Rapid Divergence of Two Classes of *Haemophilus ducreyi*

Emily E. Ricotta\textsuperscript{a§}, Nan Wang\textsuperscript{b¶}, Robin Cutler\textsuperscript{b†}, Jeffrey G. Lawrence\textsuperscript{c}, Tricia L. Humphreys\textsuperscript{a,b,*}

\textsuperscript{a} Department of Biology, Allegheny College, Meadville, PA
\textsuperscript{b} Department of Biology, Grinnell College, Grinnell, IA
\textsuperscript{c} Department of Biological Sciences, University of Pittsburgh, Pittsburgh, PA

Running Title: Rapid Divergence of Two Classes of *H. ducreyi*

Keywords: *Haemophilus ducreyi*, chancroid, Pasteurellaceae

* Corresponding author

520 N. Main Street, Box 10
Meadville, PA 16335 USA
Phone (814) 332-2967
FAX (814) 332-2789
E-mail: thumphre@allegheny.edu

§ Present address: Johns Hopkins Bloomberg School of Public Health, W. Harry Feinstone
Department of Molecular Microbiology and Immunology, 615 N. Wolfe Street, Baltimore, MD
21205

¶ Present address: Remsen 153, 3400 N. Charles Street, Johns Hopkins University, Baltimore, MD
21218

† Present address: Division of Biology and Biomedical Sciences, 660 S. Euclid Ave.,
Washington University School of Medicine, St. Louis, MO 63110
Abstract

*Haemophilus ducreyi*, the etiologic agent of chancroid, expresses variants of several key virulence factors. While previous reports suggested that *H. ducreyi* strains formed two clonal populations, the differences between—and diversity within—these populations was unclear. To assess their variability, we examined sequence diversity at 11 *H. ducreyi* loci, including virulence and housekeeping genes, augmenting published data sets with PCR-amplified genes to acquire data for at least 10 strains at each locus. While sequences from all 11 loci place strains into 2 distinct groups, there was very little variation within each group. The difference between alleles of the two groups was variable and large at 3 loci encoding surface-exposed proteins ($0.4 < K_s < 1.3$), but consistently small at genes encoding cytoplasmic or periplasmic proteins ($K_s < 0.09$). These data suggest that the two classes have recently diverged, that recombination has introduced variant alleles into at least 3 distinct loci, and that these alleles have been confined to one of the two classes. In addition, recombination is evident among alleles within, but not between, classes. Rather than clones of the same species, these properties indicate that the two classes may form distinct species.
INTRODUCTION

*Haemophilus ducreyi* is a gram negative coccobacillus in the family *Pasteurellaceae* (2) and the causative agent of chancroid, a sexually-transmitted genital ulcer disease which facilitates transmission of HIV-1 (12, 14, 23, 45). While *H. ducreyi* was placed in the polyphyletic genus *Haemophilus* due to its requirement for hemin, *Haemophilus* species do not form a cohesive group within the *Pasteurellaceae* (6, 20); species most closely related to *H. ducreyi* include *Actinobacter pleuropneumoniae* and *Nicoletella semolina* (Fig. S1, 16).

Previous attempts to examine the diversity within the species of *H. ducreyi* using immunotyping (36) and outer membrane profiling (32) showed very few differences among strains. Molecular analyses showed some differences, e.g., ribotyping of strains from two outbreaks yielded multiple RFLP patterns (19), but designations based on these differences have not been widely adopted or used for epidemiologic studies, in part due to the difficulty in culturing this fastidious organism. The lack of a standard typing system makes epidemiologic studies more difficult, and epidemiologic data for chancroid is therefore limited.

One way to study diversity in *H. ducreyi* is to assess its population structure. Whereas the lack of recombination in some bacterial species generates stable clones (5), others recombine sufficiently rapidly to obscure clonal relationships (10, 11). While many *Haemophilus* species are naturally competent, *H. ducreyi* is considered non-transformable due to defects in the DNA uptake apparatus (16, 34); this would likely decrease rates of recombination and generate stable clones. Initial studies of the clonal structure of *H. ducreyi* showed limited diversity in specific outer membrane components, including DsrA, DltA, NcaA, Hlp, MOMP/OmpA2 and lipooligosaccharide (46). The DNA sequences and immunoreactivity suggested that *H. ducreyi*
formed two potentially clonal populations, termed class I and class II, a division supported by proteomic studies (33).

However, the relationships between the two classes is entirely unclear, as genetic differences were only assessed at loci encoding surface-exposed antigens; these loci experience positive and diversifying selection, leading to levels of diversity that are not shared at the majority of genes in the genome (24). It is not clear if these two classes represent two distinct clonal populations of *H. ducreyi* whose genotypic similarities reflect recent periodic selection (25), or two recombining populations whose similarities reflect recent, locus-specific selective sweeps (18). Here we examine the relationships among *H. ducreyi* strains using sequences of 11 genes, including five genes that encode cytoplasmic or periplasmic proteins and one that encodes a vaccine candidate. This provides a rich data set with which to perform multilocus sequence analysis (42).
MATERIALS AND METHODS

Strains and culture conditions. Strains were cultured on chocolate agar supplemented with 1% IsoVitaleX. Strains were revived from frozen stocks, incubated at 33-35°C with 10% CO₂ for 24 hours, and subcultured no more than once prior to PCR.

PCR and sequencing. Sequences of the lspA2, murC, pal, recA and wecA loci were obtained from fragments amplified by PCR using the following primers: lspA2, AAGTTTCAGCAAGAGCGGC and TATTGGCTGCAAGCTCTG (44); murC, GGCATTGCCGAAGTGTTATT and TATGATGGCCCGGTACATTT; ncaA, GGTGGATTATGTGTCAGATAATTTG and CTAAGCCGCTAAAAATTCGATG (46); pal, AGTAGTTGTCAGTTAAAAACAGATG and AAATTAGTACTCTAATACTGCACGG (41); recA, CATTATGGCAGCGGATAAAAA and TCCTCAAACGCTTTGATCAAA; wecA, CCGGAATCCAACCCATAAACAC and GATATCGGTTCGCTTTGCCTTGGTT. The specific locations of amplified regions relative to the indicated gene are indicated in Fig. S2. Genomic DNA was obtained using GeneReleaser® according to the manufacturer’s instructions. Sequencing was done at the University of Iowa DNA Facility, Eurofins MWG Operon or Yale University DNA Analysis Facility. Sequences generated by this study (Table 1) have been deposited in GenBank and assigned accession numbers 00000-00000. All other sequences were obtained from NCBI as follows: dsrA (AF187001-AF187009, AY606120-AY606127, AY612644), sapA (HQ630266-HQ630274), fgbA (HQ630257-HQ630265), hgbA (AY606114-AY606118, AY603046-AY603049), mtrC (HQ712073-HQ712081) and ncaA (AY606128-AY606130, AY612645-
AY612647). Sequences for strain 35000HP were obtained from its genomic sequence (NC_002940).

Genome sequences. The sequences of *Bacillus anthracis* Ames (NC_003997), *B. cereus* 03BB102 (NC_012472), *Bordetella pertussis* Tahoma I (NC_002929), *B. parapertussis* 12822 (NC_002928), *Escherichia coli* O157:H7 Sakai (NC_002695), *Haemophilus ducreyi* 35000HP (NC_002940), *Klebsiella pneumoniae* 342 (NC_011283), *K. variicola* At-22 (NC_013850), *Mycobacterium tuberculosis* H37Ra (NC_009525), *M. bovis* BCG (NC_008769), *Shigella dysenteriae* Sd197 (NC_007606), *Yersinia pestis* CO92 (NC_003143), and *Y. pseudotuberculosis* YIPIII (NC_010465) were downloaded from NCBI; genes were identified using the accompanying annotation.

Computational analyses. DNA sequence alignments were aligned using MUSCLE (8). Substitution rates were calculated using the method of Li, Wu and Luo (26). Mean substitution rates between entire genomes were calculated as weighted averages of Ks values for orthologues; orthologues were identified as reciprocal best-matches in BLAST searches, requiring at least 80% sequence identity. Codon usage bias was calculated as Karlin’s E (22), using all genes to represent mutational bias and Sharp’s set of 40 translational genes (37) to represent the endpoint of codon selection. Phylogenies were calculated by Dollo parsimony (9) or by maximum likelihood using PhyML (17).
RESULTS

Two classes of strains. The sequences of multiple class I and class II strains of *H. ducreyi* were previously published for the *dsrA* and *hgbA* genes (46). For the *ncaA* gene, sequences for six class II strains were available (46), whereas a sole class I sequence was provided by the genome sequence of strain 35000HP (4). To expand this data set, we determined the sequences of the *wecA*, *pal*, *murC*, *lspA2*, and *recA* loci for 9 class I and class II strains, and the *ncaA* locus for 6 class I strains of *H. ducreyi* (Table 1). This data set was further augmented by sequences of the *mtrC*, *sapA* and *fgbA* genes (3, 29, 35) kindly provided by Margaret Bauer. Thus the total data set included sequences at 11 unlinked loci for at least 10 strains with broad global distributions (Table 1); this number exceeds the standard of 7 loci used in conventional multilocus sequence typing (42) and is thus sufficient to illuminate the overall relationships among these strains.

Multiple alignments of the 11 genes provided a total of 10757 sites, of which 10122 (94.1%) were invariant (Table 2; Fig. 1). The 635 variant sites included 595 sites (93.7%) which solely distinguished class I strains from class II strains, wherein strains were monotypic within each class at these sites; 332 of these sites lay in the *dsrA* gene and 263 sites are distributed in the other 10 genes. Twenty-seven of the 40 other variant sites lay in the *dsrA* gene, and 13 were distributed across the *wecA*, *recA*, *sapA*, *murC*, *mtrC* and *hgbA* genes; the *pal*, *fgbA*, *lspA2* and *ncaA* genes lacked any variant sites beyond those that distinguished class I strains from class II strains (Fig. 1).

Each of the 11 genes supports the division of strains into 2 classes (Fig. 1). A parsimony tree was constructed using the 276 variant and informative sites at 10 loci, excluding the *dsrA* locus (see below). This tree shows strong separation of the two classes with very little homoplasy (Fig. 2); the consistency index of the tree is 0.997 with only a single site showing a reversion.
Moreover, class identity is preserved across all loci; in no case did a strain belong to one class at one locus and a different class at another locus, suggesting that gene flow between class I and class II strains is very low. Total variation within the classes is small (2 steps and 5 steps; Fig. 2A) compared to the discrimination between the classes (263 steps; Fig. 2B). Since members of each class were isolated over multiple decades and across multiple continents (Table 1), neither represents a single outbreak. Rather, the lack of variability within each class is consistent with recent periodic selection events purging variant alleles within each class.

Variation in levels of diversity. While alleles at all 11 loci examined distinguish between class I and class II strains (Figs. 1, 2), the magnitude of this difference varies between genes (Table 2). Previously published data showed large differences between class I and class II alleles at the dsrA and ncaA (46) loci. Inspection of Table 2 shows that these genes, as well as the lspA2 locus, have unusually high levels of divergence at both synonymous (K_S) and nonsynonymous (K_A) sites. In contrast, the majority of genes showed more modest differences between the classes, with divergence at synonymous sites being 10-fold lower for genes encoding cytoplasmic or periplasmic proteins than for the three aforementioned genes which encode surface-exposed proteins.

If high divergence between class I and class II strains at the lspA2, ncaA and dsrA genes were the result of a relaxation of selection, one would expect comparable divergence at synonymous (K_S) and nonsynonymous (K_A) sites (30). Similarly, if the high divergence were due to positive selection acting on spontaneous mutations, K_A would increase without a commensurate increase in K_S (48). However, since both K_A and K_S were elevated, and the K_S/K_A ratios of these genes are very similar to those of genes with low divergence, we posit that
recombination has introduced variant alleles into class I, class II or both sets of strains from a more distantly-related donor. This is not surprising as recombination is a common source of variant alleles at genes encoding surface-exposed proteins which are subject to positive, diversifying, or frequency-dependent selection (24, 47).

Recombination with the \textit{dsrA} gene of strain 35000HP. The 910 aligned sites within the \textit{dsrA} genes include 359 variant positions, 332 of which simply distinguish class I and class II strains; these sites are distributed across the length of the aligned region (class A sites, Fig. 3). The other 27 sites fall into 4 classes, all but one of which distinguish strain 35000HP from other class I strains. Notably, 25 of 26 of these sites are significantly clustered between bases 157 and 567 in the central region of the \textit{dsrA} gene. (P=1.6 \times 10^{-7}; \chi^2 test, 1 df); there is no enrichment for class A sites in this region. Setting aside the possibility that the 35000HP genome sequence suffers very poor quality at this locus in this region, these data strongly suggest that recombination has introduced the variant sites into the 35000HP \textit{dsrA} allele; it is for this reason we excluded \textit{dsrA} sequences when constructing phylogenies (Fig. 2). Given that the difference between the 35000HP alleles and other class I alleles is modest relative to the difference between alleles in class I and class II strains (Fig. 2), these data suggest that variability exists within class I strains, and that class I strains do recombine with each other. Therefore, class I and class II strains do not represent strictly clonal groups. Recombination may not be detectable among other class I strains owing to the paucity of variation among them.

Rearrangements. Beyond the substitutions outlined above, four of the loci examined herein show rearrangements between class I and class II alleles. The \textit{lspA2} gene shows a 184 bp
insertion/deletion between class I and class II alleles. This could represent a recombination event, either from externally acquired DNA or an internal rearrangement of the chromosome. LspA1, LspA2, and LspB are used by *H. ducreyi* to avoid phagocytic uptake and are critical virulence factors (43). The LspA1 and LspA2 predicted protein sequences are 86% identical to each other (44); the sequence amplified here for *lspA2* represents an area of the gene unique to the *lspA2* class I allele in 35000HP. Although class II *lspA2* sequences are more similar to class I *lspA2* sequences than they are to the class I *lspA1* gene, only a complete genome sequence will confirm the orthology of this gene. Second, the *fgbA* gene, which encodes another surface-exposed product, shows two in-frame insertion/deletion events of 174 and 21 bp (58 and 7 residues). Third, the *ncaA* alleles are homologous only at the extreme 3’ end; the entirely different 5’ portions of the class I and class II genes reflect illegitimate recombination with an unknown donor. Lastly, the 5’ ends of the *dsrA* alleles are also quite different, again bespeaking illegitimate recombination. An additional 1-bp deletion is evident in the coding region of the *mtrC* gene of class II strain 6644. In all cases, the recombinant regions were excluded from the calculations of substitution rates reported in Table 2. The appearance of these complex variants in genes encoding surface proteins supports the hypothesis that the encoded proteins experience selection for these differences.
DISCUSSION

Variant lineages were created by recombination. Although the data set presented herein is limited to few strains, it does shed significant light into the relationships between class I and class II strains of *H. ducreyi*. The 11 loci around the *H. ducreyi* chromosome fall into two groups: the *sapA, wecA, pal, mtrC, murC, recA, fgbA* and *hgbA* genes have low diversity, with an average $K_S$ between class I and class II strains of 0.052 (Table 2). In contrast, the unlinked *lspA2, ncaA* and *dsrA* genes are much more divergent, with an interclass $K_S$ ranging from 0.42 to 1.22.

These data are consistent with two scenarios (Fig. S3). First, one may consider the two classes to be very closely related, whereby the low diversity genes would typify the genetic differences between the strains. Here, multiple recombination events with outside strains introduced variant alleles into the genes encoding surface-exposed proteins. Alternatively, the *H. ducreyi* classes may be quite different, with the *dsrA, lspA2* and *ncaA* genes being typical of other chromosome genes. Here, the similarity at the remaining genes would indicate recent recombination between the two classes at these loci. We reject this alternative since the high variation in $K_S$ among high-diversity genes is more consistent with recombination with a variety of sources than with high intragenomic variation in substitution rates. While such variation could result from strong variability in levels of codon selection among genes (39, 40), these genes do not encode functions which show strong codon selection (38), and genes with high and low $K_S$ values are similarly weakly biased (Table 2). Therefore, we conclude that the two classes are relatively closely related, with variant alleles having been introduced by lateral transfer. While recombination could occur at any locus, recombinants would most likely be detected at loci experiencing strong selection for variant alleles, such as those encoding exposed proteins interacting with the immune system; recombination need not have occurred at all loci encoding
surface-exposed proteins (e.g., the $hgbA$ gene). Given the rate of synonymous substitution of 1% per million years (28, 31), the two classes diverged ~5 MYr ago, predating the separation of the human and chimpanzee lineages (21).

Are $H. ducreyi$ classes different species? The data shown above suggest that $H. ducreyi$ classes are genotypically distinct, recombination is not evident between the classes, while recombination is evident within class I. These properties have been used to delineate different bacterial species (7). It is currently not clear if class I and class II strains are phenotypically distinct in a significant manner; additional genome sequencing of a class II strain would shed light on this question. If so, then the two classes of $H. ducreyi$ may form two distinct species, rather than simply forming two clonal groups. When genes affected by recombination with outside genomes ($dsrA$, $ncaA$, $lspA2$) are ignored, the average divergence at synonymous sites between class I and class II strains of $H. ducreyi$ is 0.052 (Table 2). While this difference is modest when considering nonpathogenic species, different species of pathogens are often quite similar to sibling species as significant phenotypic differences in pathogenicity or host range motivate the taxonomic distinction. For example, $Mycobacterium tuberculosis$ is very closely related to $M. bovis$, with an average $K_S$ of only 0.003 (Table 3); the difference between the two classes of $H. ducreyi$ is more than 10 times that amount. There are a number of pathogens which are more closely related to sibling species than the two $H. ducreyi$ classes are to each other (Table 3). The long time of separation (5 MYr) for this sexually-transmitted species also suggests that the two classes of strains may have different lifestyles. A complete genome sequence of a class II strain will assist in the identification of other loci which distinguish these two taxonomic groups.
Epidemiologic Typing of *H. ducreyi*. Analysis of clonal complexes or clusters typically involves examination of a set of 7 genes under neutral selective pressure using multilocus sequence analysis (MLSA) (15, 42); the inclusion of multiple loci prevents one's conclusions from being influenced to a high degree by unusual evolutionary events that may affect an individual gene. Because six of the genes presented here, *dsrA*, *hgbA*, *ncaA*, *fgbA*, *pal*, and *lspA2*, encode gene products that interact with the host, they may experience selective pressures atypical of the genome as a whole. Therefore, we expanded the conventional dataset of 7 genes to a total of 11 genes, including genes encoding cytoplasmic and periplasmic proteins. Within this expanded dataset, all eleven genes supported the same population structure: *H. ducreyi* strains partition into two well separated groups with no evidence of recombination between them. Thus, this set – or even a subset – of genes would allow for reliable differentiation of clinical isolates and make strain typing for epidemiologic studies more feasible. Among the 11 genes analyzed here, *recA*, *sapA*, *wecA*, *mtrC* and *murC* encode periplasmic or cytoplasmic proteins likely to experience little selective pressure from the host and would be most reliable in predicting the class to which a novel strain of *H. ducreyi* would belong. Using MLSA to classify *H. ducreyi* strains offers several advantages, including the possibility to perform the assay with PCR-amplified material from clinical specimens (27), eliminating the need for culturing this fastidious organism. Moreover, the lack of recombination increases the clarity of class identification, allowing fewer loci to be used to achieve accurate typing.

Impact of diversity on vaccine development. The segregation of *H. ducreyi* strains into two distinct classes affects strategies of vaccine development. Potential vaccines must offer protection against both classes, yet the differences between them will likely complicate the
process of developing a single vaccine. For example, a monovalent HgbA vaccine has shown promise in a pig model, where immunization with HgbA₁ and Freund’s adjuvant provided protection against challenge with class I strains (1). However, immunization with HgbA₁ and monophosphoryl lipid A, an adjuvant approved for use in humans, resulted in protection against class I strains but not class II strains (13); this is not surprising, given the differences in the HgbA sequence in class I and II strains (46). Genes encoding surface proteins show large differences between class I and class II strains, including substitutions and insertion/deletion events, which can alter or eliminate epitopes. If, as we suggest above, the restriction of alleles to either class I or class II strains that we observe can be generalized to the entire genome, then development of a single vaccine that is maximally efficacious for both classes may be difficult or impossible to achieve.

ACKNOWLEDGMENTS

This work was supported by Allegheny College (EER and TLH), Grinnell College (NW, RC, and TLH), and NIH grant GM078092 (JGL). Stanley Spinola and Chris Elkins kindly provided strains. Margaret Bauer provided the fgbA mtrC, and sapA sequences. The authors would like to thank Carolyn Bosse (Grinnell College), Blair Gleeson (Allegheny College) and Aleksas Nerone (Allegheny College) for technical support and Cassie Majetic (Saint Mary's College) for thoughtful critique of the manuscript.
REFERENCES


17


Table 1. Strains used in this study.

<table>
<thead>
<tr>
<th>Strain</th>
<th>Class</th>
<th>Isolation</th>
<th>dsrA</th>
<th>fgbA</th>
<th>hgbA</th>
<th>lspA2</th>
<th>mtrC</th>
<th>murC</th>
<th>ncaA</th>
<th>pal</th>
<th>recA</th>
<th>sqpA</th>
<th>wecA</th>
</tr>
</thead>
<tbody>
<tr>
<td>35000HP</td>
<td>I</td>
<td>1975; Winnipeg, Canada</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>HD183</td>
<td>I</td>
<td>1982; Singapore</td>
<td>-</td>
<td>A</td>
<td>-</td>
<td>N</td>
<td>A</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>A</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>HD188</td>
<td>I</td>
<td>1982; Kenya</td>
<td>-</td>
<td>A</td>
<td>-</td>
<td>N</td>
<td>A</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>A</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>85-023233</td>
<td>I</td>
<td>1985; New York, USA</td>
<td>-</td>
<td>A</td>
<td>-</td>
<td>N</td>
<td>A</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>A</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>C111</td>
<td>I</td>
<td>Unknown; Nairobi, Kenya</td>
<td>-</td>
<td>-</td>
<td>A</td>
<td>N</td>
<td>-</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>-</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>6644</td>
<td>I</td>
<td>1989; Boston, USA</td>
<td>-</td>
<td>A</td>
<td>-</td>
<td>N</td>
<td>A</td>
<td>-</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>A</td>
<td>N</td>
</tr>
<tr>
<td>82-029362</td>
<td>I</td>
<td>1982; California, USA</td>
<td>-</td>
<td>A</td>
<td>-</td>
<td>-</td>
<td>A</td>
<td>N</td>
<td>N</td>
<td>-</td>
<td>-</td>
<td>A</td>
<td>-</td>
</tr>
<tr>
<td>HMC46</td>
<td>I</td>
<td>1995; Kenya</td>
<td>A</td>
<td>-</td>
<td>A</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>HMC56</td>
<td>I</td>
<td>1995; Dominican Republic</td>
<td>A</td>
<td>-</td>
<td>A</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>HMC60</td>
<td>I</td>
<td>1989; Florida, USA</td>
<td>A</td>
<td>-</td>
<td>A</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>SSMC57</td>
<td>I</td>
<td>Unknown; Bangladesh</td>
<td>-</td>
<td>A</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>DMC111</td>
<td>II</td>
<td>Unknown; Bangladesh</td>
<td>A</td>
<td>A</td>
<td>-</td>
<td>-</td>
<td>A</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>SSMC71</td>
<td>II</td>
<td>Unknown; Bangladesh</td>
<td>A</td>
<td>-</td>
<td>A</td>
<td>-</td>
<td>-</td>
<td>A</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>CIP542</td>
<td>II</td>
<td>1954; Hanoi, Vietnam</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>N</td>
<td>A</td>
<td>N</td>
<td>N</td>
<td>A</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>HMC112</td>
<td>II</td>
<td>1984; Unknown</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>N</td>
<td>A</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>A</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>DMC64</td>
<td>II</td>
<td>Unknown; Bangladesh</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>N</td>
<td>A</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>A</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>33921</td>
<td>II</td>
<td>Unknown; Nairobi, Kenya</td>
<td>A</td>
<td>A</td>
<td>-</td>
<td>N</td>
<td>A</td>
<td>N</td>
<td>A</td>
<td>N</td>
<td>N</td>
<td>A</td>
<td>N</td>
</tr>
</tbody>
</table>

1. An ‘A’ denotes a previously available sequence; an ‘N’ denotes new sequence generated in this study; a dash indicates no data available.
Table 2. Summary of nucleotide sequence information.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Position&lt;sup&gt;1&lt;/sup&gt;</th>
<th>Length&lt;sup&gt;2&lt;/sup&gt;</th>
<th>E&lt;sup&gt;3&lt;/sup&gt;</th>
<th>Substitution Rate&lt;sup&gt;4&lt;/sup&gt;</th>
<th>Location and Function</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>sapA</td>
<td>996773-998523 (59)</td>
<td>1683</td>
<td>0.468</td>
<td>0.0032 0.0199</td>
<td>Periplasm; peptide transport</td>
<td>(29)</td>
</tr>
<tr>
<td>wecA</td>
<td>1551061-1551456 (91)</td>
<td>398</td>
<td>0.432</td>
<td>0.0000 0.0207</td>
<td>Cytoplasm; synthesis of common antigen</td>
<td>This Paper</td>
</tr>
<tr>
<td>pal</td>
<td>1485485-1485823 (87)</td>
<td>367</td>
<td>0.770</td>
<td>0.0033 0.0250</td>
<td>Outer membrane; peptidoglycan-associated lipoprotein</td>
<td>This Paper</td>
</tr>
<tr>
<td>murC</td>
<td>654024-654714 (38)</td>
<td>694</td>
<td>0.561</td>
<td>0.0056 0.0310</td>
<td>Periplasm; UDP-N-acetylmuramate-L-alanine ligase</td>
<td>This Paper</td>
</tr>
<tr>
<td>recA</td>
<td>325798-326862 (19)</td>
<td>1073</td>
<td>0.693</td>
<td>0.0099 0.0433</td>
<td>Cytoplasm; recombination-promotion</td>
<td>This Paper</td>
</tr>
<tr>
<td>fgbA</td>
<td>136574-137285 (8)</td>
<td>572</td>
<td>1.144</td>
<td>0.0090 0.0490</td>
<td>Outer membrane; fibrinogen binding</td>
<td>(3)</td>
</tr>
<tr>
<td>mtrC</td>
<td>1258031-1259530 (74)</td>
<td>1505</td>
<td>0.547</td>
<td>0.0109 0.0561</td>
<td>Periplasm; efflux of antimicrobial peptides</td>
<td>(35)</td>
</tr>
<tr>
<td>hgbA</td>
<td>1686700-1689521 (99)</td>
<td>2850</td>
<td>0.467</td>
<td>0.0211 0.0851</td>
<td>Outer membrane; haemoglobin binding</td>
<td>(46)</td>
</tr>
<tr>
<td>tspA2</td>
<td>925194-925452 (54)</td>
<td>459</td>
<td>0.449</td>
<td>0.1705 0.4215</td>
<td>Outer membrane; antiphagocytic protein</td>
<td>This Paper</td>
</tr>
<tr>
<td>ncaA</td>
<td>1614393-1614596 (95)</td>
<td>203</td>
<td>0.613</td>
<td>0.0688 0.5863</td>
<td>Outer membrane; Collagen adherence</td>
<td>(46); this Paper</td>
</tr>
<tr>
<td>dsrA</td>
<td>603307-604195 (35)</td>
<td>910</td>
<td>0.726</td>
<td>0.0408 1.2232</td>
<td>Outer membrane; Serum resistance</td>
<td>(46)</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>10757</td>
<td></td>
<td>0.0112 0.0518</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

1. Position of orthologous region on the H. ducreyi 35000HP chromosome; numbers in parenthesis indicate minutes on a 0-100 scale.
2. Length of multiple alignment, including insertions relative to the H. ducreyi HP35000HP genome.
4. Divergence between typical class I and class II alleles; only regions with the specified protein-coding gene were analyzed.
5. Weighted average of the eight genes with K<sub>S</sub> < 0.1.
Table 3. Average divergence at synonymous sites among pairs of bacterial taxa.

| Taxon 1                  | Taxon 2                      | Number of genes | Mean $K_S$  
|--------------------------|------------------------------|-----------------|------------
| *Mycobacterium tuberculosis* H37Ra | *M. bovis* BCG              | 3704            | 0.003      
| *Yersinia pestis* CO92   | *Y. pseudotuberculosis* YPIII | 3404            | 0.014      
| *Bordetella pertussis* Tahoma I | *B. parapertussis* 12822     | 3156            | 0.029      
| *Klebsiella pneumoniae* 342   | *K. varicola* At-22         | 4665            | 0.033      
| *Escherichia coli* O157:H7 Sakai | *Shigella dysenteriae* Sd197 | 3066            | 0.049      
| *Haemophilus ducreyi* class I | *H. ducreyi* class II       | 8               | 0.052      
| *Bacillus anthracis* Ames  | *B. cereus* 03BB102         | 4646            | 0.067      

1. Average $K_S$ weighted by gene length.
2. Excluding the 3 genes in Table 2 with aberrantly high $K_S$ values.
**Fig. 1.** Variant and informative sites among 10 *H. ducreyi* genes. Sites that do not conform to the strict pattern discriminating class I and class II strains are indicated by gray boxes. Site positions, read vertically, are numbered from the beginning of the multiple alignment; the site of the first base of the multiple alignment in the *H. ducreyi* 35000HP genome is provided in Table 2.

**Fig. 2.** Maximum parsimony phylogeny of strain using variant and informative sites at 10 loci (data in Fig. 1). **A.** Relationships among Class I and Class II strains. **B.** Relationships among strains with all branches drawn to scale. Vertical marks denote changes.

**Fig. 3.** Distribution of variant and informative sites within the *dsrA* gene. Classes are defined as follows: **A.** Distinguish Class I and Class II; **B.** 35000HP is different from all other Class I strains; **C.** 35000HP is variant from other strains; **D.** 35000HP different from other Class I strains, but same as Class II (parallelism); **E.** Variant Class II. Horizontal bar indicates the regions containing 25 of 26 class B, C and D sites. ‘Count’ denotes the number of sites in each class.