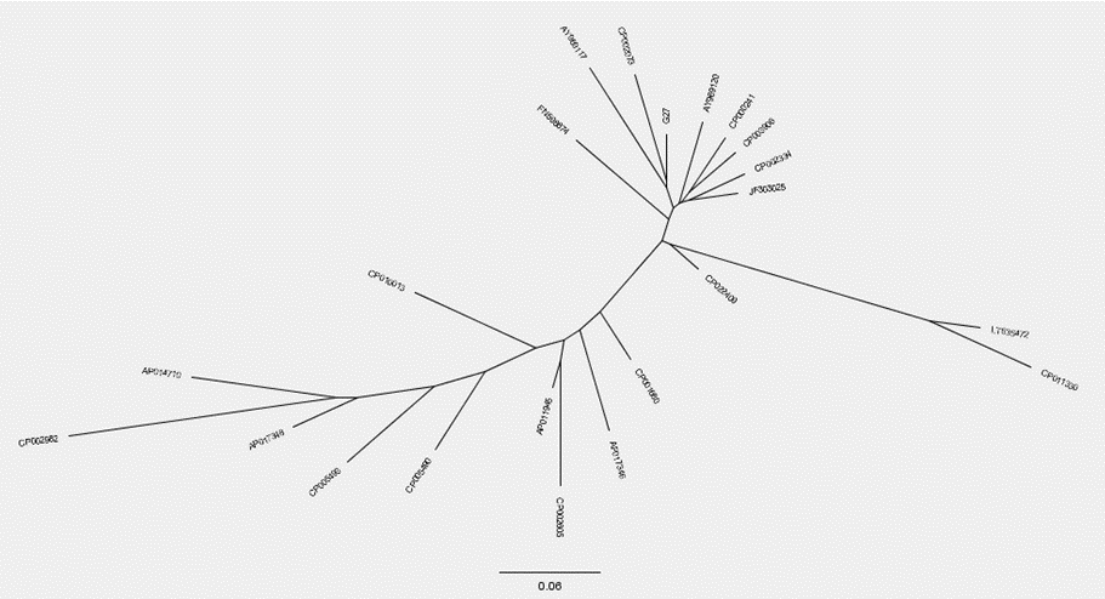


Figure 8. HopQ type-I protein phylogenetic tree

This phylogenetic tree was generated on FigTree after generating a maximum likelihood phylogeny in PhyML of approximately 20 *H. pylori* strains with type-I HopQ. These strains were chosen because they are a representation of the diversity of the HopQ type-I allele.

This tree was used in a Phylogenetic Analysis by Maximum Likelihood (PAML) which calculated the ratio of non-synonymous to synonymous substitutions (dN/dS ratio). Non-synonymous substitutions are nucleotide mutations that alter the amino acid sequence of a protein, thus changing the structure of the protein, while synonymous substitutions do not alter the amino acid sequence of a protein. Therefore, a dN/dS ratio greater than 1 indicates positive selection (driving change), a ratio less than 1 indicates

stabilizing or purifying selection, and a ratio of exactly 1 indicates neutral selection. The recurrence of positive selection is indicative of a potential host-pathogen conflict.

To test if *H. pylori* and *H. influenzae* bind to primate CEACAM1 proteins via their respective CEACAM1 N domain-binding adhesins, CEACAM1 N domains from various primates were first tagged with green fluorescent protein (GFP) for visualization and expressed in mammalian cells. They were then incubated with the bacteria of interest in a pulldown assay, a laboratory technique used to detect physical interactions between two or more proteins and a tool for confirming a predicted protein-protein interaction. After a series of washes, the bacterial pellets were resuspended in a sample buffer that denatures the proteins and makes them negatively charged. After the pulldown assay, a western blot was performed.

The western blot is a common laboratory method used in molecular biology that can detect specific protein molecules from a mixture of proteins, evaluate the size of a protein of interest, and measure the amount of protein expression. The protein molecules were first separated according to their size using gel electrophoresis, a method that separates mixtures of DNA, RNA, or proteins according to molecular sizes by applying an electric field to a gel that contains small pores. Following separation, the proteins were transferred from the gel to a blotting membrane, such that the membrane now carried all the protein bands originally on the gel. Next, the membrane went through a treatment called blocking, which prevented any nonspecific reactions from occurring during the primary antibody incubation. The membrane was then incubated with the primary antibody which specifically bound to the GFP-CEACAM1 fusion protein. Following incubation, any unbound primary antibody was washed away, and

the membrane was incubated again with a secondary antibody that specifically recognized and bound to the primary antibody. The secondary antibody was linked to a reporter enzyme that produced light, which allowed it to be easily detected via chemiluminescence and imaged via an imaging system that used a digital CCD camera to capture the emitted light as an image. These steps permitted the GFP-CEACAM1 protein bound to *H. pylori* via HopQ (or *H. influenzae* via OMP P1) to be detected from a mixture of proteins (Figure 9).

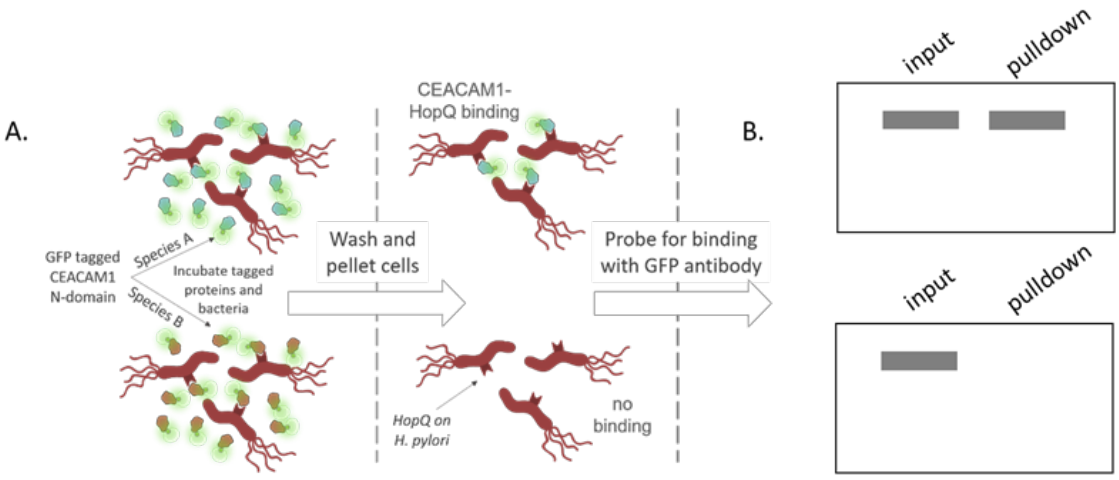


Figure 9. Bacterial pulldown assay (with *H. pylori*) and western blot (adapted from Emily Baker)

A. GFP-tagged CEACAM1 N domains from a panel of primates were individually incubated with *H. pylori.* Cells were then washed and pelleted. Binding of HopQ to CEACAM1 was determined via western blot using a GFP antibody. B. Input lanes contain 10% input (GFP-tagged CEACAM1 N domains). Pulldown lanes contain GFP-tagged CEACAM1 N domains incubated *H. pylori*. A band in the pulldown lane indicates the binding of *H. pylori* to the N domain of CEACAM1 via HopQ.

The following sections describe in detail the specific laboratory methods and materials employed:

Cell culture, transfections, and protein production:

Human embryonic kidney cell line 293T (ATCC CRL-1573) were cultured in Dulbecco's modified Eagle's medium (DMEM) containing 10% fetal bovine serum, L-glutamine, and antibiotics at 37ºC in 5% CO_2 and were sub-cultured every second day. The GFP-tagged CEACAM1 N-terminal domain proteins from various primates were expressed by cloning into a mammalian expression vector (pcDNA). Secretion of GFP-tagged CEACAM1 N-terminal domain protein from various primate species into the culture media was achieved by transient transfections of 293T cells using Lipofectamine 3000 (Life Technologies). Filtered supernatants and lysates were used for pulldown experiments.

Bacterial strains and growth conditions:

H. pylori strain G27 was grown on a horse blood agar plate under aerobic conditions (10% CO_2 and 37º) for 96 hours. All bacterial colonies were scraped into Luria broth (LB) and colony-forming units (c.f.u.) were estimated by OD_{600nm} readings according to a standard curve. *H. influenzae* strains EAGAN, 5, and 11 were grown on chocolate agar plates at 5% CO_2 and 37ºC overnight. One colony from each of the *H. influenzae* strains was precultured in brain-heart infusion broth (containing NAD and hemin) overnight until the mid-log phase (OD_{600nm} = 1) (see Future Directions for more on *H. influenzae*).

Bacterial pulldown:

Bacterial cultures were incubated with GFP-tagged CEACAM1 N-terminal domain proteins from various primates for 30 minutes at room temperature, with head-over-head rotation. After incubation, bacteria were washed 2 times with 1X PBS and boiled in 1X sodium dodecyl sulfate (SDS) sample buffer (62.5 mM Tris-HCl, pH 6.8, 2% wt/vol SDS, 10% glycerol, 50 mM dithiothreitol (DTT) and 0.01% wt/vol bromophenol blue) before sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE) and western blotting.

Western blot:

Equal volumes of CEACAM1-GFP fusion protein with bacterial culture in SDS and 10% inputs were loaded on SDS–PAGE gels. After electrophoresis, separated proteins were transferred to nitrocellulose membrane. Membranes were blocked in 5% non-fat milk for 1 hour at room temperature and incubated overnight with the primary antibody (α-GFP mouse, monoclonal, Sigma). After washing with 1X PBS, membranes were incubated with the secondary antibody (HRP-conjugated goat-α-mouse) and the ladder antibody (HRP-conjugated precision protein strep-tactin). Proteins were detected using an enhanced chemiluminescence western blotting detection kit (Advansta). Imaging was performed with a chemiluminescent western blot imager (LI-COR Biosciences).

Phylogenetic analysis:

HopQ maximum likelihood phylogeny was generated using protein sequences from approximately 20 *H. pylori* strains with the type-I HopQ allele in PhyML. A phylogenetic tree was generated using FigTree and used in a Phylogenetic Analysis by

Maximum Likelihood (PAML) that calculated the ratio of non-synonymous to synonymous substitutions (dN/dS ratio). Sites under positive selection were mapped onto three-dimensional molecular structures available from the Protein Databank (PDB) using UCSF Chimera.

Sequence alignment:

Sequences of primate CEACAM1 N domain were aligned using the MUSCLE algorithm (Edgar, 2004) and visualized with AliView.

Results

PAML generated a ratio of non-synonymous to synonymous substitutions over 1 which indicates that sites on type-I HopQ are rapidly evolving. Eight sites were found to be under positive selection; two of the sites are not on the crystal structure of type-I HopQ and are therefore not shown (Figure 10; pink residues). Importantly, some of the sites on type-I HopQ with signatures of positive selection contact sites on the N domain of CEACAM1 that are also under positive selection in primates (Figure 10; teal residues). Some of the sites on type-I HopQ with signatures of positive selection are also near sites on the N domain of CEACAM1 that frequently vary in humans, called high frequency coding polymorphisms (Figure 10; orange residues).

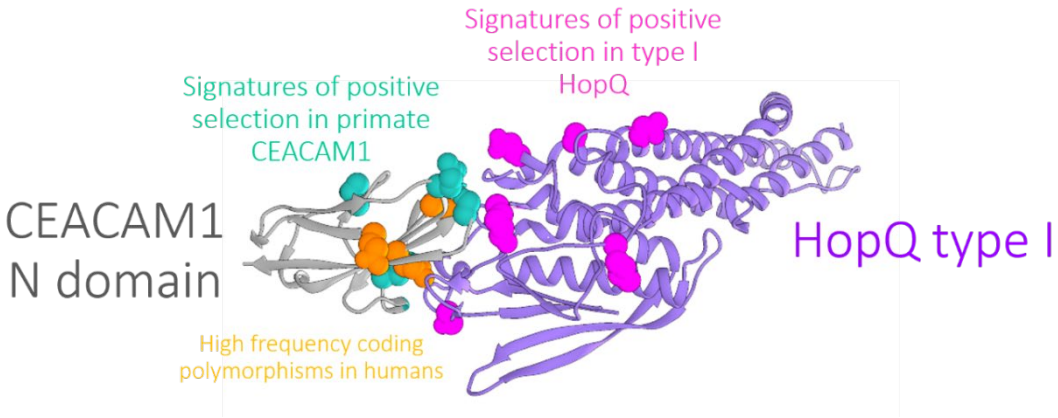


Figure 10. Signatures of positive selection in type-I HopQ

Eight residues (six shown) on type-I HopQ were identified by PAML to be under positive selection (pink residues). Teal residues indicate sites on the N domain of CEACAM1 that are under positive selection. Orange residues in the N domain of CEACAM1 are sites of high frequency coding polymorphisms in humans.

The results from the *H. pylori* binding assay show that HopQ binds to the N domains of human, chimpanzee, and gorilla CEACAM1 (Figure 11).

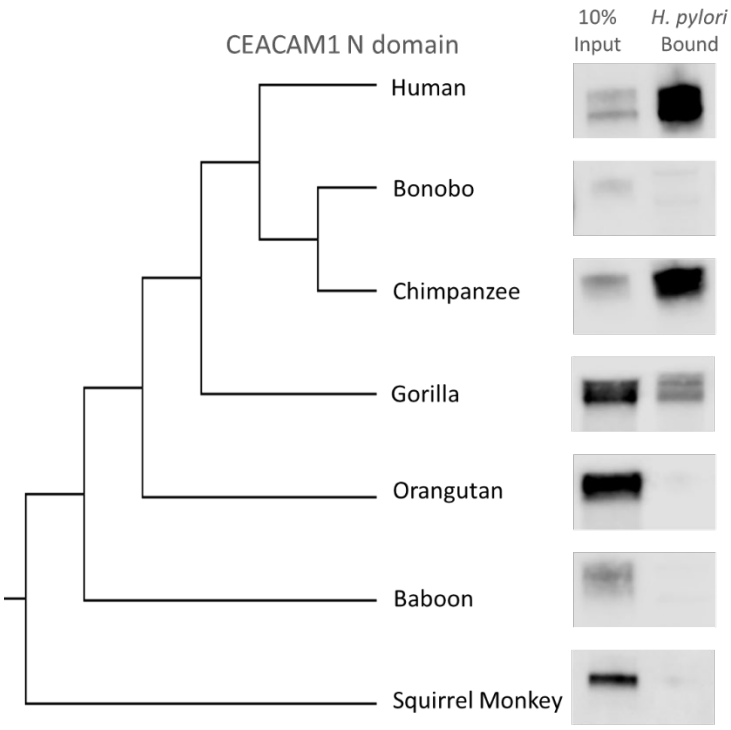


Figure 11. CEACAM1 recognition of type-I HopQ is species-specific

H. pylori bound via HopQ to the N domains of human, chimpanzee, and gorilla CEACAM1, but failed to bind to the N domains bonobo, orangutan, baboon, and squirrel monkey CEACAM1. The phylogenetic tree on the left shows the relatedness of the primates used.

The results from the *H. influenzae* binding assay show that OMP P1 from all three strains (strain 5, strain 11, and EAGAN) only binds to the N domain of human CEACAM1 (Figure 12). Therefore, was no observed difference in the binding of OMP P1 to the N domain of human CEACAM1 based on the presence or absence of a capsule.

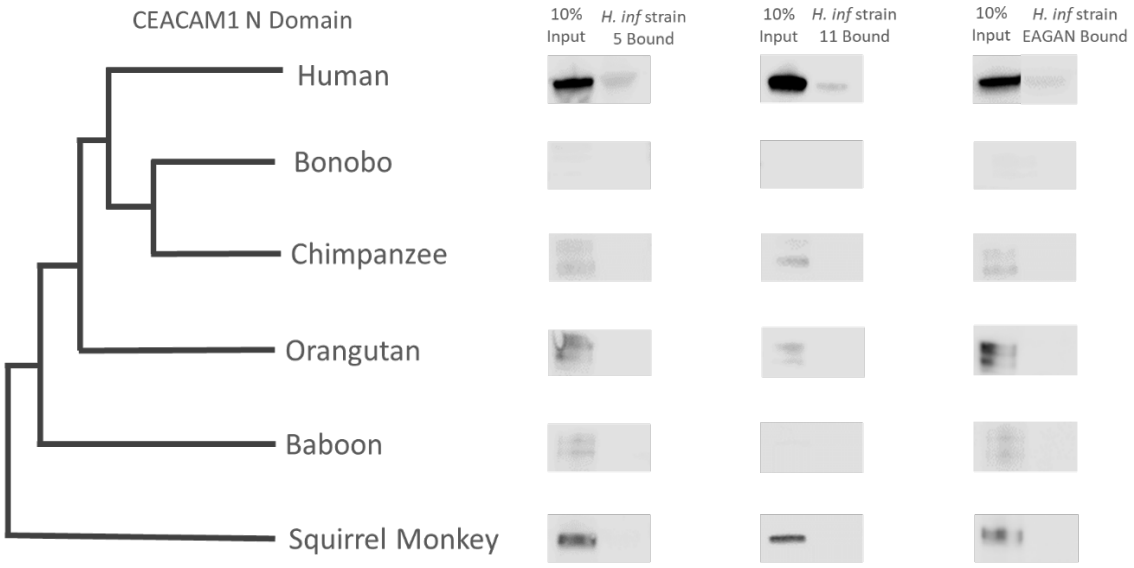


Figure 12. OMP P1 recognition of the N domain of CEACAM1 is human-specific

H. influenzae adhesin OMP P1 only binds to the N domain of human CEACAM. This was the case with regardless of presence of a capsule: strains 5 and 11 are non-typable while EAGAN is typable. The phylogenetic tree on the left shows the relatedness of the primates used.

Discussion

The discovery of signatures of positive selection on type-I HopQ was exciting because *H. pylori* is a human-associated pathogen, so selection might not be expected. However, since some of the sites in type-I HopQ that are under positive selection contact sites of high frequency coding polymorphisms in human CEACAM1, the signatures of positive selection on type-I HopQ might be indicative of *H. pylori* adaptation to human CEACAM1 variants. Importantly, some of the sites in type-I HopQ that are under positive selection contact sites on the N domain of CEACAM1 that are rapidly evolving in primates. This lends evidence to support a potential host-bacterial evolutionary "arms race" between type-I HopQ and CEACAM1 where CEACAM1 variation that prevents attachment of *H. pylori* may then be counteracted by adaptation of HopQ, setting up repeated episodes of counter-adaptation by both host and microbial populations. Despite this possibility, rapid evolution of HopQ could also be a result of its targeting by the host immune system. Since HopQ is an extracellular bacterial protein, it is exposed to antibodies and other immune proteins. Thus, host immune defense may be responsible for HopQ's evolution.

We found that the *H. pylori* adhesin HopQ type-1 (from strain G27) binds human, chimpanzee, and gorilla CEACAM1. This is surprising because *H. pylori* is considered a human-specific microbe and humans are the principal reservoir (Brown, 2000). Thus, the binding of *H. pylori* to the N domains of chimpanzee and gorilla CEACAM1 is intriguing. This could be relevant in the control of infectious disease because chimpanzees and gorillas may be susceptible to *H. pylori* infections via reverse zoonoses where *H. pylori* jumps species from humans.

Moreover, it is curious that HopQ binds the N domains of chimpanzee and gorilla CEACAM1, but not the N domain of bonobo CEACAM1 because humans are more closely related to bonobos and chimpanzees than they are gorillas. To elucidate why HopQ bound to some but not all of the N domains of primate CEACAM1 proteins, the amino acid sequences of the N domains of CEACAM1 of various primates were aligned and compared (Figure 13).

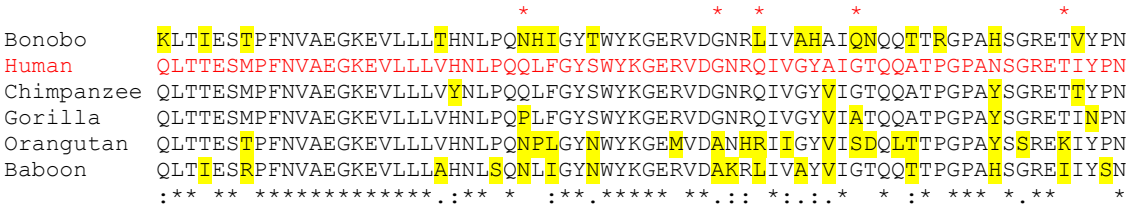


Figure 13. Amino acid sequence alignment of the N domains of CEACAM1 from various primates (adapted from Emily Baker)

The yellow represents amino acid changes in primate CEACAM1 N domains that differ from the human CEACAM1 N domain (shown in red). Red asterisks above the sequences indicate sites undergoing positive selection. Entire N domain is not shown.

The N domains of chimpanzee and gorilla CEACAM1 are very similar to the N domain of human CEACAM1: the N domains of human and chimpanzee CEACAM1 differ in only 4 amino acids (96% amino acid identity) and the N domains of human and gorilla CEACAM1 differ in 5 amino acids (95.33% amino acid identity). Due to their similarity in amino acid sequence, it is intelligible why N domains of human, chimpanzee, and gorilla CEACAM1 are bound by *H. pylori*. Interestingly, we found that the N domains of human and bonobo CEACAM1 differ in 17 amino acids (83% amino acid identity). This larger divergence from the human CEACAM1 N domain amino acid sequence might explain why *H. pylori*, a human-associated microbe, is unable to bind to the N domain of bonobo CEACAM1.

The finding that *H. influenzae* (strain 5, strain 11, and EAGAN) only binds to the N domain of human CEACAM1 was interesting because its failure to bind to N domains of non-human primate CEACAM1 proteins suggests that OMP P1 of *H. influenzae* is more species-specific than HopQ of *H. pylori*. Since the N domains of human and chimpanzee CEACAM1 differ in only 4 amino acids (Figure 13), those amino acids must be preventing OMP P1 from binding to chimpanzee CEACAM1 since it can bind to human CEACAM1. Future research might investigate which amino acids in the N domain of chimpanzee CEACAM1 prevent OMP P1 from binding since identifying the determinants of host specificity could reveal new avenues for infectious disease treatment and prevention

Bacteria are often specific to a single host species. The finding that the *H. pylori* adhesin HopQ is rapidly evolving and binds to the N domains of human, chimpanzee, and gorilla CEACAM1 not only has global health significance in terms of a reverse zoonotic threat, but also reveals new avenues to study the mechanisms underlaying host specificity and microbial adaptation. With more research, the evolutionary and molecular processes that shape host-microbe interactions, such as the engagement of CEACAM1 by HopQ, can be further elucidated to give insights into the treatment and prevention of infectious disease.

Future Directions

Since the HopQ gene has diversity that exhibits two allelic families, a logical next step is to determine if type-II HopQ also has signatures of positive selection. Since strains with different HopQ alleles differ in virulence and geographical location, comparing the evolution of the two HopQ types could be telling. In addition, the binding interactions of *H. pylori* type-II HopQ with the N domains of primate CEACAM1 proteins must be tested and compared to the binding interactions of *H. pylori* type-I HopQ with the N domains of primate CEACAM1 proteins.

A growing body of evidence suggests that many other bacteria bind to the N domain of human CEACAM1 via adhesins. In addition to *H. pylori* and *H. influenzae,* future research should test the binding interactions of other pathogens known to exploit the N domain of human CEACAM1. *Neisseria gonorrhoeae* and *Neisseria meningitidis* (bacteria which express Opa adhesins), *Moraxella catarrhalis* (bacteria which expresses USP A1 adhesin), and *Candida albicans* (a fungus that exploits human CEACAM1) are some of such pathogens (Voges et al., 2010 and Klaile et al., 2017). It is currently unknown whether these microbes bind to the N domains of non-human primate CEACAM1 proteins.

Since some strains of bacteria may express more adhesins than others, it would be beneficial to express adhesins in *E. coli* before testing their binding interactions with CEACAM1 to allow for a more controlled system. Additionally, this system would be helpful in testing the binding interactions of bacteria that express multiple different adhesins on their surface, such as certain *Neisseria* species. Site-directed mutagenesis

could be employed to mutate specific sites in adhesins to determine which sites are necessary for CEACAM1 binding (Figure 14).

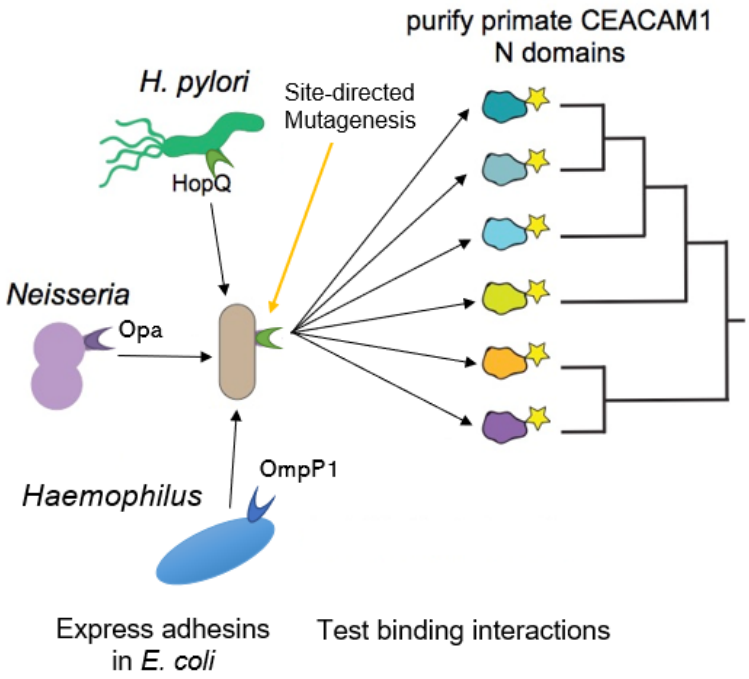


Figure 14. *E. coli* adhesin expression system

Adhesins from a multitude of pathogens that target CEACAM1 can be expressed in *E. coli* to better assess their binding interactions with CEACAM1. With this method, we can test how site-directed mutations in adhesins affect its ability to interact with primate CEACAM1.

Bibliography

Blaser, M.J., Perez-Perez G.I., Kleanthous, H., Cover, T.L., Peek, R.M., Chyou, P.H., Stemmermann, G.N., and Nomura, A. (1995) Infection with *Helicobacter pylori* Strains Possessing *cagA* Is Associated with an Increased Risk of Developing Adenocarcinoma of the Stomach. *Cancer Res* **55:** 2111-2115.

Bonsor, D.A., Zhao, Q., Schmidinger, B., Weiss, E., Wang, J., Deredge, D., Beadenkopf, R., Dow, B., Fischer, W., Beckett, D., Wintrode, P.L., Haas, R., and Sundberg, E.J. (2018) The *Helicobacter pylori* adhesin HopQ exploits the dimer interface of human CEACAMs to facilitate the translocation of the oncoprotein CagA. *The EMBO Journal* **37:** 1-17.

Brown, L.M. (2000) *Helicobacter pylori*: epidemiology and routes of transmission. *Epidemiol Rev.* **22:** 283-297.

Cao, P., and Cover, T.L. (2002) Two different Families of *hopQ* Alleles in *Helicobacter pylori. Journal of Clinical Microbiology* **40:** 4504-4511.

Daugherty, M.D., and Malik, H.S. (2012) Rules of Engagement: Molecular Insights from Host-Virus Arms Races. *Annual Review of Genetics* **46:** 677-700.

Dunne, C., Dolan, B., and Clyne, M. (2014) Factors that mediate colonization of the human stomach by *Helicobacter pylori*. *World J Gastroenterol* **20:** 5610-5624.

Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32:** 1792-1797.

Falla, T.J., Crook, D.W., Brophy, L.N., Maskell, D., Kroll, J.S., and Moxon, E.R. (1994) PCR for capsular typing of *Haemophilus influenzae*. *J Clin Microbiol.* **32:** 2382-2386.

Gray-Owen, S.D., and Blumberg R.S. (2006) CEACAM1: contact-dependent control of immunity. *Nature Reviews Immunology* **6:** 433-446.

Hatakeyama, M. (2017) Structure and function of *Helicobacter pylori* CagA, the first-identified bacterial protein involved in human cancer. *Proc Jpn Acad Ser B Phys Biol Sci* **93:** 196-219.

Javaheri, A., Kruse, T., Moonens, K., Meijías-Luque, R., Debraekeleer, A., Asche, C.I., Tegtmeyer, N., Kalali, B., Bach, N.C., Sieber, S.A., Hill, D.J., Koniger, V., Hauck, C.R., Moskalenko, R., Haas, R., Busch, D.H., Klaile, E., Slevogt, H., Schmidt, A., Backert, S., Remaut, H., Singer, B.B., and Gerhard, M. (2016) *Helicobacter pylori* adhesin HopQ engages in a virulence-enhancing interaction with human CEACAMs. *Nature Microbiology* **10:** 1-12.

Klaile, E., Müller, M.M., Schäfer, M.R., Clauder, A., Feer, S., Heyl K.A., Stock, M., Klassert, T.E., Zipfel, P.F., Singer, B.B., and Slevogt, H. (2017) Binding of *Candida albicans* to Human CEACAM1 and CEACAM6 Modulates the Inflammatory Response of Intestinal Epithelial Cells. *mBio* **8:** 1-16.

Klemm, P. and Schembri, M.A. (2000) Bacterial adhesins: function and structure. *Int J Med Microbiol* **1:** 27-35.

Kuespert, K., Pils, S., and Hauck, C.R. (2006) CEACAMs: their role in physiology and pathophysiology. *Current Opinion in Cell Biology* **18:** 565-571.

Kuespert, K., Weibel, S., and Hauck, C.R. (2007) Profiling of bacterial adhesins— host receptor recognition by soluble immunoglobulin superfamily domains. *Journal Microbiological Methods* **68:** 478-485.

LaClaire, L.L., Tondella, M.L.C., Beall, D.S., Noble, C.A., Raghunathan, P.L., Rosenstein, N.E., and Popovic, T. (2003) Identification of *Haemophilus influenzae* Serotypes by Standard Slide Agglutination Serotyping and PCR-Based Capsule Typing. *J Clin Microbiol.* **41:** 393-396.

Linz, B., Balloux, F., Moodley, Y., Manica, A., Liu, H., Roumagnac, P., Falush, D., Stamer, C., Prugnolle, F., van der Merwe, S.W., Yamaoka, Y., Graham, D.Y., Perez-Trallero, E., Wadstrom, T., Suerbaum, S., and Achtman, M. (2007) An African origin for the intimate association between humans and *Helicobacter pylori. Nature* **445:** 915-918.

Moonens, K., Hamway, Y., Nedermann, M., Reschke, M., Tegtmeyer, N., Kruse, T., Kammerer, R., Mejías-Lugue, R., Singer, B.B., Backert, S., Gerhard, M., and Remaut, H. (2018) *Helicobacter pylori* adhesin HopQ disrupts *trans* dimerization in human CEACAMs. *EMBO J* **37:** 1-17.

Mukundan, D., Ecevit, Z., Patel, M., Marrs, C.F., and Gilsdorf, J.R. (2007) Pharyngeal Colonization Dynamics of *Haemophilus influenzae* and *Haemophilus haemolyticus* in Healthy Adult Carriers. *J Clin Microbiol.* **45:** 3207-3217.

Ohno, T., Sugimoto, M., Nagashima, A., Ogiwara, H., Vilaichone, R., Mahachai, V., Graham, D.Y., and Yamaoka Y. (2009) Relationship between *Helicobacter pylori hopQ* genotype and clinical outcome in Asian and Western populations. *Journal of Gastroenterology and Hepatology* **24:** 462-468.

Osman, K.L, Jefferies, J.M., Woelk, C.H., Cleary, D.W., and Clarke, S.C. (2018) The adhesins of non-typeable Haemophilus influenzae. *Expert Review of Anti-infective Therapy* **3:** 187-196.

Parsonnet, J., Friedman, G.D., Orentreich, N., and Vogelman, H. (1997) Risk for gastric cancer in people with CagA positive or CagA negative *Helicobacter pylori* infection. *Gut* **40:** 297-301.

Rahman, R., Asombang, A.W., and Ibdah, J.A. (2014) Characteristics of gastric cancer in Asia. *World J Gastroenterol.* **20:** 4483-4490.

Tchoupa, A.K., Lichtenegger S., Reidl J., and Hauck, C.R. (2015) Outer membrane protein P1 is the CEACAM-binding adhesin of *Haemophilus influenzae*. *Molecular Microbiology* **98:** 440-455.

Tchoupa, A.K., Schuhmacher T., and Hauck C.R. (2014) Signaling by epithelial members of the CEACAM family—mucosal docking sites for pathogenic bacteria. *Cell Communication and Signaling* **12:** 27-37.

Van Valen, L. (1973) A new evolutionary law. *Evol. Theory* **1:** 1-30.

Voges, M., Bachmann, V., Kammerer, R., Gophna, U., and Hauck, C.R. (2010) CEACAM1 recognition by bacterial pathogens is species-specific. *BMC Microbiology* **10:** 1-11.

Whalen, M.B. and Massidda, O. (2015) *Helicobacter pylori*: enemy, commensal or, sometimes, friend? *J Infect Dev Ctries* **9:** 674-678.

Wroblewski, L.E., Peek, R.M., and Wilson K.T. (2010) *Helicobacter pylori* and gastric cancer: factors that modulate disease risk. *Clin Microbiol Rev.* **23:** 713-739.