TITLE:

Evolution of host-microbe cell adherence by receptor domain shuffling

Authors:

EmilyClare P. Baker[1], Ryan Sayegh[1], Kristin M. Kohler[1], Wyatt Borman[1], Claire K. Goodfellow[1,2], Eden R. Brush[1], Matthew F. Barber[1,2]

Affiliations:

1 Institute of Ecology and Evolution, University of Oregon, Eugene, OR USA

2 Department of Biology, University of Oregon, Eugene, OR USA

Correspondence to: mfbarber@uoregon.edu

ABSTRACT

Stable adherence to epithelial surfaces is required for colonization by diverse host-associated microbes. Successful attachment of pathogenic microbes via surface adhesin molecules is also the first step in many devastating infections. Despite the primacy of epithelial adherence in establishing host-microbe associations, the evolutionary processes that shape this crucial interface remain enigmatic. Carcinoembryonic antigen associated cell adhesion molecules (CEACAMs) encompass a multifunctional family of vertebrate cell surface proteins which are recurrent targets of bacterial surface adhesins at epithelial surfaces. Here we show that multiple members of the primate CEACAM family exhibit evidence of repeated natural selection at protein surfaces targeted by bacteria, consistent with pathogen-driven evolution. Inter-species diversity of CEACAM proteins, between even closely-related great apes, determines molecular interactions with a range of bacterial adhesins. Phylogenetic analyses reveal that repeated gene conversion of CEACAM extracellular domains during primate divergence plays a key role in limiting bacterial adhesin tropism. Moreover, we demonstrate that gene conversion has continued to shape CEACAM diversity within human populations, with abundant CEACAM1 variants mediating evasion of adhesins from *Neisseria gonorrhoeae*, the causative

agent of gonorrhea. Together this work reveals a mechanism by which gene conversion shapes first contact between microbes and animal hosts.

INTRODUCTION

Epithelial surfaces are typically the initial point of contact between metazoans and microbes (Brown and Clarke, 2017). As such, host factors at this barrier play an important role in facilitating or deterring microbial colonization. Bacterial attachment to epithelial surfaces is often mediated by a broad class of surface proteins termed adhesins (Kline et al., 2009). In addition to permitting the growth and colonization of commensal microbes, adhesins are also key virulence factors for many pathogenic bacteria. Adhesin-mediated adherence to host cells is often required for other downstream processes including biofilm formation, epithelial invasion, and the delivery of toxic effectors into host cells (Kline et al., 2009; Sadarangani et al., 2011) (Fig. 1). Microbial adherence can also trigger epithelial cell signaling cascades, further shaping host responses to resident and invasive microbes. Despite the fundamental importance of epithelial adherence for bacterial colonization and infectious disease pathogenesis, the dynamics of these interactions between host surface proteins and bacterial adhesions over evolutionary timescales remain a mystery. Theory predicts that exploitation of host proteins by pathogens places a significant burden on host populations, driving selection for beneficial mutations in these proteins that limit microbial invasion or virulence.

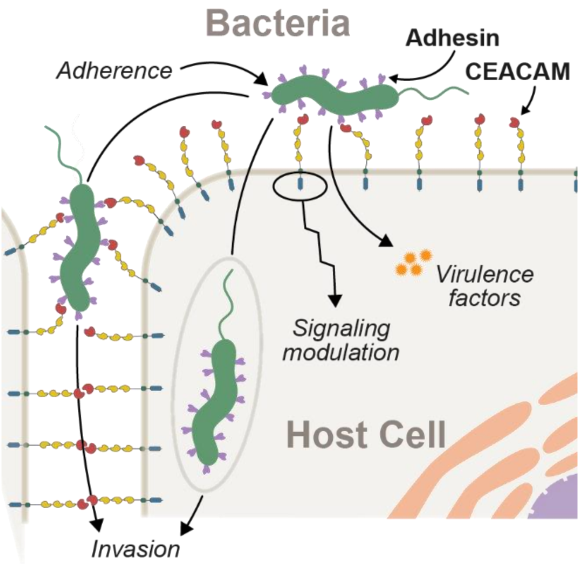


Figure 1. Interactions between epithelial CEACAMs and bacterial adhesins. Bacterial attachment to host cells via adhesin proteins (purple) facilitates epithelial adherence. Adhesins also contribute to pathogenicity by promoting invasion, modulation of host cell signaling pathways, and by promoting the delivery of virulence factors into the host cell cytoplasm.

From a microbial perspective, host defenses can also pose an existential threat resulting in reciprocal adaptation to enhance colonization, growth, and transmission. These cycles of conflict can lead to so-called Red Queen dynamics where each population must continuously adapt simply to maintain its relative fitness (Aleru and Barber, 2020; Brockhurst et al., 2014; Hamilton et al., 1990; Van Valen, 1973). However, pathogens hijack many host factors not directly involved in immunity, possibly limiting their adaptive potential in response to pathogen interaction. For example, epithelial surface proteins are not only essential for interacting with the environment but also serve crucial cellular and physiological functions including barrier maintenance, cell-cell communication, as well as coordinating host physiological and developmental pathways (Kuespert et al., 2006). Consequently, it remains unclear the extent to which such proteins are able to adapt in the face of pathogen antagonism.

A major target of bacterial adhesins on vertebrate epithelia are the carcinoembryonic antigen-related cell adhesion molecule (CEACAM) family of proteins (Gray-Owen and Blumberg, 2006). Collectively, CEACAMs are expressed on nearly all vertebrate epithelial surfaces including the microbe-rich surfaces of the urogenital, respiratory, and gastrointestinal tracts. Epithelial CEACAMs play a variety of roles in cell adhesion as well as intra- and intercellular signaling (Gray-Owen and Blumberg, 2006; Kuespert et al., 2006; Tchoupa et al., 2014). A subset of CEACAMs are also expressed on other cell types, including T-cells and neutrophils where they play important roles in immune signaling and pathogen recognition. CEACAMs typically consist of an extracellular N-terminal IgV-like domain (also termed the N-domain), a variable number of IgC-like domains, and either a membrane anchor or a cytoplasmic signaling domain (Fig. S1A). Protein-protein interactions involving CEACAMs have been shown to primarily occur through the N-domain (Kuespert et al., 2007; Markel et al., 2004). While the functions of many CEACAM proteins remain obscure, mammalian CEACAM1, CEACAM5 (also known as CEA), and CEACAM6 have been shown to contribute to immunoregulation, cell-cycle progression, and development (Gray-Owen and Blumberg, 2006; Kuespert et al., 2006; Tchoupa et al., 2014).

A growing number of bacterial genera have been found to target CEACAM proteins to promote epithelial adherence and host colonization, including *Neisseria*, *Haemophilus*, *Escherichia*, *Fusobacterium*, *Streptococcus*, and *Helicobacter* (Brewer et al., 2019; Gray-Owen and Blumberg, 2006; Javaheri et al., 2016;

Königer et al., 2016; van Sorge et al., 2021). The distinct protein structures and binding mechanisms of these adhesins indicates that CEACAM recognition has arisen independently multiple times during bacterial evolution. While capable of causing serious infections, many of the bacteria that bind CEACAMs also colonize the host as benign commensals. Bacterial CEACAM recognition can lead to several distinct outcomes (Fig.1). First, adherence to epithelial CEACAMs can provide a stable habitat to support bacterial growth and proliferation. In mice, for example, expression of human CEACAM1 is sufficient to establish stable colonization by otherwise human-restricted strains of *Neisseria meningitidis* (Johswich et al., 2013). Second, CEACAM binding may facilitate bacterial dissemination through the host epithelium (Wang et al., 1998). Third, in the case of the bacterium *Helicobacter pylori*, CEACAM-adhesin interactions promote the translocation of virulence factors into host cells via the type 4 secretion system (T4SS) leading to severe gastritis and stomach ulcers in humans (Javaheri et al., 2016; Königer et al., 2016). Finally, bacterial adhesins can potentiate CEACAM mediated signaling cascades to manipulate cellular functions, including preventing immune cell activation (Gur et al., 2019a, 2019b; Sadarangani et al., 2011), increasing cellular adhesion to prevent shedding of infected cells (Muenzner et al., 2016, 2010), and activation of apoptosis (Dje N'Guessan et al., 2007).

Previous work has indicated that mammalian CEACAMs have undergone repeated gene gain and loss as well as experienced high levels of sequence divergence (Adrian et al., 2019; Gibbs et al., 2007; Kammerer and Zimmermann, 2010; Pavlopoulou and Scorilas, 2014). These findings, coupled with the observation that many CEACAM-binding bacteria possess a narrow host range, suggests that host genetic variation may be a major determinant of bacterial colonization. In the case of CEACAM3, which is expressed exclusively in neutrophils and aids in destruction of CEACAM-binding bacteria, there is compelling evidence that residues at the interface of adhesin binding are evolving rapidly in a manner consistent with positive selection (Adrian et al., 2019). However, the consequences of epithelial CEACAM evolution for microbial interactions remain unclear. In this study, we investigate patterns of CEACAM divergence in primates and propose how CEACAM evolution and human polymorphisms have shaped interactions with pathogenic bacteria.

RESULTS

The CEACAM gene family exhibits repeated episodes of positive selection in primates

To assess patterns of primate CEACAM gene evolution, we compiled sequences of human CEACAM orthologs present in publicly available genome databases. In total nineteen representative species were analyzed including four New World monkeys, ten Old World monkeys, and five hominid species (Table S1). Some orthologs of human CEACAMs were not identified in a subset of primate genomes, likely due to losses or gains of specific CEACAMs along different lineages or incomplete genome assembly. With the exception of CEACAM3, for which additional exons annotated in Old World monkeys were included (detailed in Materials and Methods), only genomic sequences that aligned to annotated human exons were used for subsequent phylogenetic analyses. To determine if primate CEACAMs have been subject to positive selection, protein-coding sequences were analyzed using the PAML NS sites package (Yang, 2007). This program uses a maximum likelihood framework to estimate the rate of evolution of each gene or codon, expressed as the ratio of normalized nonsynonymous (dN) to synonymous (dS) nucleotide substitutions (dN/dS or ω), under different models of evolution. An excess of nonsynonymous substitutions relative to synonymous substitutions between orthologs can suggest that beneficial mutations have been repeatedly fixed by positive selection. A comparison of models that allow and disallow sites evolving under positive selection ($\omega > 1$) can determine the likelihood that a particular protein coding sequence has been evolving under positive selection. We found that eight of the twelve primate CEACAM paralogs in our dataset possess genetic signals of positive selection (Table S1) including CEACAM1, CEACAM3, CEACAM5 and CEACAM6 which have previously been shown to interact with bacterial adhesins (Gray-Owen and Blumberg, 2006). In addition, we also identified elevated ω values for CEACAM7, CEACAM8, CEACAM18 and CEACAM20.

To identify specific amino acid positions that contribute to signatures of positive selection, we analyzed CEACAM sequences using the Bayes Empirical Bayes analysis as implemented in the PAML NS sites package, as well as the programs FUBAR and MEME from the HyPhy software package (Table S2). To control for the potential impact of recombination on these inferences, we used the program GARD to identify potential breakpoints in our datasets and perform phylogenetic analyses using GARD-informed phylogenies for separate gene segments. Our analyses collectively revealed that sites with elevated ω were concentrated in the N-domain of many CEACAM proteins (Fig. 2A; Fig. S1A). Sites under positive selection in CEACAM18 and CEACAM20 were more dispersed throughout the protein, not localizing to a specific domain. The statistical support for positive selection of CEACAM18 and CEACAM20 in primates was also modest compared to that for other CEACAM proteins.

We next sought to determine the functional impact of divergence at rapidly evolving sites in the CEACAM N-domain. Residues that contribute to protein-protein interactions have been extensively annotated for CEACAM1, involving both host factors and bacterial adhesins. Overlaying sites under positive selection with known adhesin and host protein binding sites (Table S3) revealed extensive overlap between all three categories (Fig. 2B) and demonstrates that sites with elevated ω tend to cluster on the protein binding surface. Mapping rapidly-evolving CEACAM1 residues onto a co-crystal structure of human CEACAM1 and the HopQ adhesin from *H. pylori* (Moonens et al., 2018), a known interaction partner, confirmed that multiple sites fall along the binding interface of the two proteins (Fig. 2C). In summary, these results demonstrate that multiple primate CEACAM orthologs exhibit signatures of repeated positive selection within the N-domain which facilitates bacterial and host protein interactions.

CEACAM divergence in primates impairs recognition by multiple bacterial adhesins

To assess how rapid divergence of primate CEACAMs influences recognition by bacterial adhesins, we focused on CEACAM1 which is widely-expressed across different cell types (Gray-Owen and Blumberg, 2006) and has numerous well-documented microbial interactions (Table S3). Recombinant GFP-tagged CEACAM1 N-domain proteins from a panel of primate species were expressed and purified from mammalian cells (See Methods and Tables S4 & S5). Previous studies have demonstrated that the CEACAM N-domain is both necessary and sufficient to mediate interactions with bacterial adhesins (Javaheri et al., 2016; Kuespert et al., 2007; Markel et al., 2004). We focused our experiments on CEACAM1 binding to two distinct classes of bacterial adhesins: HopQ encoded by *Helicobacter pylori*, and the Opa family adhesins expressed by *Neisseria* species.

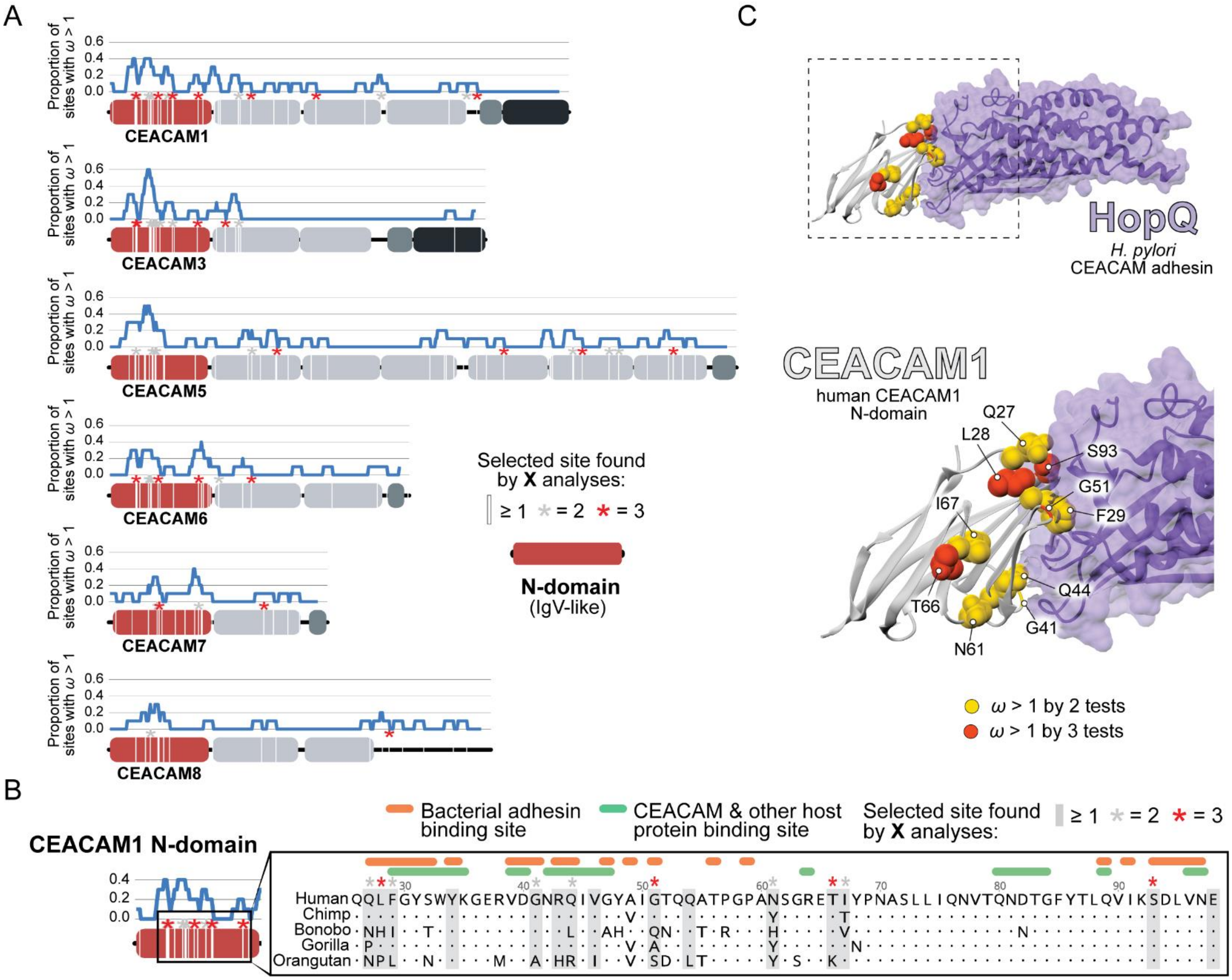


Figure 2. Rapid evolution of primate CEACAM N-domains. A) Sites in CEACAM proteins exhibiting elevated ω. Domain structure of CEACAMs outlined in red (N-domain) and gray (all other domains). All rapidly evolving sites identified by at least one phylogenetic analysis (PAML, FUBAR, or MEME) are marked by a white line, sites identified by two or three tests signified by gray and red asterisks respectively. Blue line shows the proportion of rapidly evolving sites identified across a ten amino acid sliding window. B) Multiple sequence alignment of hominid CEACAM1 residues 26-98. Sites identified as evolving under positive selection and sites known to influence adhesin and host protein binding are highlighted (Table S3). C) Protein co-crystal structure of human CEACAM1 and the HopQ adhesin from H. pylori strain G27 (PDB ID: 6GBG). CEACAM1 sites identified as evolving under positive selection by two or more tests highlighted.

The HopQ adhesin is a *H. pylori*-specific outer membrane protein that appears to be universally encoded by *H. pylori* strains and whose interaction with human CEACAM1 has been well-characterized (Bonsor et al., 2018; Javaheri et al., 2016; Königer et al., 2016; Moonens et al., 2018). For our assays we used the common *H. pylori* laboratory strains G27 (Baltrus et al., 2009), J99 (Alm et al., 1999), and Tx30a (ATCC® 51932), which have previously been confirmed to bind human CEACAM1 (Javaheri et al., 2016). The HopQ proteins encoded by these strains encompass the two major divisions of HopQ diversity, termed Type I and Type II (Cao and Cover, 2002; Javaheri et al., 2016). Strains G27 and J99 both encode a single copy of a Type I HopQ adhesin, while Tx30a encodes a Type II HopQ adhesin. All strains include extensive divergence in the CEACAM1 binding region (Bonsor et al., 2018; Moonens et al., 2018). Opa proteins are a highly diverse class of adhesins encoded by *Neisseria* species that are structurally distinct from the HopQ adhesin (Bonsor et al., 2018; Fox et al., 2014; Moonens et al., 2018; Sadarangani et al., 2011). Despite their limited sequence identity, both Opa52 and Opa74 are known to bind human CEACAM1 (Roth et al., 2013). Because *Neisseria* species typically encode

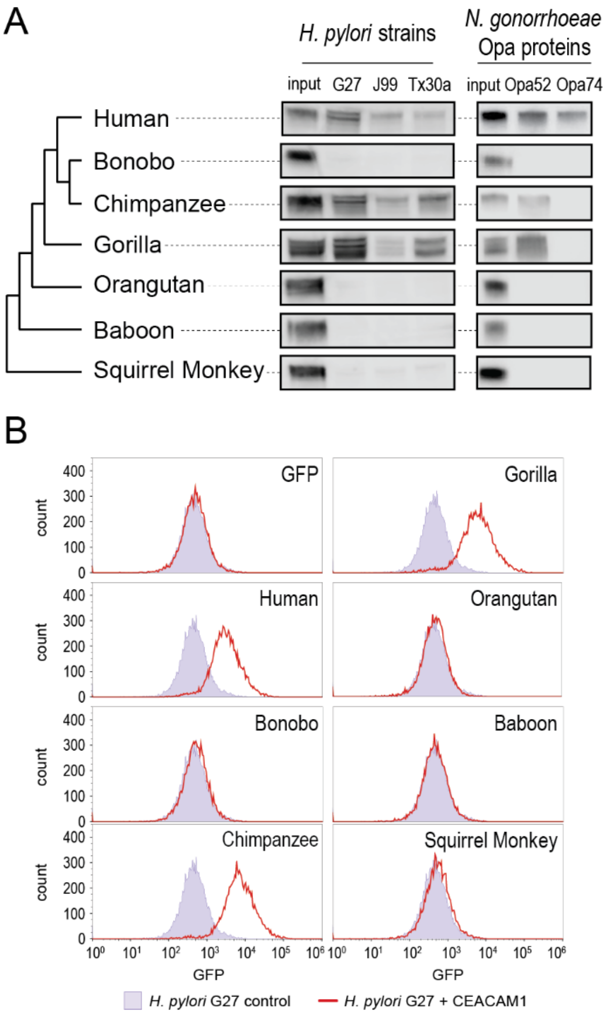


Figure 3. CEACAM1 divergence in great apes restricts bacterial adhesin recognition. A) Binding between primate GFP tagged CEACAM1 N-domain orthologs and bacteria determined by pulldown assays and visualized by western blotting. Input is 10% CEACAM1 protein used in bacterial pulldowns. Primate species relationships indicated by phylogenetic tree to the left. B) Pulldown experiments of *H. pylori* strain G27 incubated with CEACAM1 N-domain constructs or GFP alone assayed by flow cytometry. Binding indicated by GFP fluorescence.

multiple unique phase-variable Opa variants, individual Opa genes from *N. gonorrhoeae* were cloned and expressed heterologously in K12 *Escherichia coli*, which does not bind to CEACAM proteins.

To assess pairwise interactions between primate CEACAMs and bacterial adhesins, we incubated recombinant CEACAM1 N-domain proteins with individual bacterial strains. Bacterial cells were washed, pelleted, and the presence of bound CEACAM1 protein was assessed by western blot. We observe that all bacterial strains tested bind to the human CEACAM1 N-domain, consistent with previous studies (Fig. 3A). Incubation of *H. pylori* strain G27 with GFP alone fails to yield detectable signal, confirming that binding is CEACAM-dependent (Fig 3B). Furthermore, a *Δhopq* mutant of strain G27 does not exhibit significant CEACAM1 binding, consistent with previous reports that HopQ is the sole CEACAM-binding adhesin present in these strains (Fig. S2).

Examining non-human CEACAM1 bacterial binding, the chimpanzee CEACAM1 N-domain, which differs from the human protein at four amino acid positions, binds to all adhesin-expressing strains except Opa74. Gorilla CEACAM1, which differs from the human N-domain at five sites (three non-overlapping with chimpanzee) is also unable to bind Opa74 but does bind *H. pylori* strains and Opa52. Orangutan CEACAM1 is unable to interact with any bacterial strains, nor do baboon and squirrel monkey. We noted that despite the limited species divergence between bonobos and chimpanzees, bonobo CEACAM1 does not bind any of the tested bacterial strains (Fig. 3A). Previous studies have found the results of CEACAM-binding assays to be consistent between western blotting and by flow cytometry (Adrian et al., 2019; Javaheri et al., 2016; Königer et al., 2016; Kuespert et al., 2007). We confirmed this for our system with *H. pylori* strain G27, using flow cytometry to detect specific binding of GFP-tagged CEACAMs on the bacterial cell surface (Fig. 3B). These results demonstrate that CEACAM1 N-domain divergence between closely-related primate species, even within the great apes, determines bacterial recognition in an adhesin-specific manner.

Recurrent gene conversion of primate CEACAM N-domains

The inability of *H. pylori* strains or *N. gonorrhoeae* adhesins to bind bonobo CEACAM1 was surprising given bonobo's close phylogenetic relationship to both humans and chimpanzees. While archaic humans are believed to have diverged from our primate relatives at least 5 million years ago, the major divergence between chimpanzees and bonobos occurred only one to two million years ago (Prado-Martinez et al., 2013). Closer inspection revealed that the bonobo CEACAM1 N-domain sequence is unusually divergent from that of both humans and chimpanzees, while other regions of the coding sequence show higher degrees of identity (Fig. S3). To investigate bonobo CEACAM1 evolution further, we first validated the bonobo CEACAM1 N-domain sequence present in our bonobo reference genome through comparison of assemblies and sequencing reads from multiple bonobo individuals. Having confirmed the identity of the bonobo CEACAM1 reference sequence, we compared this gene to sequences from other hominids. Relative to its orthologs in humans and chimpanzees, bonobo CEACAM1 differs at nearly 20% of sites in the N-domain whereas humans and chimpanzees differ at only about 4% of sites. In contrast, outside of the N-domain bonobo CEACAM1 diverges from humans and chimpanzees at approximately 2% of sites, while human and chimpanzee CEACAM1 differ at around 1% of sites. We also noted that the number of divergent sites between bonobo and human in the N-domain (18 residues) is nearly identical to the number of divergent sites between bonobo and chimpanzee (20 residues), despite the closer phylogenetic relationship between bonobos and chimpanzees. In fact, the divergence between the bonobo and chimpanzee CEACAM1 N-domains is greater than that between chimpanzee and the earliest diverging member of the hominid clade, orangutan (81% versus 83% amino acid identity respectively). A comparison of N-domain sequences for CEACAM5, another rapidly evolving CEACAM, further highlights the extreme divergence of bonobo CEACAM1. Between human CEACAM5 and the bonobo and chimpanzee CEACAM5 sequences there are only ten and nine amino acid changes respectively, while bonobo and chimpanzee differ at only five sites along the entire length of the N-domain (Fig. S3B).

The degree of divergence within the N-domain of bonobo CEACAM1 suggests processes other than sequential accumulation of single nucleotide mutations could be responsible. One mechanism by which this could occur is through gene conversion, a form of homologous recombination in which genetic material from one location replaces sequence in a non-homologous location, often with substantial sequence similarity (Chen et al., 2007). Gene conversion is thought to be an important source of genetic novelty and a mechanism that can accelerate adaptation (Bittihn and Tsimring, 2017; Daugherty and Zanders, 2019). To determine if inter-locus recombination has shaped the evolution of CEACAM genes in primates, we looked for evidence of discordance between species and gene trees. Gene-species tree discordance can be an indication of multiple evolutionary processes, including a history of gene conversion between paralogs. In a maximum likelihood-based phylogeny of full-length CEACAM coding sequences, clades containing single CEACAM paralogs were inferred with robust statistical support (Fig. 4A & Fig. S4). In general, the relationships between CEACAM homologs are inferred with high confidence and reflected species relationships as expected for the divergence of orthologous coding sequences. To determine if there have been domain-specific instances of gene conversion, we constructed phylogenetic trees of specific CEACAM domains. Typically, we expect paralog sequences to form clearly defined clades reflecting species divergence. This is the pattern we observe for full-length CEACAM coding sequences, indicating that overall the paralogs have remained distinct since their initial duplication and have steadily diverged between species. Specific CEACAM domain sequences generally follow this pattern (Fig. 4B, Sup. Figs. S5-7). However, the N-domains of CEACAM1, CEACAM3, CEACAM5 and CEACAM6 deviate strikingly from this norm and form a single monophyletic group (hereafter called CCM$_{1356}$), albeit one with low bootstrap support (Fig. 4B, S5). Within the CCM$_{1356}$ clade we observe that rather than clustering by paralog, N-domains are split into subclades representing the three major primate lineages (Fig. 4C, Extended Fig. 1). In general, the close phylogenetic relationship of sequences within these clades is well-supported. This topology suggests that these CEACAM N-domains are more similar to paralogous domains within the same species or primate lineage than they are to their respective orthologs across species. Several well-supported nodes provide further evidence that gene conversion is driving concerted evolution within the CCM$_{1356}$ clade (Fig. 4C). Certain pairs of N-domains, such as CEACAM3 and CEACAM1 in gorilla and CEACAM1 and CEACAM5 in orangutan, form monophyletic groups with strong bootstrap support. As these relationships are not observed for the other domains of these CEACAM proteins, this suggests conversion events affecting only the N-domains of these CEACAMs occurred in these species. New World monkeys provide the most striking phylogenetic evidence of gene conversion among primates. For each of the four New World monkey species examined, the N-domains of CEACAM1, CEACAM5, and CEACAM6 are all more closely related within species than to their orthologs in other species, suggesting gene conversion has independently acted on the N-domains of these three CEACAMs at least four times within this single clade (Fig. 4C). These findings are consistent

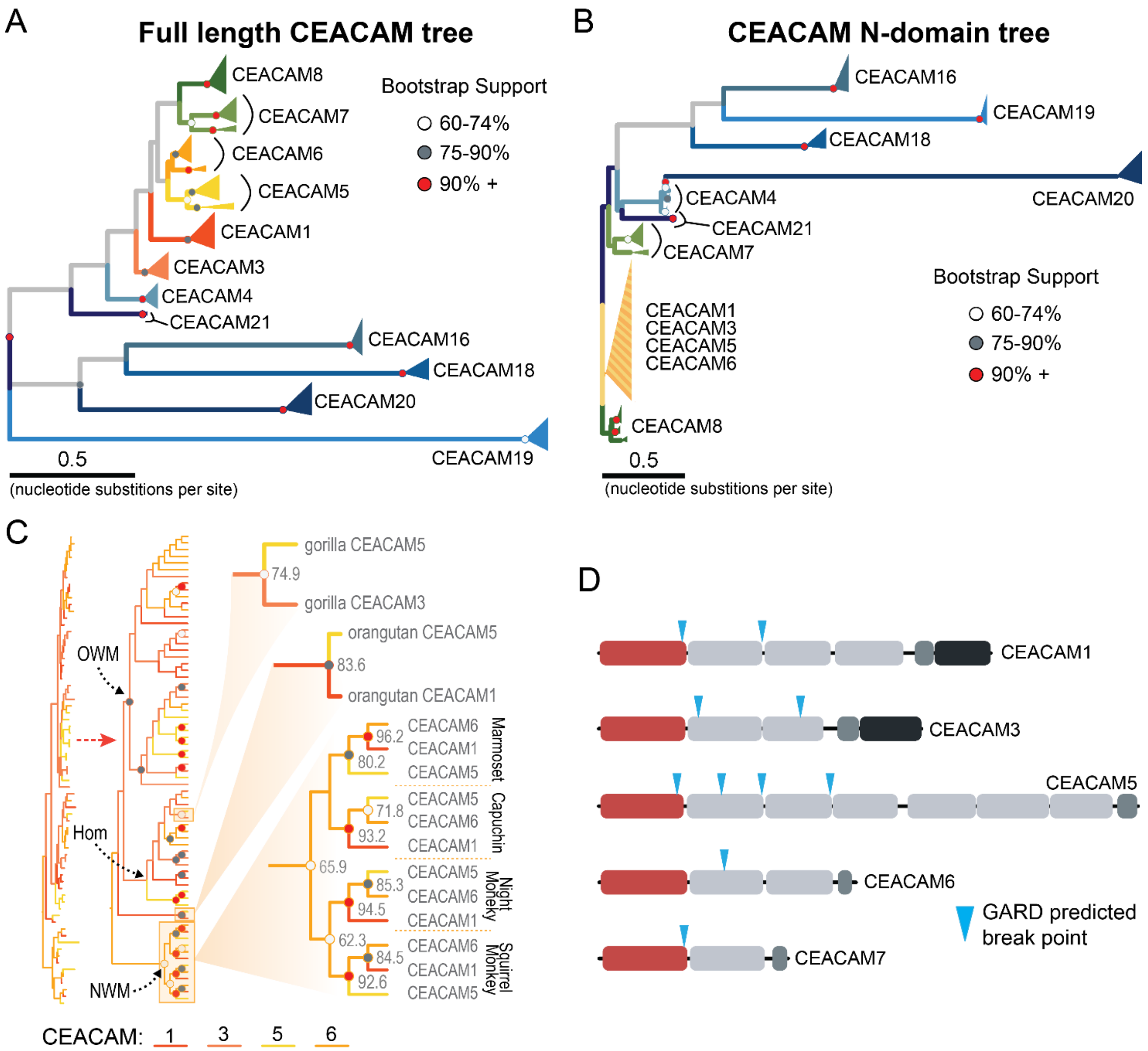


Figure 4. Recurrent episodes of gene conversion among adhesin-binding CEACAMs. A) Maximum-likelihood based phylogenetic reconstruction of full-length primate CEACAM protein coding sequences. B) Phylogenetic reconstruction of the IgV-like (N-domain) of primate CEACAM proteins. C) Expanded view of the clade containing the N-domains of CEACAM1, CEACAM3, CEACAM5 and CEACAM6 from panel B. Arrows indicate nodes designating clades for Old World monkeys (OWM), hominids (Hom) and New World monkeys (NWM). Specific subclades, gorilla CEACAM3 and CEACAM5, orangutan CEACAM5 and CEACAM1, and New World monkeys are further magnified and highlighted with bootstrap support at nodes. D) Domain structures of CEACAM proteins predicted to have undergone recombination by GARD analysis with sites of predicted breakpoints highlighted (blue arrows). CEACAM N-domains are denoted in red.

with N-domains of CEACAMs 1, 3, 5, and 6 undergoing widespread concerted evolution, likely facilitated by gene conversion.

To further test for evidence of gene conversion acting on primate CEACAM family members, we applied the GARD algorithm from the HyPhy software package. GARD detects topological changes between trees inferred from segments of a gene alignment, assesses the likelihood they are consistent with recombination, and identifies potential breakpoints. Consistent with our phylogenetic examination of CEACAM homologs, GARD detects strong evidence of recombination for CEACAM1, CEACAM3, CEACAM5 and CEACAM6 (Fig. 4D). In all cases, breakpoints were identified at the C-terminus of the N-domain or in immediately adjacent IgC domains. This pattern is consistent with repeated N-domain gene conversion between CCM_{1356} paralogs (Fig. 4D) and is also in line with our phylogenetic reconstructions of CEACAM IgC domains (Sup Fig.S6). In addition to CEACAM1, CEACAM3, CEACAM5, and CEACAM6, GARD also indicates a recombination breakpoint for CEACAM7 that would encompass the N-domain. While we do not detect discordance in our N-domain gene tree that implicates gene conversion involving CEACAM7, there is a single instance in the IgC domain tree of a gorilla CEACAM5 IgC domain grouping more closely with homologs of the IgC domain of CEACAM7 (Fig. S5). A breakpoint in this region is also consistent with CEACAMs with rapid N-domain evolution being involved in gene conversion events as well as previous analyses (Zid and Drouin, 2013). Together these results support a model in which gene conversion between rapidly diverging CEACAMs has contributed to N-domain diversification during primate evolution.

Rapidly evolving regions of CEACAM1 are sufficient to block bacterial adhesin recognition

Phylogenetic analyses confirm that the bonobo CEACAM1 N-domain is not closely related to other primate CEACAM1 sequences but fail to strongly support its relationship to any other single CEACAM N-domain. Reasoning that the extant bonobo CEACAM1 gene may have arisen from multiple iterative recombination events, we performed a BLAST search of genomes on the NCBI database using base pairs 103-303 of the bonobo CEACAM1 sequence (corresponding to resides 1-67 of the N-domain) as our query. Human and chimpanzee are roughly 86% identical to bonobo CEACAM1 in this region versus 99% identical (a single nucleotide change) in the remaining 120 base pairs (Fig. S3A). This search identifies orangutan

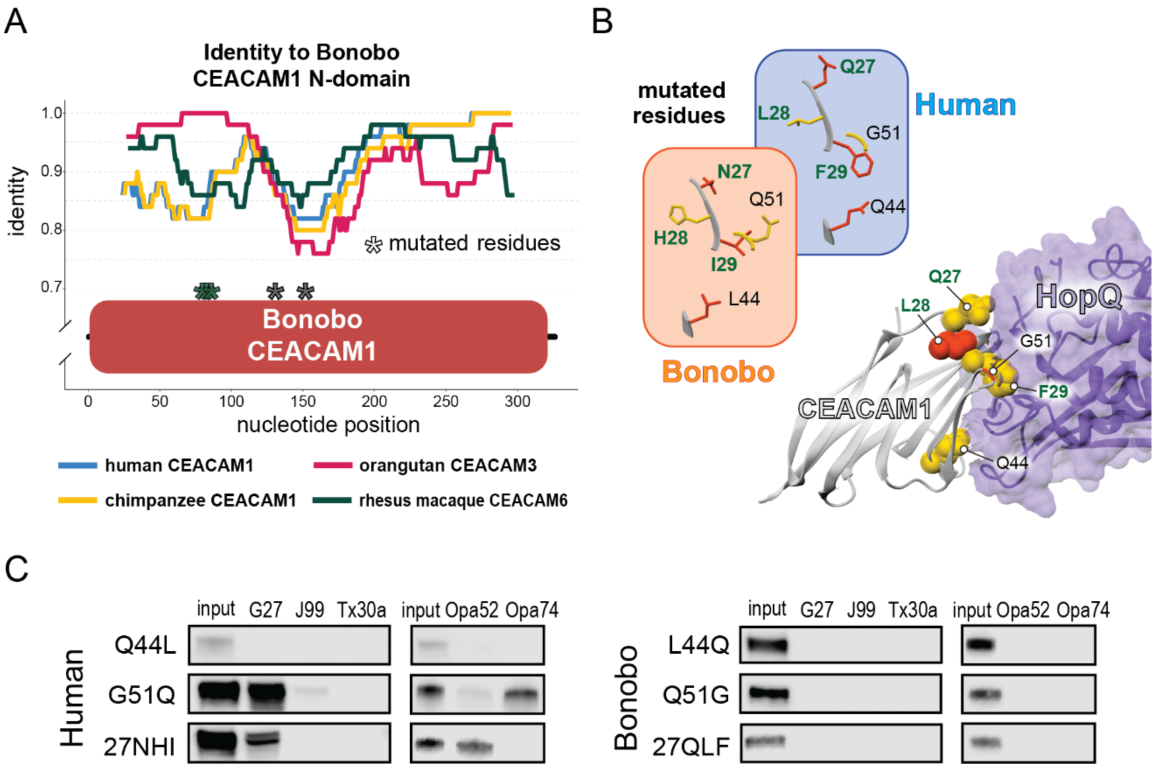


Figure 5. Rapid divergence of the bonobo CEACAM1 N-domain impairs bacterial adhesin recognition. A) Graph shows a fifty base pair sliding window of identity between bonobo CEACAM1 N-domain sequence and other CEACAM sequences. Asterisks mark locations of residues mutated for adhesin binding assays. B) Windows show amino acids and their structures at sites selected for mutational analysis in humans and bonobos. Lower right is a protein co-crystal structure of human CEACAM1 and *H. pylori* G27 HopQ with sites selected for mutagenesis highlighted. C) Binding between chimeric human and bonobo CEACAM1 N-domain constructs and bacterial strains assayed by pulldown experiments and visualized by western blotting.

CEACAM3 as the closest match. While the similarity between the first 120bp of bonobo CEACAM1 and orangutan CEACAM3 is striking and the final third of the nucleotide sequence is nearly identical to human and chimpanzee CEACAM1, other segments of bonobo CEACAM1 are still quite divergent from all other N-domain sequences (Fig. 5A). A BLAST search of this region in bonobo CEACAM1 (base pairs 221-380) indicates the greatest similarity is with the analogous region from rhesus macaque CEACAM6. However, the increased similarity of macaque CEACAM6 in this region compared to other CEACAMs is marginal (Fig. 5A).

The extreme divergence of the bonobo CEACAM1 N-domain from other CEACAM1 homologs in even its closest relatives could indicate that this particular sequence has been evolving independently of other N-domain alleles for a long period of time as a result of balancing selection. This has been observed for other genes involved in host-pathogen conflicts, most notably major histocompatibility complex (MHC) alleles. In this case, we might expect to identify alleles similar to bonobo CEACAM1 currently circulating in other hominid populations, and likewise alleles similar to CEACAM1 sequences observed in humans and chimpanzees may be found in the larger bonobo population. In a search of human genetic variation data available through the International Genome Sample Resource (IGSR) accessed through the Ensembl webserver (www.ensembl.org) there is no evidence for any alleles with similarity to bonobo CEACAM1 circulating within human populations. Searching population data from the Great Apes Genome Project (Prado-Martinez et al., 2013), alleles similar to bonobo CEACAM1 are not found for chimpanzees, gorillas, or orangutans. Likewise, CEACAM1 alleles similar to those found in humans and chimpanzees are not observed in any of the bonobo genomes from the same dataset. Given the information at hand, it is difficult to precisely determine the series of mutational events that produced the bonobo CEACAM1 allele or determine the likely origin point of this allele in the diversification of hominids. However, these results are consistent with multiple independent instances of gene conversion giving rise to bonobo CEACAM1, with subsequent fixation of this haplotype in bonobo populations since their divergence from chimpanzees over the last million years.

Given the large number of residue changes between human and bonobo CEACAM1, we next sought to determine if a subset of rapidly-evolving sites are sufficient to either impair or restore recognition by bacterial adhesins. To test this, we generated CEACAM1 N-domain proteins in which a subset of residues between humans and bonobos were swapped. We focused on sites that are identical in humans and chimpanzees but differ in bonobos and which exhibit high ω across primates, resulting in a total of five tested sites (Fig. 5A & B). Of these residues we chose to mutate adjacent amino acids 27-29 as a single group. This patch of sites is highly variable among the rapidly-evolving CEACAMs, particularly CEACAM1, CEACAM3 and CEACAM5 (Fig. S8). None of the “humanized” mutants in the bonobo CEACAM1 background were sufficient to confer binding (Fig. 5C). In contrast, introduction of bonobo residue 44 into human CEACAM1 (mutation

Q44L) prevents binding by *H. pylori* and Opa expressing strains, while introduction of bonobo variable sites 27-29 abolishes binding to Opa74 (Fig. 5C). Mutation G51Q has no appreciable impact on binding by *H. pylori* strain G27 or Opa74, but blocks binding by strain Tx30a and reduces binding to J99 and Opa52. Collectively these results reveal that multiple single positions in human CEACAM1 exhibiting signatures of positive selection are sufficient to impair recognition by multiple bacterial adhesins. Moreover, these findings also demonstrate how instances of gene conversion between CEACAM paralogs could serve as large-effect adaptive mutations during conflicts with pathogens.

Abundant human CEACAM1 polymorphisms impair bacterial recognition

Pervasive evidence of positive selection acting on CEACAMs in primates raises the question as to whether CEACAM variants that evade pathogen recognition are currently segregating in human populations. To explore the existence of human CEACAM variants and their consequences for bacterial interactions, we queried human single nucleotide polymorphism (SNP) and haplotype data for rapidly evolving CEACAM paralogs available from the International Genome Sample Resource accessed through the Ensembl genome browser (See Methods). We found that variation in the N-domains of CEACAM6, CEACAM7 and CEACAM8 predominantly consists of polymorphisms not shared with other CEACAM proteins and found on isolated haplotypes. In contrast, CEACAM1, CEACAM3 and CEACAM5 N-domain variation is composed primarily of extended haplotypes (Sup Figs. S9-12). Furthermore, these extended haplotypes increase similarity between CEACAM1, CEACAM3 and CEACAM5, consistent with possible gene conversion events. Indeed, some haplotypes not only have changes at nonsynonymous sites that increase similarity with these CEACAMs, but also include multiple shared synonymous changes. These observations suggest that gene conversion among CEACAMs has occurred relatively recently and may be ongoing in human populations.

A search of polymorphisms for CEACAM1 in human populations reveals three high-frequency nonsynonymous variants within the N-domain: Q1K (rs8111171), A49V (rs8110904) and Q89H (rs8111468) (Fig. 6A). The haplotype containing all three alternative alleles is the most frequent non-reference CEACAM1 haplotype annotated, occurring in 14% of the human population overall and in up to 43% of individuals in African populations (Fig. 6A). In total, nearly 17% of all sequenced individuals carry at least one of these high

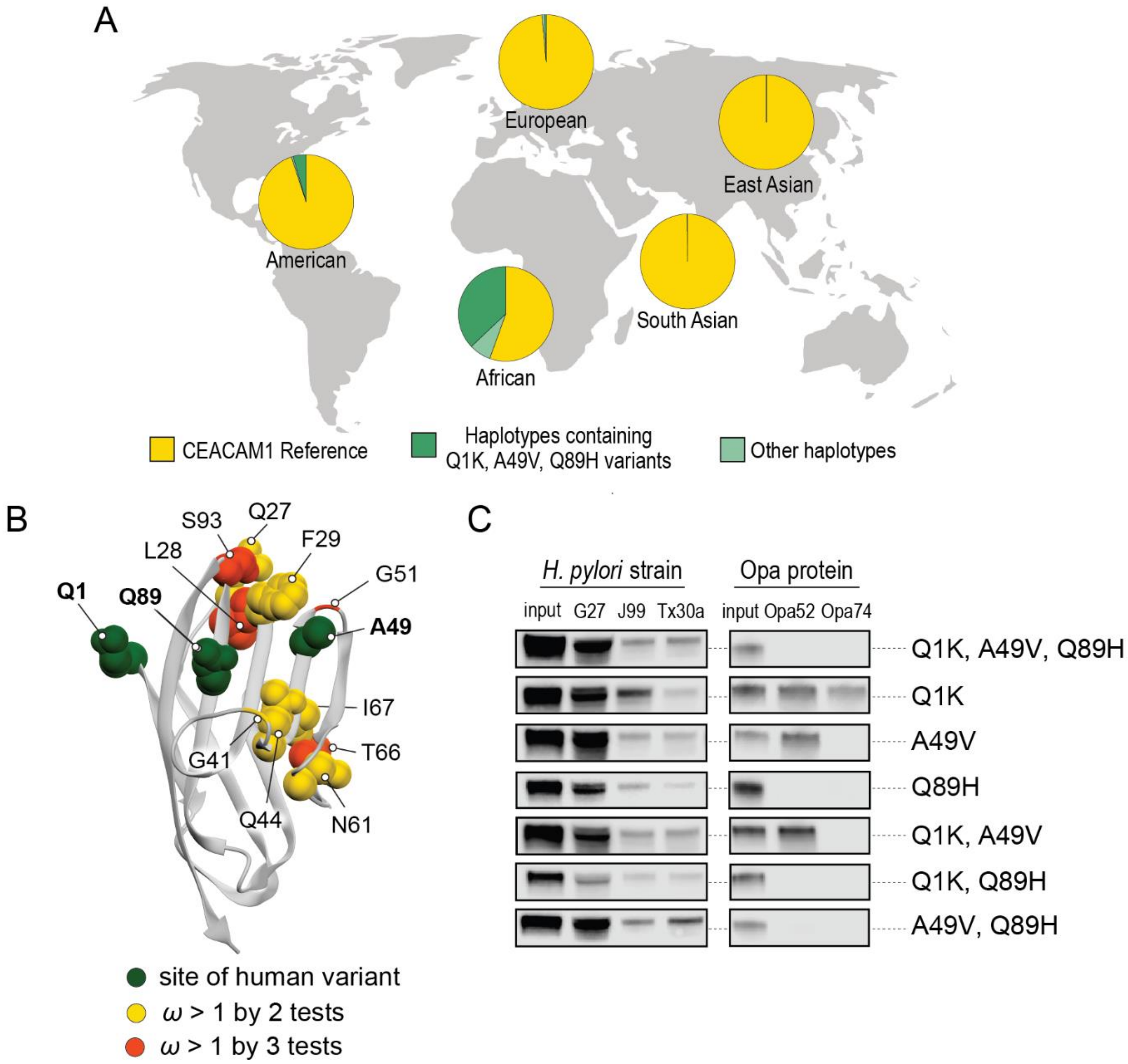


Figure 6. Abundant human CEACAM1 variants restrict pathogen binding. A) Frequency of haplotypes containing high frequency human variants Q1K, A49V and Q89H across human populations (map from BioRender.com). B) CEACAM1 crystal structure highlighting high frequency human variants and sites found to be evolving under positive selection across simian primates. C) Binding between combinations of high frequency human variants in the human CEACAM1 reference background and bacterial strains assayed by pulldown experiments and determined by western blotting. D) Sequence identity between the human CEACAM1-CEACAM3 hybrid allele and the human CEACAM1 and CEACAM3 reference alleles.

frequency SNPs (Fig. S10). Of the three variants, A49V and Q89H both lie within regions of CEACAM1 known to interact with bacterial adhesins suggesting they may alter bacterial adherence (Fig. 6B). To determine if these high-frequency CEACAM1 polymorphisms affect bacterial recognition, we generated recombinant CEACAM1 N-domain variant proteins for use in our adhesin binding assays. None of the variants are able to abolish CEACAM1 binding to our panel of *H. pylori* strains (Fig. 6C). In contrast, *Neisseria* Opa expressing strains exhibit highly variable recognition of multiple human CEACAM1 variants. The Q1K mutation alone has no impact on binding, while A49V abolishes recognition by Opa74, and variant Q89H abrogates binding to both Opa52 and Opa74 (Fig. 6C). Combinatorial CEACAM1 variants reveal that these mutations behave in a dominant manner, with Q89H dominant over A49V (Fig. 6C). Together these results demonstrate that high frequency human polymorphisms in CEACAM1 are sufficient to impair binding by specific classes of bacterial adhesins present in human pathogens. These findings further suggest that high-frequency CEACAM variants could alter human colonization or infection by pathogenic *Neisseria*, including causative agents of gonorrhea and meningitis.

DISCUSSION

Our investigation of species-specific bacterial adherence to CEACAM1 revealed an unforeseen example of extreme genetic divergence within the great apes. The bonobo CEACAM1 gene could represent a rapid succession of single residue changes combined with multiple recombination events arising in bonobos under strong selection and/or a population bottleneck. Alternatively, this allele may be ancient and have been subject to balancing selection or incomplete lineage sorting in ancestral hominid populations. We also considered that the source of the bonobo CEACAM1 sequence may not be from functional CEACAM genes, but a pseudogenized CEACAM sequence or a pregnancy-specific glycoprotein (PSG), a family of proteins closely related to CEACAMs. However, a BLAST search of relevant NCBI databases (see Methods) fails to identify any new genomic regions in bonobos or other primates with greater sequence identity than what had already been found. While there are multiple possible explanations for the highly divergent nature of bonobo CEACAM1, absent further evidence the origin of this particular allele remains obscure. What is clear from the example of bonobo CEACAM1, however, is the extent to which gene conversion can rapidly generate diversity between closely related species and the impact of such variation on interactions with microbes.

During the course of investigating the origin of the bonobo CEACAM1 sequence we discovered evidence that gene conversion has shaped the evolution of many CEACAMs across primates. While we identify several instances of likely gene conversion, results from phylogenetic analyses probably represent an underestimate of the true number of recombination events that have occurred among rapidly evolving CEACAMs in primates. Repeated episodes of gene conversion can obscure past instances of recombination and hinder their identification by gene-species tree discordance. In turn, GARD analyses and recombination detection programs in general tend to miss many recombination events (Bay and Bielawski, 2011; Kosakovsky Pond et al., 2006). One particularly interesting example in orangutans implicates multiple conversion events impacting CEACAM1, CEACAM5 and CEACAM8. (Fig. S8). Phylogenetic analyses indicate a species-specific conversion event between CEACAM1 and CEACAM5 in orangutans. Prior to the CEACAM1-CEACAM5 conversion, however, residues 29-64 in either CEACAM1 or CEACAM5 were likely replaced by the homologous sequence from CEACAM8. Evidence for this event includes not only the multiple nonsynonymous substitutions shared with orangutan CEACAM8, but a shared synonymous substitution in both orangutan CEACAM1 and CEACAM5 only observed in hominid CEACAM8 homologs. Despite this evidence, neither our phylogenetic analyses nor GARD analyses suggest CEACAM8 has been involved in gene conversion. Like for CEACAM7, the involvement of CEACAM8 in intra-paralog gene conversion is consistent with CEACAMs with rapid N-domain evolution participating in gene conversion events. Overall, the rapid shuffling of genetic variation among CEACAM genes that we observe could greatly augment the potential for host adaptation in the face of microbial antagonism.

It has been suggested that gene conversion between CEACAM paralogs works to preserve the ability of CEACAM3 to effectively mimic bacterially antagonized CEACAMs and thereby maintain its function as a decoy receptor (Zid and Drouin, 2013; Zimmermann, 2019). Indeed, our results and those of Adrian *et al.* (Adrian et al., 2019) support the importance of gene conversion for maintaining the similarity of CEACAM3 to other bacterial adhesin binding CEACAMs in apes and Old World monkeys. However, gene conversion cannot only serve to maintain CEACAM3's mimicry function. There is no evidence that New World monkeys encode a CEACAM3 homolog, yet within this group gene conversion appears to be frequent between

CEACAM1, CEACAM5, and CEACAM6 (Fig. 4C). Additionally, we observe multiple conversion events in hominids that do not involve CEACAM3 (Fig. 4C & Fig. S12).

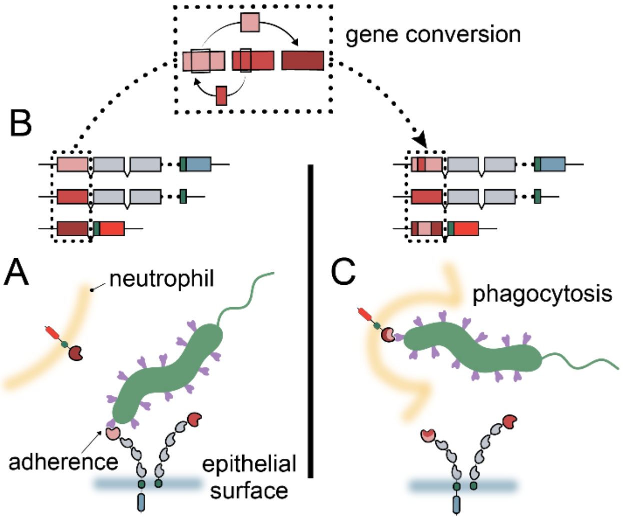


Figure 7. Model of CEACAM evolution in primates. A) Bacterial adhesins recognize a subset of epithelial CEACAM proteins and avoid binding with decoy CEACAM receptors present on neutrophils. B) Gene conversion facilitates the shuffling of regions of the CEACAM N-domain that alter binding to bacterial adhesins. C) Through gene conversion outlined in B, epithelial CEACAM proteins avoid binding by bacterial adhesins while the CEACAM decoy receptor gains binding triggering bacterial clearance through phagocytosis.

We propose that gene conversion among epithelial CEACAMs reflects a general mechanism of pathogen evasion (Fig. 7), allowing beneficial mutations to be exchanged among CEACAMs that interact with pathogens. Gene conversion allows beneficial sets of mutations to spread, whether between decoys and targets or among antagonized epithelial CEACAMs, more rapidly than conversion of residues through independent mutational events (Bittihn and Tsimring, 2017). In addition, in the case of decoys and targets the ability of sequences to be exchanged back-and-forth limits the residues available for differentiation by pathogens and provides decoys the ability to gain binding to CEACAM antagonists through exchanges from epithelial CEACAMs. Finally, the interchangeability of the protein binding domain among antagonized CEACAMs effectively provides multiple copies of the same domain, increasing the evolutionary space available to CEACAMs to evolve and test alleles, further enhancing the pace at which beneficial alleles may evolve. In this way gene conversion could provide an important mechanism by which the host can keep pace with rapidly evolving pathogenic microbes. It is also notable that the immense variation between Opa alleles, like CEACAMs, has been shown to involve a combination of rapid substitutions and recombination between extracellular loop domains that recognize host factors and serve as potential antigens for the host adaptive immune system. In this regard CEACAM-Opa interactions reflect an unusual evolutionary dynamic in which recombination likely plays a crucial role in reciprocal adaptation.

In addition to exploring the role of CEACAM gene conversion among primates, we provide evidence that this process continues to shape CEACAM diversity within human populations. The three human CEACAM1 variants we test in our adhesin binding assay are part of a group of related CEACAM1 haplotypes that increase sequence similarity to CEACAM3 and/or CEACAM5 (Figs. S9 & S10). Extended haplotypes that increase similarly to CEACAM1 at both synonymous and nonsynonymous positions in the N-domain are also found for CEACAM3 and CEACAM5 in humans (Figs. S11-S12). Indeed, haplotypes consisting of variants of putative recombination events are the most common non-reference alleles for CEACAM1, CEACAM3, and CEACAM5 (Fig. 6A, S10-12). Variant sites in these proteins tend to lie along the protein binding interface of the N-domain and often impact residues known to influence adhesin recognition. The relationships between these different CEACAM haplotypes appears to be complex, as many different combinations of partial variant haplotypes exist for each CEACAM paralog. The haplotype structures we observe suggest these CEACAM variants are the result of one or more recombination events between paralogous sequences, likely followed by further recombination with the major CEACAM allele.

Important questions remain regarding the rapid evolution of a subset of primate CEACAM proteins. Among these questions is why CEACAM7 and CEACAM8 show similar patterns of evolution to bacterially antagonized CEACAMs despite no known instances of bacterial antagonism. The simplest explanation is that CEACAM7 and CEACAM8 are themselves the targets of as yet unidentified pathogen antagonists (Sintsova et al., 2015). Alternatively, their rapid evolution may reflect pressure to maintain binding with rapidly evolving CEACAMs (Gray-Owen and Blumberg, 2006; Skubitz and Skubitz, 2008), could merely be a result of their genomic proximity to rapidly evolving CEACAMs prone to gene conversion (Zid and Drouin, 2013) or could be the result of some as yet unknown evolutionary pressures.

Another intriguing aspect of rapid CEACAM evolution is the impact rapid divergence might have outside of interactions with pathogenic microbes. Given the extensive overlap of CEACAM binding sites among unrelated bacterial adhesins, the ramifications of rapid CEACAM evolution likely extend beyond the adhesins of pathogens to those of commensal and beneficial microbes as well. For commensal microbes which rely on these interaction surfaces, pathogen-driven evolution could significantly alter their ability to colonize the host. The impact of CEACAM divergence on composition of the host microbiome and/or the evolution of commensal strains warrants further investigation.

Studies of other “housekeeping” proteins targeted by pathogens have found that sites under positive selection typically do not overlap with sites involved in essential host functions (Barber and Elde, 2014; Demogines et al., 2013). This is clearly not the case for CEACAMs, where we observed extensive overlap between sites involved in host protein interactions, sites targeted by bacterial adhesins and sites undergoing rapid evolution (Fig. 2B). How CEACAMs are able to rapidly evolve while maintaining their other essential host protein interactions remains a mystery. Future studies on CEACAM protein functions, interaction networks, and pathogen antagonism will likely clarify these outstanding questions regarding rapidly evolving CEACAMs.

Collectively our study provides evidence that repeated adaptation among primate CEACAMs has shaped host-specific cell adherence by diverse pathogenic bacteria. We find that over half of the CEACAM paralogs found in humans display signatures of positive selection across the primate lineage, localized primarily to the extracellular N-domain. We further discovered that rapid evolution of CEACAM N-domains has been facilitated by extensive “shuffling” of sequences between a subset of CEACAM paralogs through repeated gene conversion. The diversification of primate CEACAM N-domain sequences has likely had significant consequences for interactions between primates and bacteria. Consistent with observations across other primate species, we also provide evidence that gene conversion events impact bacterial pathogen recognition of CEACAMs in contemporary human populations. Together this work reveals how dynamic evolutionary processes have shaped bacterial-host associations with consequences for infectious disease susceptibility.

MATERIALS AND METHODS

Primate comparative genetics

Sequence identification

Orthologs for human CEACAM genes were identified through BLAST searches of primate reference genomes available through the NCBI BLAST webserver (Boratyn et al., 2013). Full length genomic regions for annotated human CEACAMs were used as query sequences. A full record of CEACAM orthologs identified and a partial record of BLAST results, including date accessed, query coverage and identity, as well as information on synteny, are listed in supplementary ExcelS1. Orthology was established by sequence identity, reciprocal best-BLAST hit, as well as intron structure and synteny. In total, we were able to extract 186 primate CEACAM sequences for analysis. We could not identify orthologs of every human CEACAM in every primate species, in some cases because of lineage specific gains and losses and in some cases likely because of incomplete genome assembly. As a result, the number of primate orthologs available for evolutionary analysis and phylogenetic reconstruction for each human CEACAM range from 11-19 (ExcelS1).

Sequence alignment & trimming

Orthologous protein coding sequences were extracted from CEACAM genes as follows. Multiple sequence alignments of the full-length gene were done using MAFFT alignment software as implemented in Geneious Prime 2020.2.2 with default settings. Alignments were manually corrected to correspond to human exon splice sites. Regions corresponding to human exons were then extracted, realigned, and minimally trimmed so all sequences were in-frame and orthologous codons aligned. So as not to exclude any protein coding regions from evolutionary analysis all human exons for a given CEACAM were concatenated and treated as a single protein coding sequence. Consequently, representations of CEACAM proteins in figures are not necessarily indicative of mature peptides, but rather represent all parts of the CEACAM protein that could potentially have been subject to positive selection. Gaps in alignments were removed for evolutionary analyses but were retained for tree building.

CEACAM3 exons

Almost all Old World monkey CEACAM3 genes analyzed had two extra exons annotated compared to humans. These exons are located between the exon encoding the N-domain and the transmembrane domain and are predicted by InterProScan (Quevillon et al., 2005), as implemented in Geneious Prime 2020.2.2, to encode the IgC-like domains typical of this region of CEACAM proteins. The majority of Old World monkeys have two exons annotated and all primates, including hominids, have strongly conserved sequences in this region, though hominids all encode premature stop codons. With the exception of the second IgC exon in colobus, these exons would allow for the translation of full length CEACAM proteins. While exon annotation differences between primate CEACAM genes is not unusual, the conservation of these sequences across primates containing a CEACAM3 gene, including in hominids where they are not annotated, was striking. To the best of our knowledge CEACAM3 transcripts for humans or other primates including either of these extra IgC domains have not been reported and indeed, the exon closest to the N-domain likely does not encode a functional protein in most hominids as a result of a premature stop codon. However, the strong conservation of these sequences across primates could indicate these exons encode functional protein segments in at least some species. For this reason, these exons and their orthologous sequences in hominids were included in downstream evolutionary analyses.

CEACAM5 trimming

The differences in number and likely arrangement of IgC domains in primate CEACAM5 orthologs prevented alignment of all full length CEACAM5 genes into a single multiple sequence alignment for extracting human orthologous protein coding sequences. Instead, sequences were first aligned in three groups; New World monkeys, leaf-eating monkeys (black-and-white colobus, black snub-nosed monkey and golden snub-nosed monkey), and the remaining Old World monkey sequences with the hominid sequences. There were enough similarities with human exons for orthologous exon sequences to be assigned and extracted for New World monkeys and the Old World monkey/Hominid group, but not for the leaf-eating monkeys group. For leaf-eating monkeys the predicted exons in common between species in this group were extracted. After extracting coding sequences for each group individually the extracted sequences were then aligned in a single multiple sequence alignment. However, the large gaps caused by missing IgC sequences relative to human CEACAM5 posed a problem for evolutionary analyses which require gaps to be removed from sequences prior to analysis. We were concerned that choices made regarding which sequences were removed would unduly influence the results of evolutionary analyses or result in lower coverage of the evolutionary history of the entire coding sequence. To account for this, three strategies of trimming alignment gaps were carried out and the results of each used in separate evolutionary analyses. For the first strategy every species whose sequence contained gaps corresponding to missing IgC domains was removed. These species were black-and-white colobus, black snub-nosed monkey, golden snub-nosed monkey, drill, sooty mangabey, and common marmoset. This resulted in the longest sequence for analysis (2 Kb) including 6 predicted IgC domains, but the smallest number of species represented (12). In the second strategy primate sequences with gaps corresponding to the largest number of missing IgC domains (four) were removed, while those with only two missing domains were retained, and the alignment region containing the sequence gap caused by the missing domains removed, giving a smaller alignment (1.4 Kb, with four IgC domains), but more species (16). For this strategy sooty mangabey, and common marmoset were removed from the analysis. For the third strategy all species for which complete CEACAM5 gene sequences could be identified were retained and all gaps corresponding to missing IgC domains removed. This gave the smallest sequence (0.9 Kb, retaining two IgC domains), but provided the largest number of represented species (18). Evolutionary analyses for these strategies are included in Fig S1, TableS1 and Table S2.

Alignment comparison between MAFFT and MUSCLE

To confirm that our alignment method was not biasing the assignment of orthology of coding sequences to human exons, we compared the results of alignments of extracted exons using MAFFT (Katoh and Standley, 2013) and the alternative program MUSCLE (Edgar, 2004), both as implemented in Geneious Prime 2020.2.2. With the exceptions of CEACAM7 and CEACAM5 there were no drastic changes between alignments performed using MAFFT and those done using MUSCLE. Upon inspection the discrepancy between MAFFT and MUSCLE alignments for CEACAM7 could be attributed to an approximately 7 Kb insertion in the orangutan CEACAM7 gene relative to all other primates. Upon removing this insertion alignments with both MAFFT and MUSCLE were in agreement. Discrepancies between alignments of

CEACAM5 with MAFFT and MUSCLE were due to differences in how the programs aligned sequences corresponding to IgC domains, likely as a result of differences in the number and possibly the arrangement of sequences coding for IgC domains between primates. MAFFT and MUSCLE alignments were carried out for each of the three different trimmed versions of CEACAM5 (see above) and each set of sequence alignments was tested using each of the evolutionary analysis methods. All other evolutionary analyses were carried out using sequences trimmed according to MAFFT alignments.

The results of CEACAM5 evolutionary analyses were largely similar regardless of which alignment or trimming method was employed, identifying similar patterns of selection (sites under selection concentrated in the N-domain) and many of the same sites under selection. Results presented in the paper are for dataset 1 (ds1) which contains the largest number of domains and using the MAFFT alignment to match the method used for other CEACAM analyses presented. Results for alternative CEACAM5 trimming and alignment methods are included in Fig S1, TableS1 and Table S2.

Bonobo CEACAM1 N-domain sequence verification

Bonobo genomic DNA was not available for direct sequencing of CEACAM1, so currently available bonobo genome sequence data was used for sequence verification. While the genome assembly from which bonobo CEACAM sequences were identified for evolutionary analyses did not have reads available, a more recent assembly of a different bonobo individual became available during the course of this study which did deposit sequencing reads along with a *de novo* genome assembly (Mao et al., 2021). The CEACAM1 genomic region of the newer assembly was 99% identical to the older version while the coding sequences differ at only a single nucleotide outside of the N-domain. Furthermore, examining the reads used to assemble the newer genome we confirmed that multiple reads covered the length of the bonobo CEACAM1 N-domain and included the highly diverged nucleotides of the binding region in contiguous reads. Additionally, we examined CEACAM1 sequences for the thirteen bonobo individuals sequenced as part of the Great Apes Genome Project (Prado-Martinez et al., 2013). Genomes for these individuals were constructed using a reference based assembly method to the human genome. The assembled sequences largely supported the highly diverged N-domain seen in the reference genome; however there was a 31 bp region that was identical to the human CEACAM1 sequence rather than the two *de novo* bonobo sequences. Examining reads from these individuals failed to support human sequences at this position and in fact supported the more divergent sequence seen in the bonobo *de novo* assemblies. Nucleotide BLAST searches on the NCBI webserver for bonobo N-domain sequences were performed with query sequences searching against the RefSeq Genome Database (refseq_genomes) for the organism groups "Homo/Pan/Gorilla groups" (taxid:207598) and "Primates" (taxid: 9443), while excluding "bonobos" (taxid:9597).

Identification of human CEACAM N-domain variation

Human haplotype data for CEACAM1, CEACAM3, CEACAM5, CEACAM6, CEACAM7, and CEACAM8, available through the International Genome Sample Resource (https://www.internationalgenome.org/) was accessed through the Ensemble genome browser (https://www.ensembl.org/). For each CEACAM the haplotypes identified for the Matched Annotation from NCBI and EMBL-EBI (MANE) Select v0.92 transcript were used. All coding sequence haplotypes for the MANE Select transcript were downloaded and analyzed in excel as well as in R using custom scripts.

Phylogenetic analyses

PAML/FUBAR/MEME/GARD

Evolutionary analyses were performed individually for each group of human CEACAM coding sequence orthologs. Only CEACAM21 was excluded from evolutionary analyses, since it was found only in hominid genomes and has likely been lost in the pan lineage (ExcelS1) resulting in only three closely related sequences being available for comparison, insufficient for robust phylogenetic based evolutionary analysis. CEACAM21 sequences were included in subsequent phylogenetic reconstructions.

CEACAM coding sequences were tested for evidence of positive selection using the PAML NS sites program under the codon model F3x4 (Yang, 2007). To determine the likelihood a gene was evolving under positive selection, log-likelihood tests were performed comparing the models of selection M1&M2 as well as M7&M8 (Table S1). Sites evolving under positive selection were identified by PAML using the Bayes

Empirical Bayes analysis as implemented in the NS sites package for evolutionary Model 2, which has been shown to be more robust to error due to recombination than the alternative, Model 8, when identifying sites under selection (Anisimova et al., 2003). In addition, sites under selection were identified (Table S2) using the HyPhy package programs FUBAR and MEME (Murrell et al., 2013, 2012) as implemented on the Datamonkey web servers (www.datamonkey.org and classic.datamonkey.org respectively) (Delport et al., 2010; Kosakovsky Pond and Frost, 2005; Pond et al., 2005; Weaver et al., 2018). For FUBAR and initial MEME analyses, species trees of the relevant primates were provided to inform analyses of evolution. HyPhy GARD analyses (classic.datamonkey.org) were used to identify evidence of recombination and the number and approximate locations of breakpoints (Kosakovsky Pond et al., 2006). When GARD detected evidence of recombination, updated GARD informed phylogenies were used for MEME analyses to account for errors in calling sites under selection due to recombination. Prior to running MEME and GARD analyses the "automatic model selection tool" provided by classic.datamonkey.org was used to determine the most appropriate model of selection under which to run analyses. For PAML, sites with posterior probability >0.95 were considered to have strong support to be evolving under positive selection (Yang et al., 2005), while >0.9 posterior probability supported sites found by FUBAR (Murrell et al., 2013) and p-values ≤0.05 supported sites found by MEME (Murrell et al., 2012).

Tree building

Phylogenetic trees were constructed using our panel of primate CEACAM coding sequences identified as described above. Multiple sequence alignments on which tree constructions were based were done by translation alignment using default settings of the MAFFT sequence alignment software as implemented in Geneious Prime 2020.2.2. For domain specific phylogenetic reconstruction domains were identified using InterProScan (Quevillon et al., 2005) in Geneious Prime. Assignments for immunoglobulin-like domains, that is the IgV-like (N-domain) and IgC domains were based on predictions by the Superfamily database (Wilson et al., 2009) and cytoplasmic domain assignments were based on the PHOBIUS database (Käll et al., 2004). Transmembrane domains were excluded from analyses due to their particularly small sequence length, which can make tree building unreliable due to limited phylogenetically informative sites. Indeed, relatively short sequence lengths for the other domains, typically around 300 bps or less, along with often high sequence similarity likely decreased the reliability and statistical support for our domain trees. However, even with these limitations in many cases relationships between domains were resolved with high bootstrap support, particularly for peripheral nodes and clades and for CEACAMs not found to be evolving rapidly. Phylogenetic reconstructions were done using the PhyML 3.0 web browser (http://www.atgc-montpellier.fr/phyml/) with default settings and confidence testing by 1000 bootstrap replicates (Guindon et al., 2010).

Data visualization

Visualization of evolutionary analyses, phylogenetic trees, sequence identity, and haplotype frequencies was done in R (R Core Team, 2019) using the R packages BiocManager (Morgan, 2019), treeio (Wang et al., 2019), ggplot2 (Wickham, 2016), ggtree (Yu et al., 2018, 2017), evobiR (Blackmon and Adams, 2015), and ggforce (Pedersen, 2021). Protein structures were visualized using the UCSF Chimera package version 1.13.1. Chimera is developed by the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco (supported by NIGMS P41-GM103311) (Pettersen et al., 2004).

CEACAM1 Binding Assays

Recombinant CEACAM1 expression plasmid construction

Plasmids encoding primate CEACAM1 N-domains were constructed by assembly PCR and ligation independent cloning (LIC) into the pcDNA3 GFP LIC vector (6D) (a gift from Scott Gradia; Addgene plasmid #30127). A detailed description of the assembly PCRs is provided in the Supplementary Methods and the DNA oligomers and templates are described in Supplementary Tables S4 and S5. Briefly, oligonucleotides were designed to assemble expression cassettes containing the human IgK signal sequence followed by a primate CEACAM1 N-terminal domain, and finally a STREPII tag and Tobacco Etch Virus (TEV) protease site. LIC cloning of the primate CEACAM1 N-terminal expression cassettes into pcDNA2 GFP LIC (6D) was performed following the protocol provided by the California Institute for Quantitative Biosciences at Berkeley (https://qb3.berkeley.edu/facility/qb3-macrolab/projects/lic-cloning-protocol/). Mutations to introduce bonobo CEACAM1 residues and population variants into the human CEACAM1 reference as well human CEACAM1 residues into the bonobo CEACAM1 reference sequence were done by site directed mutagenesis using mutation specific primers designed using the Agilent QuikChange Primer Design tool (https://www.chem.agilent.com/store/primerDesignProgram.jsp), then transformed into One Shot™ Top10 chemically competent cells for amplification and sequence verification. Plasmids were extracted for further use using the ZymoPURE™ II Plasmid Maxiprep kit.

Recombinant CEACAM1 expression plasmids were transfected into Human HEK293T cells using the Lipofectamine™ 3000 transfection kit following manufacturers instructions. Two days post transfection cell supernatant was collected and filter sterilized and cells were collected and lysed. Expression of proteins was confirmed by western blotting (detailed below).

Bacterial strains & culture

H. pylori strains G27 (Baltrus et al., 2009) J99 (Alm et al., 1999), Tx30a (ATCC® 51932), and the G27 HopQ deletion strain (*omp27::cat-sacB* in NSH57) (Yang et al., 2019) were cultured microaerobically at 37°C on Columbia agar plates supplemented with 5% horse blood 0.2% beta cyclodextrin, 0.01% amphotericin B, and 0.02% vancomycin. To assay binding between recombinant primate CEACAM1 N-domain proteins and *H. pylori* strains, *H. pylori* strains were grown for two to five days on solid media, collected and suspended in Brain Heart Infusion Media. 500uL of bacterial suspension were then incubated with 100uL of CEACAM protein for thirty minutes, rotating on a nutator. Bacteria were then washed twice with cold PBS. Samples to be visualized by western blotting were pelleted and resuspended in 1x Laemmli Buffer. Samples to be examined by flow cytometry were suspended in 0.5-1 mL of PBS.

The use of *Escherichia coli* to express MS11 and VP1 *Neisseria gonorrhoeae* Opa proteins was described previously (Roth et al., 2013). For this project plasmids expressing Opa proteins, Opa52 (Kupsch et al., 1993) and Opa74 (Roth et al., 2013), were synthesized in the pET-28a vector background by GeneScript. Synthesizing Opa expression plasmids bypassed the subcloning described in previous works that allowed outer membrane expression, so an N-terminal signal sequence from the OMP A protein, native to the outer membrane of *E. coli,* was added by the manufacturer to express Opa proteins on the outer membrane of *E. coli*. NcoI and HindIII restriction sites were used to add OMP A and Opa sequences to the pET-28a plasmid. Opa expression vectors were transformed into *E. coli* DH5α cells for maintenance, replication and sequence verification. Plasmids were extracted for further use using the Zymo Research Zyppy™ Plasmid miniprep kit. For pulldown experiments Opa expression plasmids were transformed into BL21(DE3) *E. coli* cells to allow for inducible expression of Opa proteins. Cells were grown to an optical density of OD_{600} 0.4-0.6, then IPTG (Isopropyl β- d-1-thiogalactopyranoside) was added to a concentration of 100mM to induce expression of Opa proteins. Bacterial cells were left to induce for three hours at 37°C. For pulldown assays 300μL of induced *E. coli* cell culture was incubated with 100μL of CEACAM1 protein construct as processed as described for *H. pylori*. All *E. coli* cells were cultured at 37°C in LB (Luria-Bertani) broth.

Western blotting and flow cytometry

Pulldown experiments assayed by western blotting were visualized using a commercially available mixture of Mouse α-GFP clones 7.1 and 13.1 (Sigma-Aldrich) for the primary antibody incubation followed by secondary incubation with goat α-mouse conjugated to horseradish peroxidase (Jackson ImmunoResearch) and visualized by WesternBright™ ECL HRP Substrate (Thomas Scientific). For pulldowns visualized by western blotting CEACAM1 protein input samples were prepared by mixing 20uL of protein with 20uL 2x Laemmli then boiled at 95°C for five minutes and centrifuged at max speed for five minutes, before visualization by western blotting along side pulldown samples. GFP fluorescence of primate CEACAM1 constructs bound to *H. pylori* strain G27 was also measured by flow cytometry, with 10,000 events per sample measured. Flow cytometry data was analyzed using FlowJo v10.5.3.

AUTHOR CONTRIBUTIONS

E.P.B. and M.F.B. conceived the study with input from K.M.K. Phylogenetic analyses were performed by E.P.B. with assistance from K.M.K. CEACAM-adhesin binding experiments were performed by E.P.B. with assistance from R.S., E.B., C.K.G., and W.B. Original manuscript was prepared by E.P.B. with input from M.F.B. All authors contributed to editing of the original manuscript.

ACKNOWLEDGMENTS

We are grateful to members of the Barber lab, especially Caitlin Kowalski, for helpful discussions and feedback on the manuscript. This work was supported by National Institutes of Health grants R35GM133652 (to M.F.B.) and F32AI147565 (to E.P.B). We thank Karen Guillemin for strains and assistance with *H. pylori* as well as Nina Salama for sharing *H. pylori* strains. We thank Andrew Kern and CJ Battey for helpful discussions and assistance with population genetic analyses. We thank Nels Elde for thoughtful discussions related to this work. We thank Jacob Laser for early assistance with CEACAM phylogenetics. We also thank Scott Gradia for the gift of vector pcDNA3 GFP LIC.

COMPETING INTERESTS STATEMENT

The authors declare no competing interests.

REFERENCES

Adrian J, Bonsignore P, Hammer S, Frickey T, Hauck CR. 2019. Adaptation to Host-Specific Bacterial Pathogens Drives Rapid Evolution of a Human Innate Immune Receptor. *Curr Biol* **29**:616-630.e5. doi:10.1016/j.cub.2019.01.058

Aleru O, Barber MF. 2020. Battlefronts of evolutionary conflict between bacteria and animal hosts. *PLoS Pathog* **16**:e1008797. doi:10.1371/journal.ppat.1008797

Alm RA, Ling LS, Moir DT, King BL, Brown ED, Doig PC, Smith DR, Noonan B, Guild BC, deJonge BL, Carmel G, Tummino PJ, Caruso A, Uria-Nickelsen M, Mills DM, Ives C, Gibson R, Merberg D, Mills SD, Jiang Q, Taylor DE, Vovis GF, Trust TJ. 1999. Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen Helicobacter pylori. *Nature* **397**:176–180. doi:10.1038/16495

Anisimova M, Nielsen R, Yang Z. 2003. Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. *Genetics* **164**:1229–1236. doi:10.1017/CBO9780511808999

Baltrus DA, Amieva MR, Covacci A, Lowe TM, Merrell DS, Ottemann KM, Stein M, Salama NR, Guillemin K. 2009. The complete genome sequence of Helicobacter pylori strain G27. *J Bacteriol* **191**:447–448. doi:10.1128/JB.01416-08

Barber MF, Elde NC. 2014. Escape from bacterial iron piracy through rapid evolution of transferrin. *Science* **346**:1362–1366. doi:10.1126/science.1259329

Bay RA, Bielawski JP. 2011. Recombination detection under evolutionary scenarios relevant to functional divergence. *J Mol Evol* **73**:273–286. doi:10.1007/s00239-011-9473-0

Bittihn P, Tsimring LS. 2017. Gene Conversion Facilitates Adaptive Evolution on Rugged Fitness Landscapes. *Genetics* **207**:1577–1589. doi:10.1534/genetics.117.300350

Blackmon H, Adams RH. 2015. evobiR: Comparative and population genetic analyses.

Bonsor DA, Zhao Q, Schmidinger B, Weiss E, Wang J, Deredge D, Beadenkopf R, Dow B, Fischer W, Beckett D, Wintrode PL, Haas R, Sundberg EJ. 2018. The Helicobacter pylori adhesin protein HopQ exploits the dimer interface of human CEACAMs to facilitate translocation of the oncoprotein CagA. *EMBO J* **37**:e98664. doi:10.15252/embj.201798664

Boratyn GM, Camacho C, Cooper PS, Coulouris G, Fong A, Ma N, Madden TL, Matten WT, McGinnis SD, Merezhuk Y, Raytselis Y, Sayers EW, Tao T, Ye J, Zaretskaya I. 2013. BLAST: a more efficient report with usability improvements. *Nucleic Acids Res* **41**:W29-33. doi:10.1093/nar/gkt282

Brewer ML, Dymock D, Brady RL, Singer BB, Virji M, Hill DJ. 2019. Fusobacterium spp. target human CEACAM1 via the trimeric autotransporter adhesin CbpF. *J Oral Microbiol* **11**:1–16. doi:10.1080/20002297.2018.1565043

Brockhurst MA, Chapman T, King KC, Mank JE, Paterson S, Hurst GDD. 2014. Running with the Red Queen: the role of biotic conflicts in evolution. *Proc Biol Sci* **281**. doi:10.1098/rspb.2014.1382

Brown RL, Clarke TB. 2017. The regulation of host defences to infection by the microbiota. *Immunology* **150**:1–6. doi:10.1111/imm.12634

Cao P, Cover TL. 2002. Two different families of hopQ alleles in Helicobacter pylori. *J Clin Microbiol* **40**:4504–4511. doi:10.1128/jcm.40.12.4504-4511.2002

Chen J-M, Cooper DN, Chuzhanova N, Férec C, Patrinos GP. 2007. Gene conversion: mechanisms, evolution and human disease. *Nat Rev Genet* **8**:762–775. doi:10.1038/nrg2193

Daugherty MD, Zanders SE. 2019. Gene conversion generates evolutionary novelty that fuels genetic conflicts. *Current Opinion in Genetics & Development* **58–59**:49–54. doi:10.1016/j.gde.2019.07.011

Delport W, Poon AFY, Frost SDW, Kosakovsky Pond SL. 2010. Datamonkey 2010: a suite of phylogenetic analysis tools for evolutionary biology. *Bioinformatics* **26**:2455–2457. doi:10.1093/bioinformatics/btq429

Demogines A, Abraham J, Choe H, Farzan M, Sawyer SL. 2013. Dual Host-Virus Arms Races Shape an Essential Housekeeping Protein. *PLoS Biol* **11**:1001571. doi:10.1371/journal.pbio.1001571

Dje N'Guessan P, Vigelahn M, Bachmann S, Zabel S, Opitz B, Schmeck B, Hippenstiel S, Zweigner J, Riesbeck K, Singer BB, Suttorp N, Slevogt H. 2007. The UspA1 Protein of Moraxella catarrhalis Induces CEACAM-1–Dependent Apoptosis in Alveolar Epithelial Cells. *J Infect Dis* **195**:1651–1660. doi:10.1086/514820

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**:1792–1797. doi:10.1093/nar/gkh340

Fox DA, Larsson P, Lo RH, Kroncke BM, Kasson PM, Columbus L. 2014. Structure of the Neisserial outer membrane protein Opa$_{60}$: loop flexibility essential to receptor recognition and bacterial engulfment. *J Am Chem Soc* **136**:9938–9946. doi:10.1021/ja503093y

Gibbs R a., Rogers J, Katze MG, Bumgarner R, Weinstock GM, Mardis ER, Remington K a., Strausberg RL, Venter JC, Wilson RK, Batzer M a., Bustamante CD, Eichler EE, Hahn MW, Hardison RC, Makova KD, Miller W, Milosavljevic A, Palermo RE, Siepel A, Sikela JM, Attaway T, Bell S, Bernard KE, Buhay CJ, Chandrabose MN, Dao M, Davis C, Delehaunty KD, Ding Y, Dinh HH, Dugan-Rocha S, Fulton L a., Gabisi RA, Garner TT, Godfrey J, Hawes AC, Hernandez J, Hines S, Holder M, Hume J, Jhangiani SN, Joshi V, Khan ZM, Kirkness EF, Cree A, Fowler RG, Lee S, Lewis LR, Li Z, Liu Y-S, Moore SM, Muzny D, Nazareth LV, Ngo DN, Okwuonu GO, Pai G, Parker D, Paul H a., Pfannkoch C, Pohl CS, Rogers Y-H, Ruiz SJ, Sabo A, Santibanez J, Schneider BW, Smith SM, Sodergren E, Svatek AF, Utterback TR, Vattathil S, Warren W, White CS, Chinwalla AT, Feng Y, Halpern AL, Hillier LW, Huang X, Minx P, Nelson JO, Pepin KH, Qin X, Sutton GG, Venter E, Walenz BP, Wallis JW, Worley KC, Yang S-P, Jones SM, Marra M a., Rocchi M, Schein JE, Baertsch R, Clarke L, Csürös M, Glasscock J, Harris RA, Havlak P, Jackson AR, Jiang H, Liu Y, Messina DN, Shen Y, Song HX-Z, Wylie T, Zhang L, Birney E, Han K, Konkel MK, Lee J, Smit AF a., Ullmer B, Wang H, Xing J, Burhans R, Cheng Z, Karro JE, Ma J, Raney B, She X, Cox MJ, Demuth JP, Dumas LJ, Han S-G, Hopkins J, Karimpour-Fard A, Kim YH, Pollack JR, Vinar T, Addo-Quaye C, Degenhardt J, Denby A, Hubisz MJ, Indap A, Kosiol C, Lahn BT, Lawson H a., Marklein A, Nielsen R, Vallender EJ, Clark AG, Ferguson B, Hernandez RD, Hirani K, Kehrer-Sawatzki H, Kolb J, Patil S, Pu L-L, Ren Y, Smith DG, Wheeler D a., Schenck I, Ball EV, Chen R, Cooper DN, Giardine B, Hsu F, Kent WJ, Lesk A, Nelson DL, O'brien WE, Prüfer K, Stenson PD, Wallace JC, Ke H, Liu X-M, Wang P, Xiang AP, Yang F, Barber GP, Haussler D, Karolchik D, Kern AD, Kuhn RM, Smith KE, Zwieg AS. 2007. Evolutionary and biomedical insights from the rhesus macaque genome. *Science* **316**:222–234. doi:10.1126/science.1139247

Gray-Owen SD, Blumberg RS. 2006. CEACAM1: Contact-dependent control of immunity. *Nat Rev Immunol* **6**:433–446. doi:10.1038/nri1864

Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol* **59**:307–321. doi:10.1093/sysbio/syq010

Gur C, Maalouf N, Gerhard M, Singer BB, Emgård J, Temper V, Neuman T, Mandelboim O, Bachrach G. 2019a. The Helicobacter pylori HopQ outermembrane protein inhibits immune cell activities. *Oncoimmunology*. doi:10.1080/2162402X.2018.1553487

Gur C, Maalouf N, Shhadeh A, Berhani O, Singer BB, Bachrach G, Mandelboim O. 2019b. Fusobacterium nucleatum supresses anti-tumor immunity by activating CEACAM1. *Oncoimmunology* **8**:1–6. doi:10.1080/2162402X.2019.1581531

Hamilton WD, Axelrod R, Tanese R. 1990. Sexual reproduction as an adaptation to resist parasites (a review). *Proc Natl Acad Sci U S A* **87**:3566–3573. doi:10.1073/pnas.87.9.3566

Javaheri A, Kruse T, Moonens K, Mejías-luque R, Debraekeleer A, Asche CI, Tegtmeyer N, Kalali B, Bach NC, Sieber SA, Hill DJ, Königer V, Hauck CR, Moskalenko R, Haas R, Busch DH, Klaile E, Slevogt H, Schmidt A, Backert S, Remaut H, Singer BB, Gerhard M. 2016. Helicobacter pylori adhesin HopQ engages in a virulence-enhancing interaction with human CEACAMs. *Nature Microbiology* **17**:16189. doi:10.1038/nmicrobiol.2016.189

Johswich KO, McCaw SE, Islam E, Sintsova A, Gu A, Shively JE, Gray-Owen SD. 2013. In Vivo Adaptation and Persistence of Neisseria meningitidis within the Nasopharyngeal Mucosa. *PLoS Pathog* **9**. doi:10.1371/journal.ppat.1003509

Käll L, Krogh A, Sonnhammer ELL. 2004. A combined transmembrane topology and signal peptide prediction method. *J Mol Biol* **338**:1027–1036. doi:10.1016/j.jmb.2004.03.016

Kammerer R, Zimmermann W. 2010. Coevolution of activating and inhibitory receptors within mammalian carcinoembryonic antigen families. *BMC Biol* **8**. doi:10.1186/1741-7007-8-12

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* **30**:772–780. doi:10.1093/molbev/mst010

Kline KA, Fälker S, Dahlberg S, Normark S, Henriques-Normark B. 2009. Bacterial adhesins in host-microbe interactions. *Cell Host Microbe* **5**:580–592. doi:10.1016/j.chom.2009.05.011

Königer V, Holsten L, Harrison U, Busch B, Loell E, Zhao Q, Bonsor DA, Roth A, Kengmo-Tchoupa A, Smith SI, Mueller S, Sundberg EJ, Zimmermann W, Fischer W, Hauck CR, Haas R. 2016. Helicobacter pylori exploits human CEACAMs via HopQ for adherence and translocation of CagA. *Nature Microbiology* **2**. doi:10.1038/nmicrobiol.2016.188

Kosakovsky Pond SL, Frost SDW. 2005. Not so different after all: A comparison of methods for detecting amino acid sites under selection. *Mol Biol Evol* **22**:1208–1222. doi:10.1093/molbev/msi105

Kosakovsky Pond SL, Posada D, Gravenor MB, Woelk CH, Frost SDW. 2006. Automated phylogenetic detection of recombination using a genetic algorithm. *Mol Biol Evol* **23**:1891–1901. doi:10.1093/molbev/msl051

Kuespert K, Pils S, Hauck CR. 2006. CEACAMs : their role in physiology and pathophysiology. *Curr Opin Cell Biol* **18**:565–571. doi:10.1016/j.ceb.2006.08.008

Kuespert K, Weibel S, Hauck CR. 2007. Profiling of bacterial adhesin — host receptor recognition by soluble immunoglobulin superfamily domains. *J Microbiol Methods* **68**:478–485. doi:10.1016/j.mimet.2006.10.003

Kupsch EM, Knepper B, Kuroki T, Heuer I, Meyer TF. 1993. Variable opacity (Opa) outer membrane proteins account for the cell tropisms displayed by Neisseria gonorrhoeae for human leukocytes and epithelial cells. *EMBO J* **12**:641–650. doi:10.1002/j.1460-2075.1993.tb05697.x

Mao Y, Catacchio CR, Hillier LW, Porubsky D, Li R, Sulovari A, Fernandes JD, Montinaro F, Gordon DS, Storer JM, Haukness M, Fiddes IT, Murali SC, Dishuck PC, Hsieh P, Harvey WT, Audano PA, Mercuri L, Piccolo I, Antonacci F, Munson KM, Lewis AP, Baker C, Underwood JG, Hoekzema K, Huang T-H, Sorensen M, Walker JA, Hoffman J, Thibaud-Nissen F, Salama SR, Pang AWC, Lee J, Hastie AR, Paten B, Batzer MA, Diekhans M, Ventura M, Eichler EE. 2021. A high-quality bonobo genome refines the analysis of hominid evolution. *Nature* **594**:77–81. doi:10.1038/s41586-021-03519-x

Markel G, Gruda R, Achdout H, Katz G, Nechama M, Blumberg RS, Kammerer R, Zimmermann W, Mandelboim O. 2004. The Critical Role of Residues 43R and 44Q of Carcinoembryonic Antigen Cell Adhesion Molecules-1 in the Protection from Killing by Human NK Cells. *The Journal of Immunology* **173**:3732–3739. doi:10.4049/jimmunol.173.6.3732

Moonens K, Hamway Y, Neddermann M, Reschke M, Tegtmeyer N, Kruse T, Kammerer R, Mejías-Luque R, Singer BB, Backert S, Gerhard M, Remaut H. 2018. Helicobacter pylori adhesin HopQ disrupts trans dimerization in human CEACAMs. *EMBO J* **37**:e98665. doi:10.15252/embj.201798665

Morgan M. 2019. BiocManager: Access the Bioconductor Project Package Repository.

Muenzner P, Bachmann V, Zimmermann W, Hentschel J, Hauck CR. 2010. Human-Restricted Bacterial Pathogens Block Shedding of Epithelial Cells by Stimulating Integrin Activation. *Science* **357**:90. doi:10.1126/science.1185231

Muenzner P, Kengmo Tchoupa A, Klauser B, Brunner T, Putze J, Dobrindt U, Hauck CR, Blanke SR. 2016. Uropathogenic E. coli Exploit CEA to Promote Colonization of the Urogenital Tract Mucosa. doi:10.1371/journal.ppat.1005608

Murrell B, Moola S, Mabona A, Weighill T, Sheward D, Kosakovsky Pond SL, Scheffler K. 2013. FUBAR: A fast, unconstrained bayesian AppRoximation for inferring selection. *Mol Biol Evol* **30**:1196–1205. doi:10.1093/molbev/mst030

Murrell B, Wertheim JO, Moola S, Weighill T, Scheffler K, Kosakovsky Pond SL. 2012. Detecting individual sites subject to episodic diversifying selection. *PLoS Genet* **8**. doi:10.1371/journal.pgen.1002764

Pavlopoulou A, Scorilas A. 2014. A comprehensive phylogenetic and structural analysis of the carcinoembryonic antigen (CEA) gene family. *Genome Biol Evol* **6**:1314–1326. doi:10.1093/gbe/evu103

Pedersen TL. 2021. ggforce: Accelerating “ggplot2.”

Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE. 2004. UCSF Chimera--a visualization system for exploratory research and analysis. *J Comput Chem* **25**:1605–1612. doi:10.1002/jcc.20084

Pond SLK, Frost SDW, Muse SV. 2005. HyPhy: hypothesis testing using phylogenies. *Bioinformatics* **21**:676–679. doi:10.1093/bioinformatics/bti079

Prado-Martinez J, Sudmant PH, Kidd JM, Li H, Kelley JL, Lorente-Galdos B, Veeramah KR, Woerner AE, O’Connor TD, Santpere G, Cagan A, Theunert C, Casals F, Laayouni H, Munch K, Hobolth A, Halager AE, Malig M, Hernandez-Rodriguez J, Hernando-Herraez I, Prüfer K, Pybus M, Johnstone L, Lachmann M, Alkan C, Twigg D, Petit N, Baker C, Hormozdiari F, Fernandez-Callejo M, Dabad M, Wilson ML, Stevison L, Camprub C, Carvalho T, Ruiz-Herrera A, Vives L, Mele M, Abello T, Kondova I, Bontrop RE, Pusey A, Lankester F, Kiyang JA, Bergl RA, Lonsdorf E, Myers S, Ventura M, Gagneux P, Comas D, Siegismund H, Blanc J, Agueda-Calpena L, Gut M, Fulton L, Tishkoff SA, Mullikin JC, Wilson RK, Gut IG, Gonder MK, Ryder OA, Hahn BH, Navarro A, Akey JM, Bertranpetit J, Reich D, Mailund T, Schierup MH, Hvilsom C, Andrés AM, Wall JD, Bustamante CD, Hammer MF, Eichler EE, Marques-Bonet T. 2013. Great ape genetic diversity and population history. *Nature* **499**:471–475. doi:10.1038/nature12228

Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, Lopez R. 2005. InterProScan: protein domains identifier. *Nucleic Acids Res* **33**:W116-20. doi:10.1093/nar/gki442

R Core Team. 2019. R: A Language and Environment for Statistical Computing.

Roth A, Mattheis C, Muenzner P, Unemo M, Hauck CR. 2013. Innate recognition by neutrophil granulocytes differs between Neisseria gonorrhoeae strains causing local or disseminating infections. *Infect Immun* **81**:2358–2370. doi:10.1128/IAI.00128-13

Sadarangani M, Pollard AJ, Gray-Owen SD. 2011. Opa proteins and CEACAMs: pathways of immune engagement for pathogenic Neisseria. *FEMS Microbiology Reviews* **35**:498–514. doi:10.1111/j.1574-6976.2010.00260.x

Sintsova A, Wong H, MacDonald KS, Kaul R, Virji M, Gray-Owen SD. 2015. Selection for a CEACAM receptor-specific binding phenotype during Neisseria gonorrhoeae infection of the human genital tract. *Infect Immun* **83**:1372–1383. doi:10.1128/IAI.03123-14

Skubitz KM, Skubitz APN. 2008. Interdependency of CEACAM-1, -3, -6, and -8 induced human neutrophil adhesion to endothelial cells. *J Transl Med* **6**:78. doi:10.1186/1479-5876-6-78

Tchoupa AK, Schuhmacher T, Hauck CR. 2014. Signaling by epithelial members of the CEACAM family – mucosal docking sites for pathogenic bacteria. *Cell Commun Signal* **12**:1–10. doi:10.1186/1478-811X-12-27

van Sorge NM, Bonsor DA, Deng L, Lindahl E, Schmitt V, Lyndin M, Schmidt A, Nilsson OR, Brizuela J, Boero E, Sundberg EJ, van Strijp JAG, Doran KS, Singer BB, Lindahl G, McCarthy AJ. 2021. Bacterial protein domains with a novel Ig-like fold target human CEACAM receptors. *EMBO J* e106103. doi:10.15252/embj.2020106103

Van Valen L. 1973. A new evolutionary law. *Evolutionary Theory* **1**:1–30.

Wang J, Gray-Owen SD, Knorre A, Meyer TF, Dehio C. 1998. Opa binding to cellular CD66 receptors mediates the transcellular traversal of Neisseria gonorrhoeae across polarized T84 epithelial cell monolayers. *Mol Microbiol* **30**:657–671. doi:10.1046/j.1365-2958.1998.01102.x

Wang L-G, Lam TT-Y, Xu S, Dai Z, Zhou L, Feng T, Guo P, Dunn CW, Jones BR, Bradley T, Zhu H, Guan Y, Jiang Y, Yu G. 2019. treeio: an R package for phylogenetic tree input and output with richly annotated and associated data. *Molecular Biology and Evolution*. doi:10.1093/molbev/msz240

Weaver S, Shank SD, Spielman SJ, Li M, Muse SV, Kosakovsky Pond SL. 2018. Datamonkey 2.0: A Modern Web Application for Characterizing Selective and Other Evolutionary Processes. *Mol Biol Evol* **35**:773–777. doi:10.1093/molbev/msx335

Wickham H. 2016. ggplot2: Elegant Graphics for Data Analysis.

Wilson D, Pethica R, Zhou Y, Talbot C, Vogel C, Madera M, Chothia C, Gough J. 2009. SUPERFAMILY--sophisticated comparative genomics, data mining, visualization and phylogeny. *Nucleic Acids Res* **37**:D380-6. doi:10.1093/nar/gkn762

Yang DC, Blair KM, Taylor JA, Petersen TW, Sessler T, Tull CM, Leverich CK, Collar AL, Wyckoff TJ, Biboy J, Vollmer W, Salama NR. 2019. A Genome-Wide Helicobacter pylori Morphology Screen Uncovers a Membrane-Spanning Helical Cell Shape Complex. *J Bacteriol* **201**. doi:10.1128/JB.00724-18

Yang Z. 2007. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Mol Biol Evol* **24**:1586–1591. doi:10.1093/molbev/msm088

Yang Z, Wong WSW, Nielsen R. 2005. Bayes empirical bayes inference of amino acid sites under positive selection. *Mol Biol Evol* **22**:1107–1118. doi:10.1093/molbev/msi097

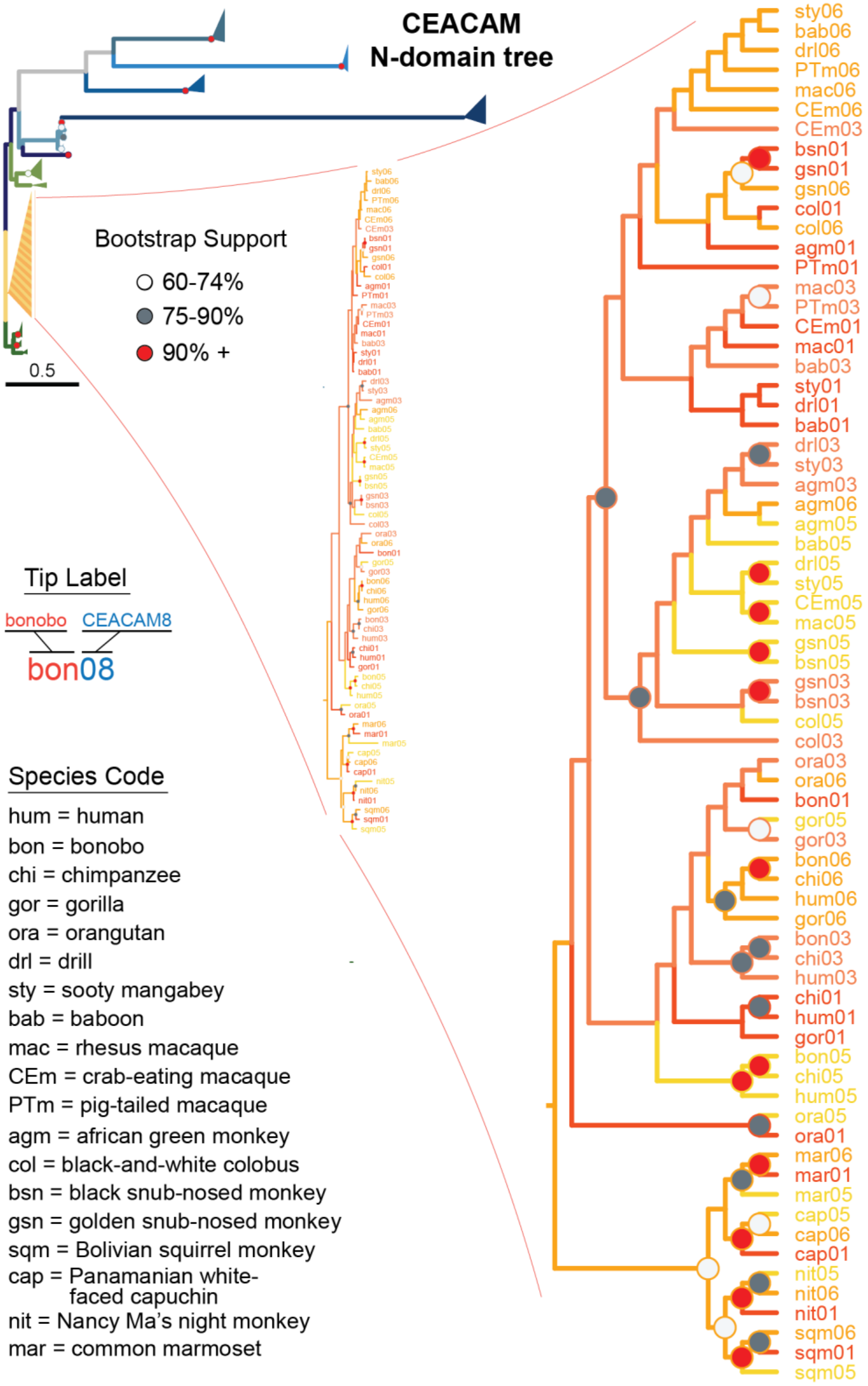
Yu G, Lam TT-Y, Zhu H, Guan Y. 2018. Two Methods for Mapping and Visualizing Associated Data on Phylogeny Using ggtree. *Mol Biol Evol* **35**:3041–3043. doi:10.1093/molbev/msy194

Yu G, Smith DK, Zhu H, Guan Y, Lam TT. 2017. ggtree : An r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol Evol* **8**:28–36. doi:10.1111/2041-210x.12628

Zid M, Drouin G. 2013. Gene conversions are under purifying selection in the carcinoembryonic antigen immunoglobulin gene families of primates. *Genomics* **102**:301–309. doi:10.1016/j.ygeno.2013.07.003

Zimmermann W. 2019. Evolution: Decoy Receptors as Unique Weapons to Fight Pathogens. *Current Biology*. doi:10.1016/j.cub.2018.12.006

EXTENDED AND SUPPLEMENTARY FIGURES



Extended Fig. 1

Expanded view of CEACAM1, CEACAM3, CEACAM5 and CEACAM6 clade from Fig. 4B.

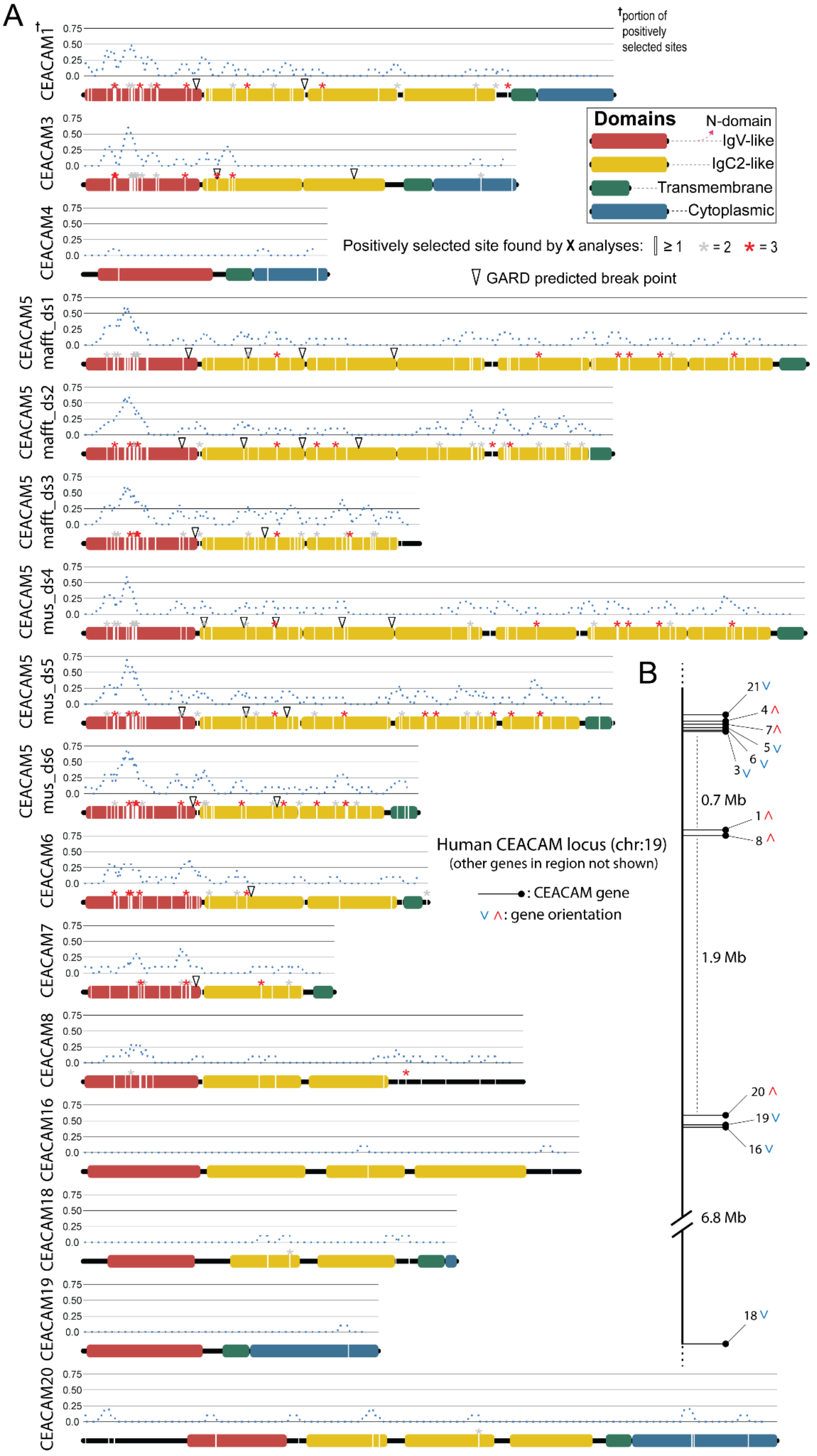

A
†portion of positively selected sites
0.75
0.50
0.25
0.0
CEACAM1
CEACAM3
CEACAM4
CEACAM5 mafft_ds1
CEACAM5 mafft_ds2
CEACAM5 mafft_ds3
CEACAM5 mus_ds4
CEACAM5 mus_ds5
CEACAM5 mus_ds6
CEACAM6
CEACAM7
CEACAM8
CEACAM16
CEACAM18
CEACAM19
CEACAM20
Domains
N-domain
IgV-like
IgC2-like
Transmembrane
Cytoplasmic
Positively selected site found by X analyses: ≥ 1 * = 2 * = 3
∇ GARD predicted break point
B
Human CEACAM locus (chr:19)
(other genes in region not shown)
●: CEACAM gene
∨ ∧: gene orientation
21 ∨
4 ∧
7 ∧
5 ∨
6 ∨
3 ∨
0.7 Mb
1 ∧
8 ∧
1.9 Mb
20 ∧
19 ∨
16 ∨
6.8 Mb
18 ∨

Fig. S1

Sites with elevated dN/dS in all human CEACAM proteins. A) Sites in CEACAM proteins identified as evolving rapidly in specific domains by one (white line), two (gray asterisks) or three (red asterisks) evolutionary analyses. Dotted blue line indicates the proportion of sites identified as evolving rapidly across a ten amino acid sliding window. Open triangles show GARD predictions of the approximate locations of recombination breakpoints. B) Location of human CEACAM genes along chromosome 19. Other genes on chromosome 19 are not shown.

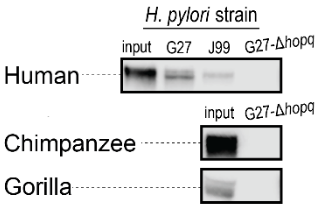


Fig. S2

Binding assay to assess interactions between *H. pylori* strain G27 *Δhopq* and GFP-tagged CEACAM1 N-domain constructs for human, chimpanzee, and gorilla, by pulldown experiments and visualization by western blot.

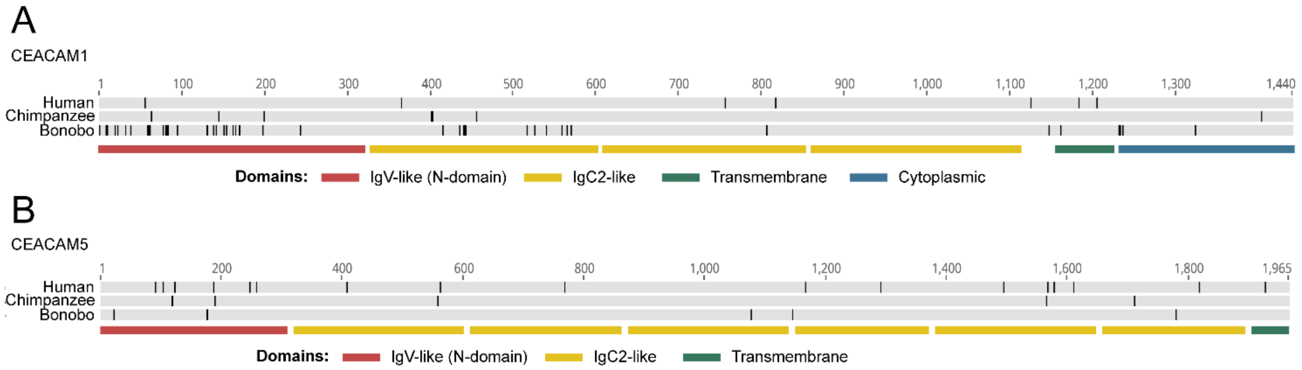


Fig. S3

Bonobo CEACAM sequence alignments. Human, chimpanzee and bonobo CEACAM1 (A) and CEACAM5 (B) alignments by MAFFT translation alignment implemented in Geneious Prime 2020.2.2. Black lines mark differences from consensus. Lower lines show location of CEACAM domains.

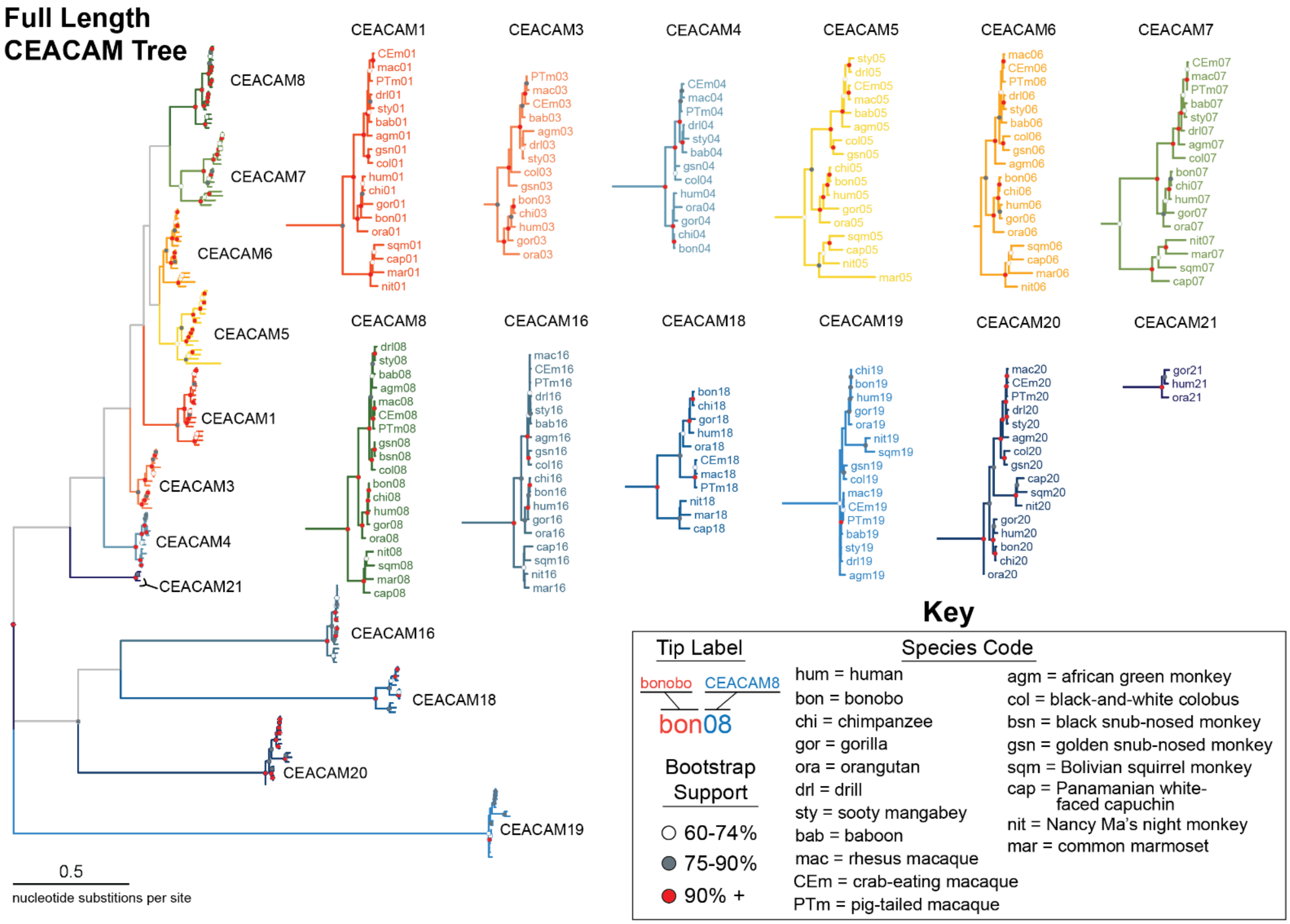
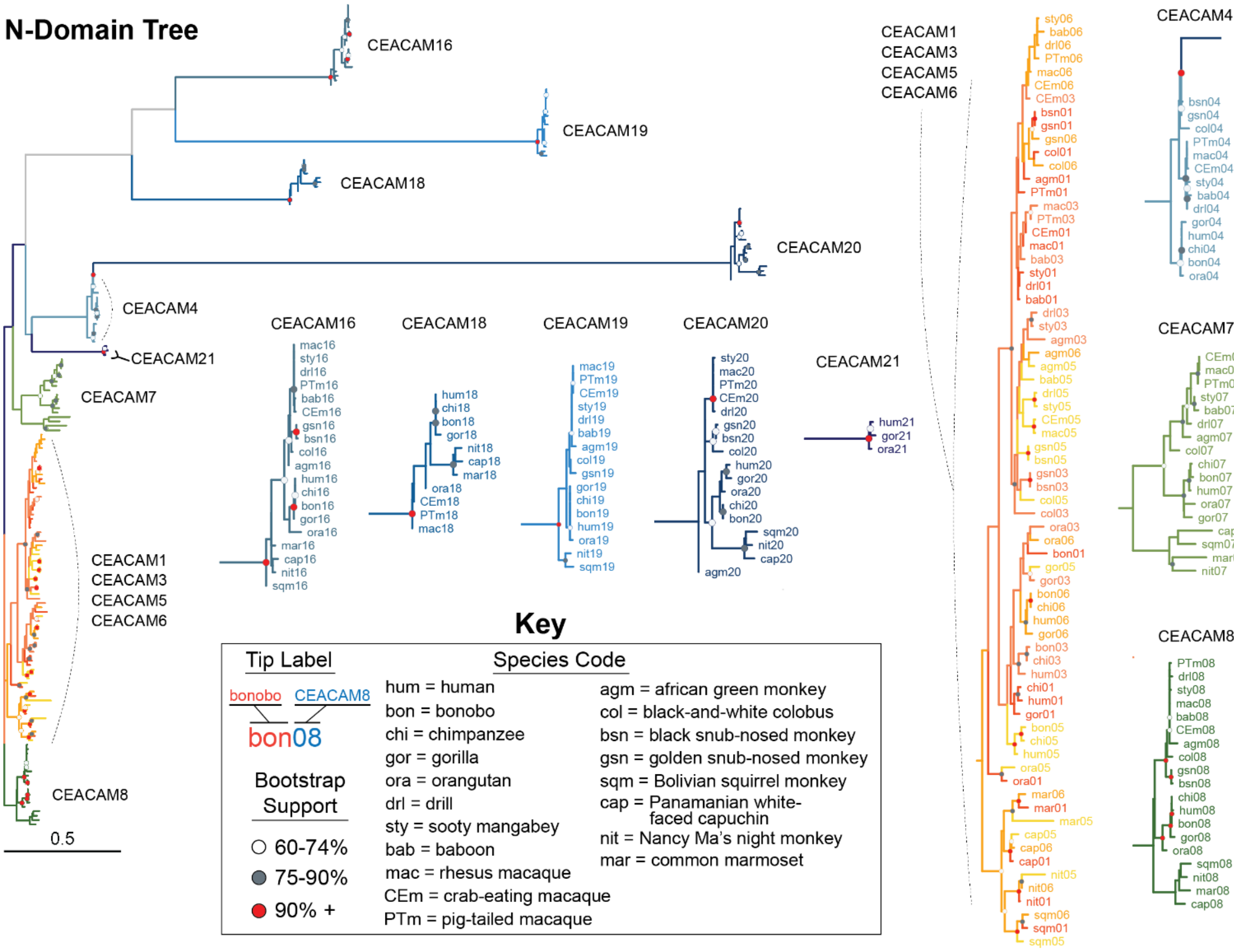

Full Length
CEACAM Tree
CEACAM8
CEACAM7
CEACAM6
CEACAM5
CEACAM1
CEACAM3
CEACAM4
CEACAM21
CEACAM16
CEACAM18
CEACAM20
CEACAM19
0.5
nucleotide substitutions per site
CEACAM1
CEm01 mac01 PTm01 drl01 sty01 bab01 agm01 gsn01 col01 hum01 chi01 gor01 bon01 ora01 sqm01 cap01 mar01 nit01
CEACAM3
PTm03 mac03 CEm03 bab03 agm03 drl03 sty03 col03 gsn03 bon03 chi03 hum03 gor03 ora03
CEACAM4
CEm04 mac04 PTm04 drl04 sty04 bab04 gsn04 col04 hum04 ora04 gor04 chi04 bon04
CEACAM5
sty05 drl05 CEm05 mac05 bab05 agm05 col05 gsn05 chi05 bon05 hum05 gor05 ora05 sqm05 cap05 nit05 mar05
CEACAM6
mac06 CEm06 PTm06 drl06 sty06 bab06 col06 gsn06 agm06 bon06 chi06 hum06 gor06 ora06 sqm06 cap06 mar06 nit06
CEACAM7
CEm07 mac07 PTm07 bab07 sty07 drl07 agm07 col07 bon07 chi07 hum07 gor07 ora07 nit07 mar07 sqm07 cap07
CEACAM8
drl08 sty08 bab08 agm08 mac08 CEm08 PTm08 gsn08 bsn08 col08 bon08 chi08 hum08 gor08 ora08 nit08 sqm08 mar08 cap08
CEACAM16
mac16 CEm16 PTm16 drl16 sty16 bab16 agm16 gsn16 col16 chi16 bon16 hum16 gor16 ora16 cap16 sqm16 nit16 mar16
CEACAM18
bon18 chi18 gor18 hum18 ora18 CEm18 mac18 PTm18 nit18 mar18 cap18
CEACAM19
chi19 bon19 hum19 gor19 ora19 nit19 sqm19 gsn19 col19 mac19 CEm19 PTm19 bab19 sty19 drl19 agm19
CEACAM20
mac20 CEm20 PTm20 drl20 sty20 agm20 col20 gsn20 cap20 sqm20 nit20 gor20 hum20 bon20 chi20 ora20
CEACAM21
gor21 hum21 ora21
Key
Tip Label
bonobo
CEACAM8
bon08
Bootstrap Support
○ 60-74%
● 75-90%
● 90% +
Species Code
hum = human
bon = bonobo
chi = chimpanzee
gor = gorilla
ora = orangutan
drl = drill
sty = sooty mangabey
bab = baboon
mac = rhesus macaque
CEm = crab-eating macaque
PTm = pig-tailed macaque
agm = african green monkey
col = black-and-white colobus
bsn = black snub-nosed monkey
gsn = golden snub-nosed monkey
sqm = Bolivian squirrel monkey
cap = Panamanian white-faced capuchin
nit = Nancy Ma's night monkey
mar = common marmoset

Fig. S4

Maximum likelihood-based phylogeny of full length CEACAM protein coding sequences as represented in Fig. 4A, but with clades expanded. Clades encompassing individual CEACAM orthologs are shown isolated and expanded.



1 Fig. S5

2 Maximum likelihood-based phylogeny of CEACAM IgV-like (N-domain) sequences as represented in Fig. 4B,

3 but with clades expanded. Clades encompassing individual CEACAM orthologs along with the CEACAM1,

4 CEACAM3, CEACAM5 and CEACAM6 clade are shown isolated and expanded.

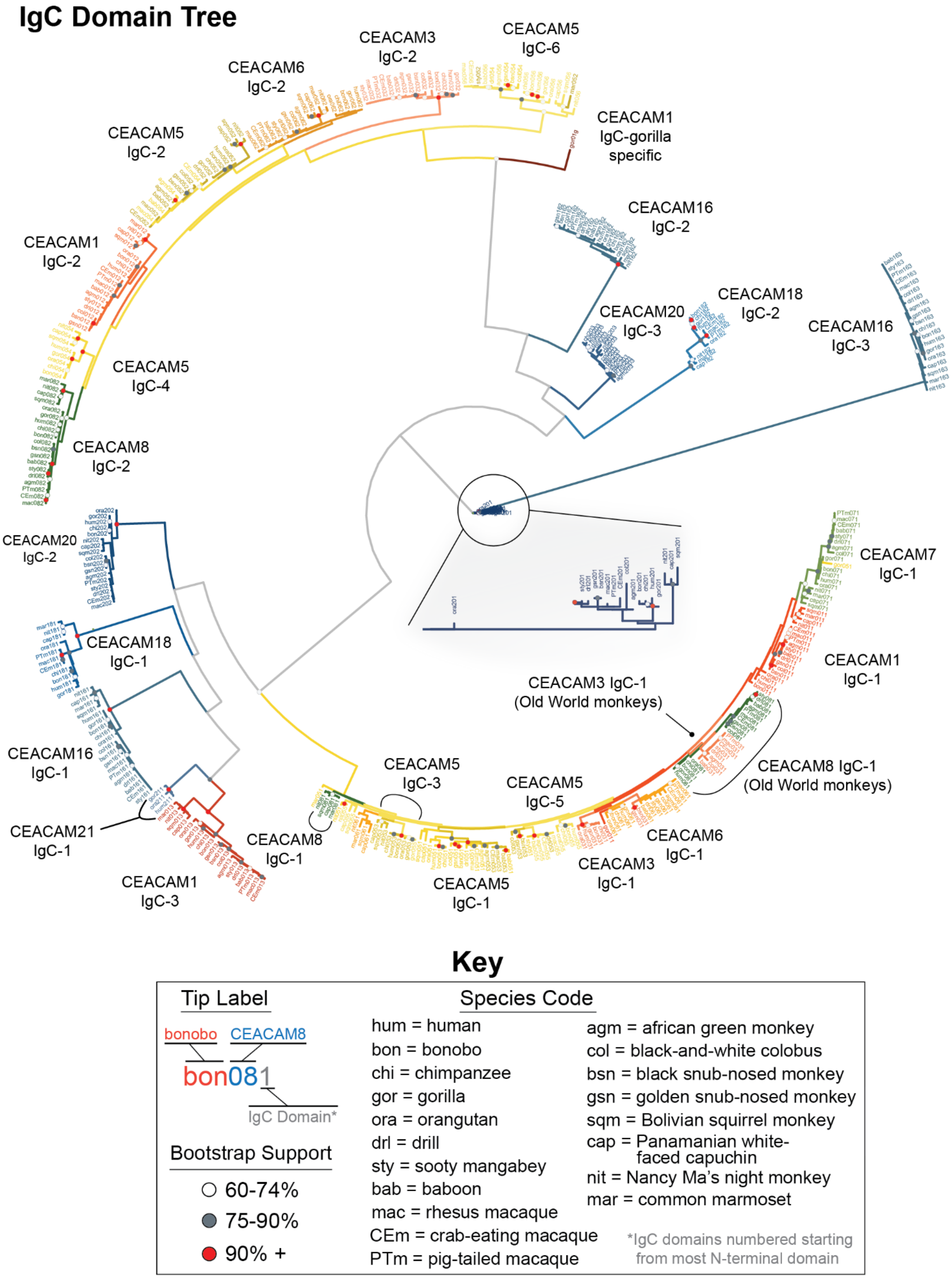

IgC Domain Tree
CEACAM6 IgC-2
CEACAM3 IgC-2
CEACAM5 IgC-6
CEACAM1 IgC-gorilla specific
CEACAM5 IgC-2
CEACAM16 IgC-2
CEACAM1 IgC-2
CEACAM20 IgC-3
CEACAM18 IgC-2
CEACAM16 IgC-3
CEACAM5 IgC-4
CEACAM8 IgC-2
CEACAM20 IgC-2
CEACAM7 IgC-1
CEACAM18 IgC-1
CEACAM1 IgC-1
CEACAM3 IgC-1 (Old World monkeys)
CEACAM16 IgC-1
CEACAM5 IgC-3
CEACAM5 IgC-5
CEACAM8 IgC-1 (Old World monkeys)
CEACAM21 IgC-1
CEACAM8 IgC-1
CEACAM6 IgC-1
CEACAM3 IgC-1
CEACAM1 IgC-3
CEACAM5 IgC-1
Key
Tip Label
bonobo
CEACAM8
bon081
IgC Domain*
Bootstrap Support
○ 60-74%
● 75-90%
● 90% +
Species Code
hum = human
bon = bonobo
chi = chimpanzee
gor = gorilla
ora = orangutan
drl = drill
sty = sooty mangabey
bab = baboon
mac = rhesus macaque
CEm = crab-eating macaque
PTm = pig-tailed macaque
agm = african green monkey
col = black-and-white colobus
bsn = black snub-nosed monkey
gsn = golden snub-nosed monkey
sqm = Bolivian squirrel monkey
cap = Panamanian white-faced capuchin
nit = Nancy Ma's night monkey
mar = common marmoset
*IgC domains numbered starting from most N-terminal domain

Fig. S6

Maximum likelihood-based phylogeny of CEACAM IgC-like domain sequences. Expanded view of CEACAM20 clade shown.

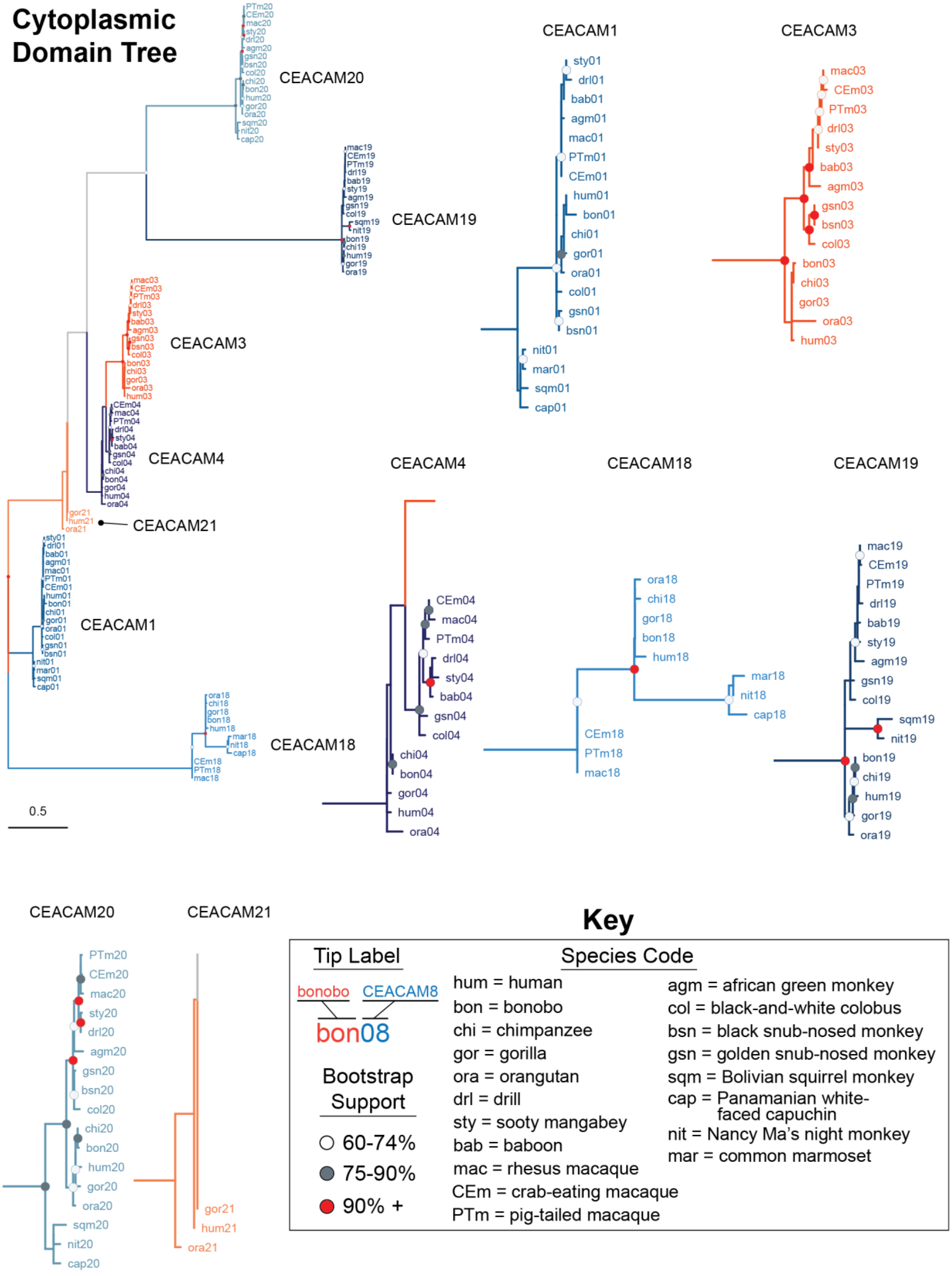

Cytoplasmic Domain Tree
CEACAM20
CEACAM19
CEACAM3
CEACAM4
CEACAM21
CEACAM1
CEACAM18
0.5
CEACAM1
sty01
drl01
bab01
agm01
mac01
PTm01
CEm01
hum01
bon01
chi01
gor01
ora01
col01
gsn01
bsn01
nit01
mar01
sqm01
cap01
CEACAM3
mac03
CEm03
PTm03
drl03
sty03
bab03
agm03
gsn03
bsn03
col03
bon03
chi03
gor03
ora03
hum03
CEACAM4
CEm04
mac04
PTm04
drl04
sty04
bab04
gsn04
col04
chi04
bon04
gor04
hum04
ora04
CEACAM18
ora18
chi18
gor18
bon18
hum18
mar18
nit18
cap18
CEm18
PTm18
mac18
CEACAM19
mac19
CEm19
PTm19
drl19
bab19
sty19
agm19
gsn19
col19
sqm19
nit19
bon19
chi19
hum19
gor19
ora19
CEACAM20
PTm20
CEm20
mac20
sty20
drl20
agm20
gsn20
bsn20
col20
chi20
bon20
hum20
gor20
ora20
sqm20
nit20
cap20
CEACAM21
gor21
hum21
ora21
Key
Tip Label
bonobo
CEACAM8
bon08
Bootstrap Support
○ 60-74%
● 75-90%
● 90% +
Species Code
hum = human
bon = bonobo
chi = chimpanzee
gor = gorilla
ora = orangutan
drl = drill
sty = sooty mangabey
bab = baboon
mac = rhesus macaque
CEm = crab-eating macaque
PTm = pig-tailed macaque
agm = african green monkey
col = black-and-white colobus
bsn = black snub-nosed monkey
gsn = golden snub-nosed monkey
sqm = Bolivian squirrel monkey
cap = Panamanian white-faced capuchin
nit = Nancy Ma's night monkey
mar = common marmoset

Fig. S7

Maximum likelihood-based phylogeny of CEACAM cytoplasmic domain sequences. Clades encompassing individual CEACAM orthologs are shown isolated and expanded.

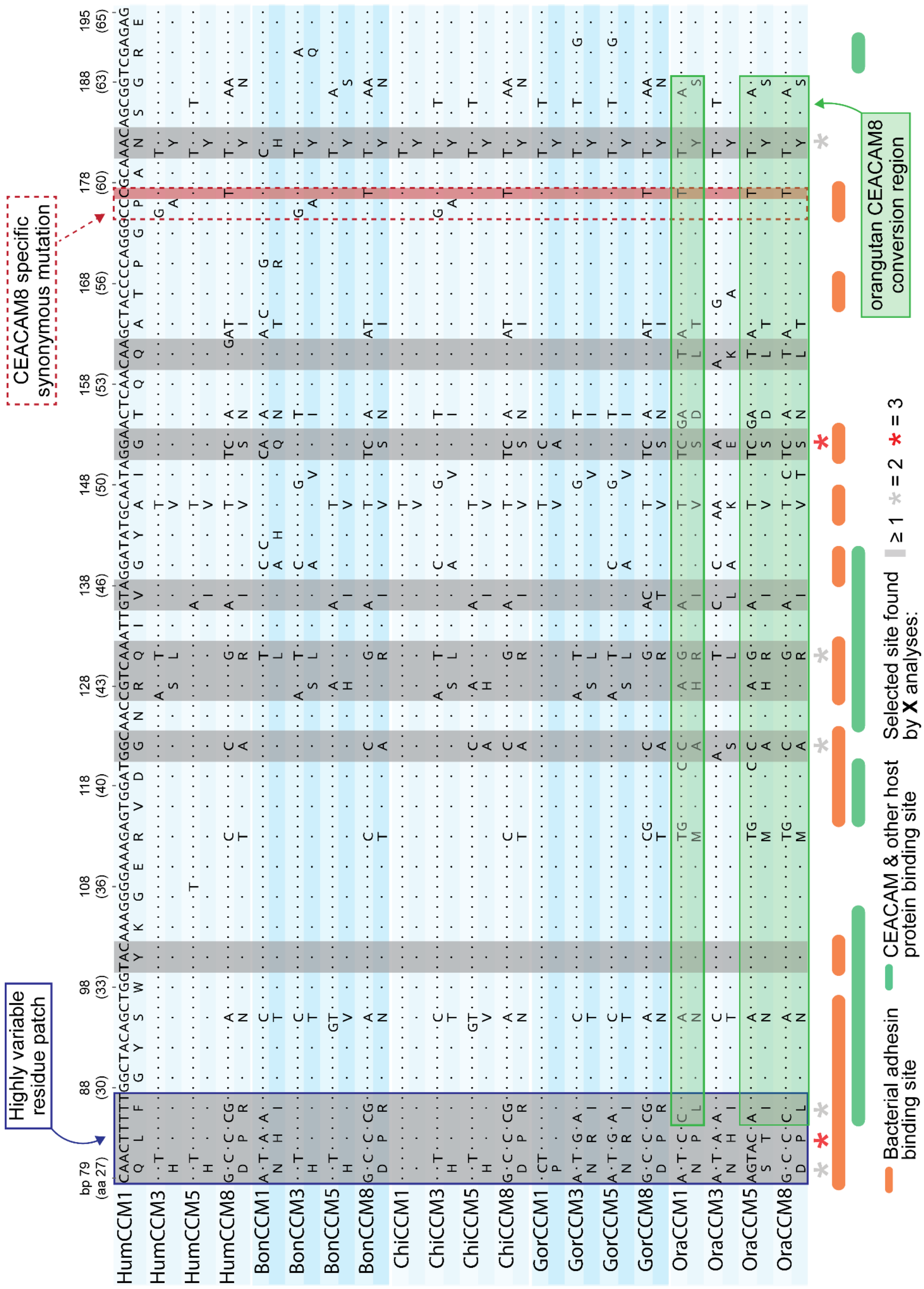

Highly variable residue patch
CEACAM8 specific synonymous mutation
orangutan CEACAM8 conversion region
bp 79 (aa 27)
88 (30)
98 (33)
108 (36)
118 (40)
128 (43)
138 (46)
148 (50)
158 (53)
168 (56)
178 (60)
188 (63)
195 (65)
CAACTTTTTGGCTACAGCTGGTACAAAGGGGAAAGAGTGGATGGCAACCGTCAAATTGTAGGATATGCAATAGGAACTCAACAAGCTACCCCAGGGCCCGCAAACAGCGGTCGAGAG
Q L F G Y S W Y K G E R V D G N R Q I V G Y A I G T Q Q A T P G P A N S G R E
HumCCM1
HumCCM3
HumCCM5
HumCCM8
BonCCM1
BonCCM3
BonCCM5
BonCCM8
ChiCCM1
ChiCCM3
ChiCCM5
ChiCCM8
GorCCM1
GorCCM3
GorCCM5
GorCCM8
OraCCM1
OraCCM3
OraCCM5
OraCCM8
Bacterial adhesin binding site
CEACAM & other host protein binding site
Selected site found by X analyses:
≥ 1 * = 2 * = 3

Fig. S8

Multiple sequence alignment of CEACAM1, CEACAM3, CEACAM5 and CEACAM8 orthologs for human, bonobo, chimpanzee, gorilla, and orangutan. Translation of each nucleotide sequence is positioned on the line below. Sites known to influence adhesin and host protein binding (TableS3) are indicated as are sites identified as evolving under positive selection.

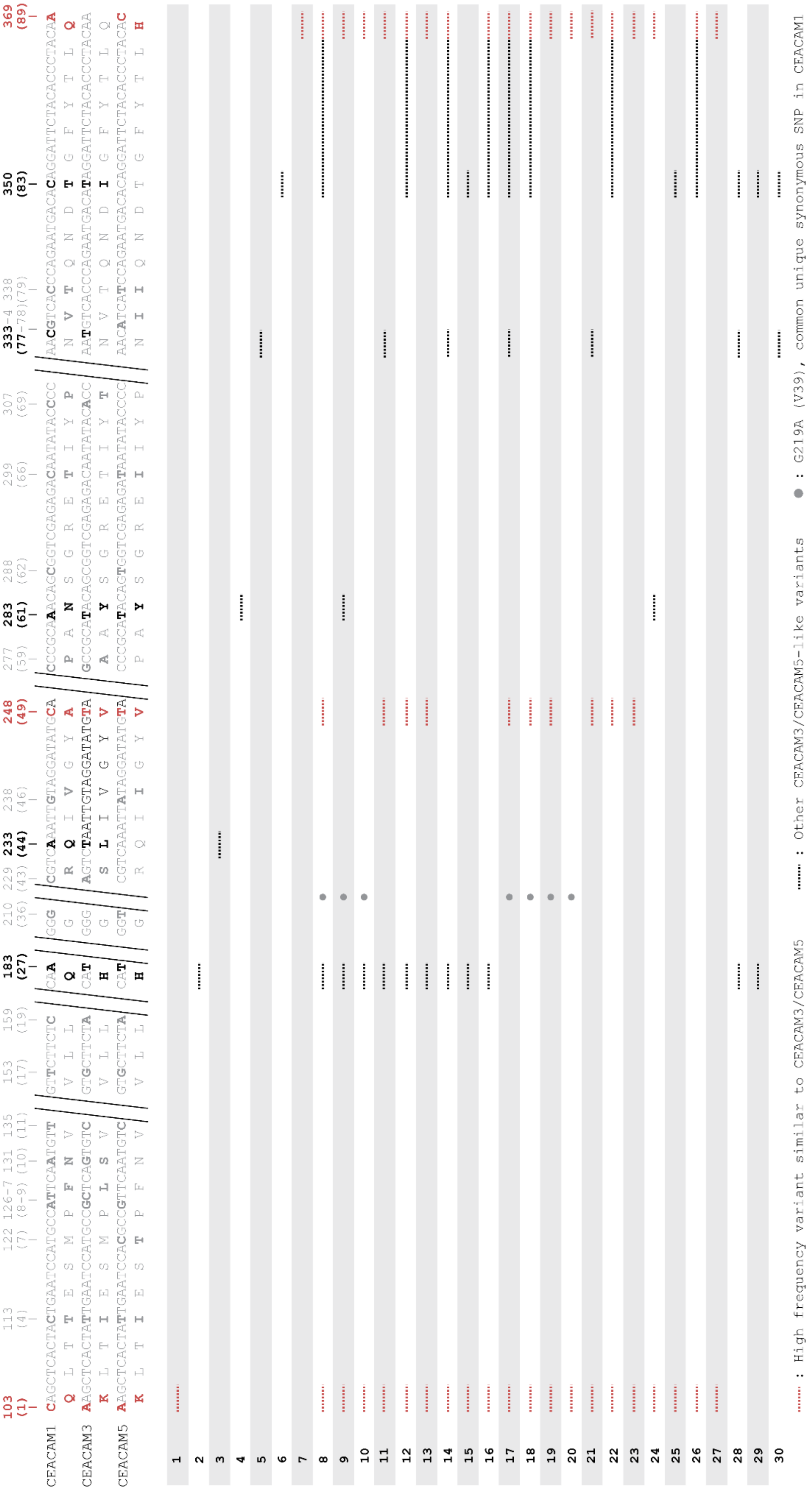

103 (1)
113 (4)
122 (7)
126-7 (8-9)
131 (10)
135 (11)
153 (17)
159 (19)
183 (27)
210 (36)
229 (43)
233 (44)
238 (46)
248 (49)
277 (59)
283 (61)
288 (62)
299 (66)
307 (69)
333-4 (77-78)
338 (79)
350 (83)
369 (89)
CEACAM1
CAGCTCACTACTGAATCCATGCCATTCAATGTT GTTCTTCTC CAA GGG CGTCAAATTGTAGGATATGCA CCCGCAAACAGCGGTCGAGAGACAATATACCCC AACGTCACCCAGAATGACACAGGATTCTACACCCTACAA
Q L T T E S M P F N V V L L Q G R Q I V G Y A P A N S G R E T I Y P N V T Q N D T G F Y T L Q
CEACAM3
AAGCTCACTATTGAATCCATGCCGCTCAGTGTC GTGCTTCTA CAT GGG AGTCTAATTGTAGGATATGTA GCCGCATACAGCGGTCGAGAGACAATATACACC AATGTCACCCAGAATGACATAGGATTCTACACCCTACAA
K L T I E S M P L S V V L L H G S L I V G Y V A A Y S G R E T I Y T N V T Q N D I G F Y T L Q
CEACAM5
AAGCTCACTATTGAATCCACGCCGTTCAATGTC GTGCTTCTA CAT GGT CGTCAAATTATAGGATATGTA CCCGCATACAGTGGTCGAGAGATAATATACCCC AACATCATCCAGAATGACACAGGATTCTACACCCTACAC
K L T I E S T P F N V V L L H G R Q I I G Y V P A Y S G R E I I Y P N I I Q N D T G F Y T L H
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
...... : High frequency variant similar to CEACAM3/CEACAM5
...... : Other CEACAM3/CEACAM5-like variants
● : G219A (V39), common unique synonymous SNP in CEACAM1

Fig. S9

Other CEACAM-like human CEACAM1 haplotypes. Alignment of human CEACAM1, CEACAM3 and CEACAM5 N-domain reference nucleotide sequences with amino acid translations below. Long invariable alignment regions were removed. Sites that differ in CEACAM3 or CEACAM5 relative to CEACAM1 are bolded. Sites found in variant CEACAM1 haplotypes are in black. Changes that encode the high frequency variants Q1K, A49V, and Q89H are in red. Below alignment each row is a unique human CEACAM1 N-domain haplotype. Lines indicate variant regions in CEACAM1. Only haplotypes that increase similarity to CEACAM3 or CEACAM5 are shown.

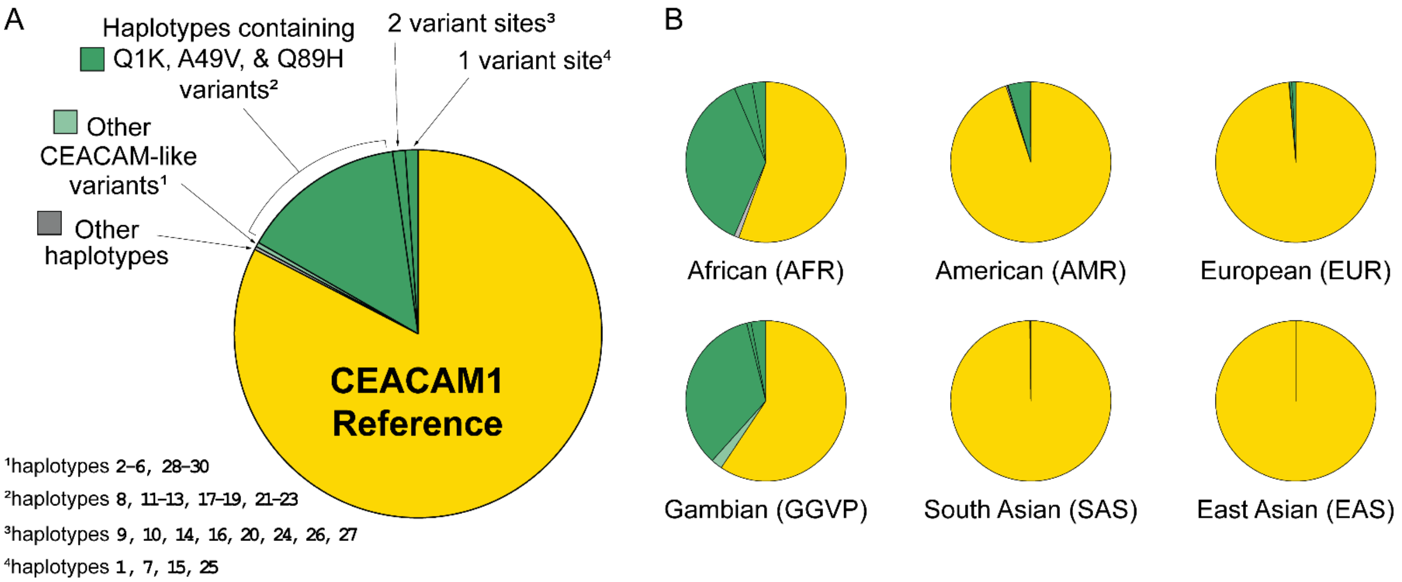


Fig. S10

Frequency of variant human CEACAM1 haplotypes. A) Overall frequency of CEACAM1 variants Q1K, 449V, Q89H and other variant haplotypes in humans. The indicated CEACAM-like haplotypes are enumerated in Fig. 9. B) Frequency of CEACAM1 variants across different human populations.

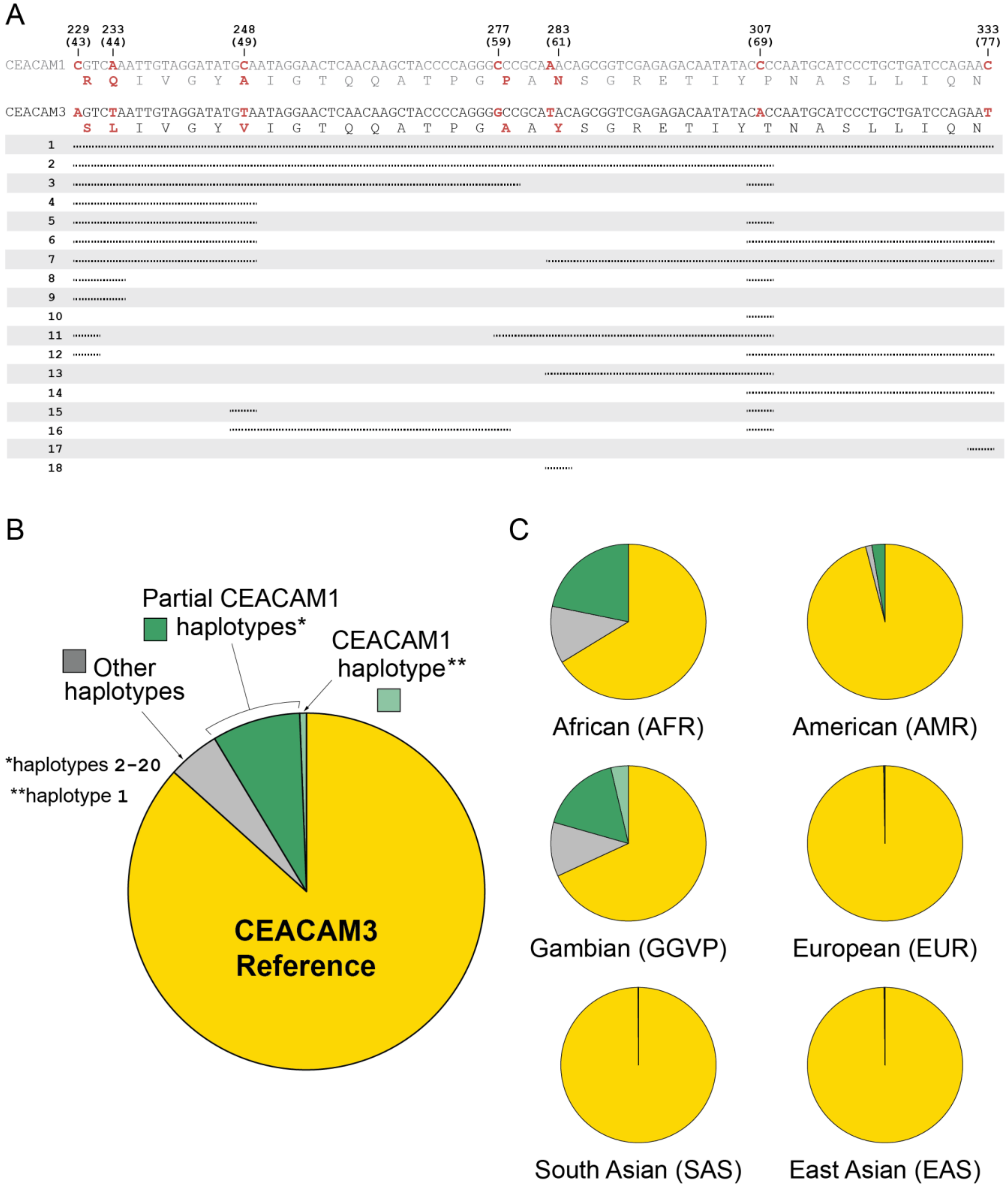


Fig. S11

Human CEACAM1-like CEACAM3 haplotypes. A) Alignment of human CEACAM1 and CEACAM3 reference sequences. Disagreements are bolded in red with the amino acid translation below each sequence. Below alignment each row represents a unique human CEACAM3 haplotype. Lines indicate variant regions that match the human CEACAM1 reference sequence. Only haplotypes that increase similarity to the human

CEACAM1 reference sequence are shown. B) Overall frequency of variant CEACAM3 haplotypes in humans. The CEACAM1-like haplotypes indicated are enumerated in panel A. C) Frequency of CEACAM3 variants across different human populations.

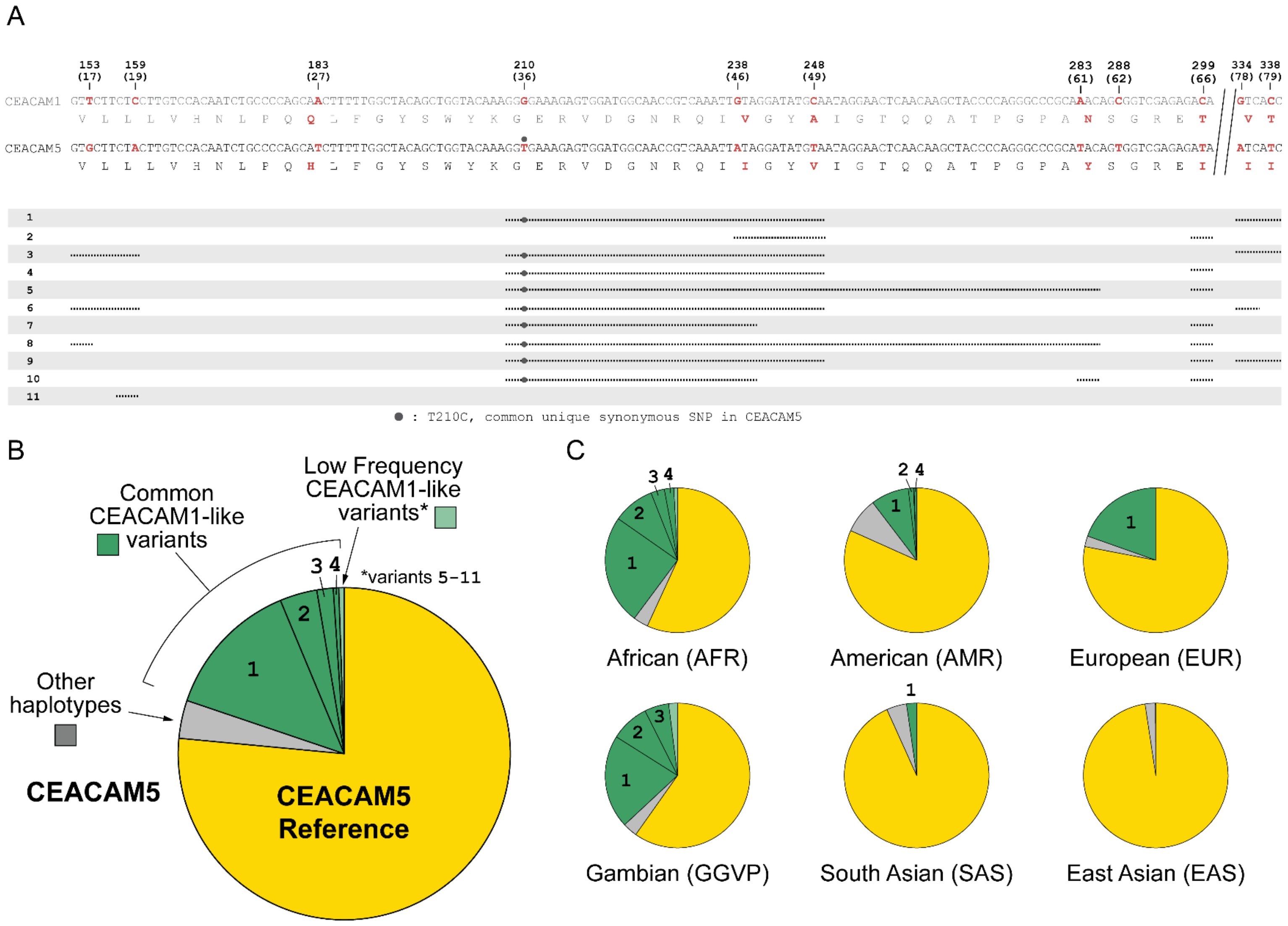

A
153 (17)
159 (19)
183 (27)
210 (36)
238 (46)
248 (49)
283 (61)
288 (62)
299 (66)
334 (78)
338 (79)
CEACAM1 GTTCTTCTCCTTGTCCACAATCTGCCCCAGCAACTTTTTGGCTACAGCTGGTACAAAGGGGAAAGAGTGGATGGCAACCGTCAAATTGTAGGATATGCAATAGGAACTCAACAAGCTACCCCAGGGCCCGCAAACAGCGGTCGAGAGACA GTCACC
V L L L V H N L P Q Q L F G Y S W Y K G E R V D G N R Q I V G Y A I G T Q Q A T P G P A N S G R E T V T
CEACAM5 GTGCTTCTACTTGTCCACAATCTGCCCCAGCATCTTTTTGGCTACAGCTGGTACAAAGGTGAAAGAGTGGATGGCAACCGTCAAATTATAGGATATGTAATAGGAACTCAACAAGCTACCCCAGGGCCCGCATACAGTGGTCGAGAGATA ATCATC
V L L L V H N L P Q H L F G Y S W Y K G E R V D G N R Q I I G Y V I G T Q Q A T P G P A Y S G R E I I I
1
2
3
4
5
6
7
8
9
10
11
● : T210C, common unique synonymous SNP in CEACAM5
B
Common CEACAM1-like variants
Low Frequency CEACAM1-like variants*
Other haplotypes
*variants 5-11
1
2
3 4
CEACAM5
CEACAM5 Reference
C
African (AFR)
American (AMR)
European (EUR)
Gambian (GGVP)
South Asian (SAS)
East Asian (EAS)
1
2
3
4

Fig. S12

Human CEACAM1-like CEACAM5 haplotypes. A) Alignment of human CEACAM1 and CEACAM5 reference sequences. Disagreements are bolded in red with the amino acid translation below each sequence. Below alignment each row represents a unique human CEACAM5 haplotype. Lines indicate variant regions that match the human CEACAM1 reference sequence. Only haplotypes that increase similarity to the human CEACAM1 reference sequence are shown. B) Overall frequency of variant CEACAM5 haplotypes in humans. The CEACAM1-like haplotypes indicated are enumerated in panel A. C) Frequency of CEACAM5 variants across different human populations.

Excel S1

Primate CEACAM sequences extracted for evolutionary analyses and phylogenetic reconstructions.

Table S1

PAML NS sites tests of selection in primate CEACAMs

Table S2

Sites identified as evolving under positive selection by evolutionary analyses and GARD predicted recombination breakpoints.

Table S3

References for sites identified as contributing to CEACAM1 binding with host proteins and bacterial adhesins as well as the specific sites identified.

Table S4

Table of oligomers, DNA templates and their order in assembly reactions used to assemble CEACAM1 N-domain expression plasmids.

Table S5

Sources of template sequences for CEACAM1 and other plasmid components used for expression plasmid construction.