Title

Helicobacter pylori diversification during chronic infection within a single host generates sub-populations with distinct phenotypes

Short title

Helicobacter pylori diversification within a single host

Authors

Laura K. Jackson[1,2], Barney Potter[3], Sean Schneider[2], Matthew Fitzgibbon[4], Kris Blair[1,2], Hajirah Farah[2,5], Uma Krishna[6], Trevor Bedford[2,3], Richard M. Peek Jr.[6], Nina R. Salama[1,2,5]

Affiliations

[1]Molecular and Cellular Biology Graduate Program, University of Washington, Seattle, WA
[2] Human Biology Division, Fred Hutchinson Cancer Research Center, Seattle, WA
[3]Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center, Seattle, WA
[4]Genomics & Bioinformatics Shared Resource, Fred Hutchinson Cancer Research Center, Seattle, WA
[5]Department of Microbiology, University of Washington School of Medicine, Seattle, WA
[6]Division of Gastroenterology, Department of Medicine, Vanderbilt University Medical Center, Nashville, TN

Abstract

Helicobacter pylori chronically infects the stomach of approximately half of the world's population. Manifestation of clinical diseases associated with *H. pylori* infection, including cancer, is driven by strain properties and host responses; and as chronic infection persists, both are subject to change. Previous studies have documented frequent and extensive within-host bacterial genetic variation. To define how within-host diversity contributes to phenotypes related to *H. pylori* pathogenesis, this project leverages a collection of 39 clinical isolates acquired prospectively from a single subject at two time points and from multiple gastric sites. During the six years separating collection of these isolates, this individual, initially harboring a duodenal ulcer, progressed to gastric atrophy and concomitant loss of acid secretion. Whole genome sequence analysis identified 2,232 unique single nucleotide polymorphisms (SNPs) across isolates and a nucleotide substitution rate of 1.3×10^{-4} substitutions/site/year. Gene ontology analysis identified cell envelope genes among the genes with excess accumulation of nonsynonymous SNPs (nSNPs). A dendrogram based on genetic similarity, clusters isolates from each time point separately. Within time points, there is segregation of subgroups with phenotypic differences in bacterial morphology, ability to induce inflammatory cytokines, and mouse colonization. Higher inflammatory cytokine induction in recent isolates maps to shared polymorphisms in the Cag PAI protein, CagY, while rod morphology in a subgroup of recent isolates mapped to eight mutations in three distinct helical cell shape determining (*csd*) genes. The presence of subgroups with unique genetic and phenotypic properties suggest complex selective forces and multiple sub-niches within the stomach during chronic infection.

Author Summary

Helicobacter pylori, one of the most common bacterial pathogens colonizing humans, is the main agent responsible for stomach ulcers and cancer. Certain strain types are associated with increased risk of disease, however many factors contributing to disease outcome remain unknown. Prior work has documented genetic diversity among bacterial populations within single individuals, but the impact of this diversity for continued bacterial infection or disease progression remains understudied. In our analysis we examined both genetic and functional features of many stomach isolates from a single individual infected over six years. During these six years the subject shifted from having excess acid production and a duodenal ulcer to lower acid production from gastric atrophy. The 39 isolates form sub-populations based on gene sequence changes that accumulated in the different isolates. In addition to having distinguishing genetic features, these sub-populations also have differences in several bacterial properties, including cell shape, ability to activate immune responses, and colonization in a mouse model of infection. This apparent functional specialization suggests that the bacterial sub-populations may have adapted to distinct sub-niches within the stomach during chronic infection.

Introduction

Helicobacter pylori is a bacterial pathogen that colonizes the human gastric mucosa of approximately half of the world's population [1]. Infections persist throughout life without intervention and can lead to gastric and duodenal ulcers, MALT lymphoma, and gastric cancer in a subset of individuals [2,3]. *H. pylori* exhibits marked genetic diversity

compared to other bacterial pathogens, can impact treatment efficacy and disease severity [4–7]. Typically, antibiotics and proton pump inhibitors are employed for treatment, but variable prevalence of antibiotic resistance across populations make implementing a single treatment regimen difficult [8,9]. Strain-specific genotypes also contribute to increased disease risk within distinct ethno-geographic populations. Individuals with strains carrying the Cag pathogenicity island (Cag PAI), encoding a type IV secretion system (T4SS) and effector toxin CagA, have an increased risk of gastric cancer [10,11]. Cag PAI encoded genes, *cagA* and *cagY* exhibit significant allelic variation between individuals and have been identified as targets of positive selection within the global population [12]. Both CagA and CagY have been shown to modulate the host inflammatory response [13,14]. Recombination events within *cagY*, which encodes a structural component of the Cag T4SS with homology to the VirB10 component of other T4S systems, modifies secretion of inflammatory cytokines from epithelial cells [15,16]. CagA alters host responses through its interaction with intracellular kinases leading to the activation of the NFκB pathway [17,18]. In addition to Cag PAI genes, certain alleles of vacuolating cytotoxin, *vacA*, and frequent phase variation as well as recombination mediated gain and loss of outer membrane protein (OMP) adhesins BabA, SabA, and HopQ, have been linked to strain differences in pathogenesis [19–22].

Several mechanisms promote genomic diversification. Although *H. pylori* does encode several transcriptional regulators, much of gene regulation occurs through genomic alterations [23]. *H. pylori* has several phase variable genes whose expression is altered due to slipped strand mispairing in homo-polymeric tracts [24,25]. In addition,

H. pylori has a somewhat elevated baseline mutation rate compared to other bacteria; 10^{6}-10^{8} substitutions/site/generation compared to the 10^{9} substitutions/site/generation reported for *Escherichia coli* [26–28]. This is due to absent mismatch repair genes and deficiencies in the exonuclease domain of Pol1 [29]. However, base-excision repair is robust, preventing hypermutator phenotypes [30]. Variation is largely driven by high rates of intra and inter-genomic recombination. Intragenomic recombination can alter protein expression via gene conversion among paralogous families of outer membrane proteins [25]. Additionally, as a naturally competent bacterium, *H. pylori* incorporates DNA from genetically distinct strains into its chromosome, further varying gene content and sequence [31,32].

The human stomach is the only known niche for *H. pylori;* therefore, the breadth of genomic diversity across global populations likely reflects adaptation to individual host stomach environments [33]. More recently, genetic diversity within a single host has also become appreciated, suggesting the existence of sub-niches within the stomach with distinct selective pressures [34–37]. *H. pylori* can colonize the epithelial surface of the inner gastric mucus layer, form cell adherent microcolonies, and penetrate into the gastric glands in both the antrum and corpus (Fig. 1) [38,39]. The antrum and corpus have distinct gland architecture and cell type composition, providing unique challenges to bacterial survival. Gastric environments also change during lifelong infection. While most acute infections start in the antrum where the pH is closer to neutral, *H. pylori* can expand into the corpus [40,41]. Changes in bacterial localization are associated with histologic changes, including loss of the parietal cells (gastric atrophy), a risk factor for the development of gastric cancer [42]. These changes are

accompanied by fluctuations in immune responses and alteration of glycosylation patterns affecting OMP-receptor binding to cell surface and mucus [24,43,44].

Prior studies of within-host genetic variation have observed signatures of diversifying selection in support of selective pressures during chronic stomach colonization [7,37,45]. However, phenotypic variation in infecting populations over time has not been well studied. To define both genetic changes that occur during infection and their functional consequences, we leveraged *H. pylori* isolates from a single individual collected at two time points spanning 6 years (1994-2000). One of multiple isolates obtained from culture of a single antral biopsy in 1994, J99, has previously been sequenced and a complete reference genome is available [46]. This same individual, who had not been successfully treated for *H. pylori*, underwent a repeat endoscopy performed in 2000 from which single biopsies from the corpus, antrum and gastric metaplasia in the duodenum were cultured and additional *H. pylori* isolates recovered. A subset of isolates from the second time point (yr 2000) were analyzed by PCR microarray as part of a study highlighting diversity of isolates from a single individual [34].

Here we combined whole genome sequence analysis of multiple isolates from this subject with extensive phenotypic characterization to explore the rates and extent of genetic and phenotypic diversification within a single host. In this individual, during six years of chronic colonization, isolates adapted to occupy at least two distinct niches within the stomach reflected by differential ability to colonize a mouse model. We identified the genetic basis for modulation of Cag-dependent inflammatory cytokine induction and morphologic diversification. Neither of these phenotypes fully account for

143 the differences in colonization of the mouse model, highlighting the multifactorial
144 selection pressures operant during chronic stomach colonization.
145
146 **Results:**
147
148 **Whole genome sequencing detects within-host genetic diversification of bacterial**
149 **populations**.
150 For this study we analyzed 39 isolates from two distinct sampling time points. At the
151 time of the original biopsy (yr 1994), the source individual had a duodenal ulcer,
152 indicative of *H. pylori* infection localized to the antrum and consistent with recovery of
153 multiple single colonies from the single antral biopsy processed for culture. Six years
154 later (yr 2000), after refusing antibiotic therapy, additional single colony isolates were
155 collected from distinct biopsy sites. At this time, this individual had corpus predominant
156 gastritis and signs of gastric atrophy, including decreased production of stomach acid,
157 indicating the spread of infection to the main body of the stomach (Fig. 1) [34,47].
158 Twelve ancestral isolates, including *H. pylori* strain J99, were all collected in 1994 from
159 the antral biopsy. From the second time point (recent, yr 2000), we analyzed 27 isolates
160 from the antrum (n=12), corpus (n=12), and duodenum (n=3).
161

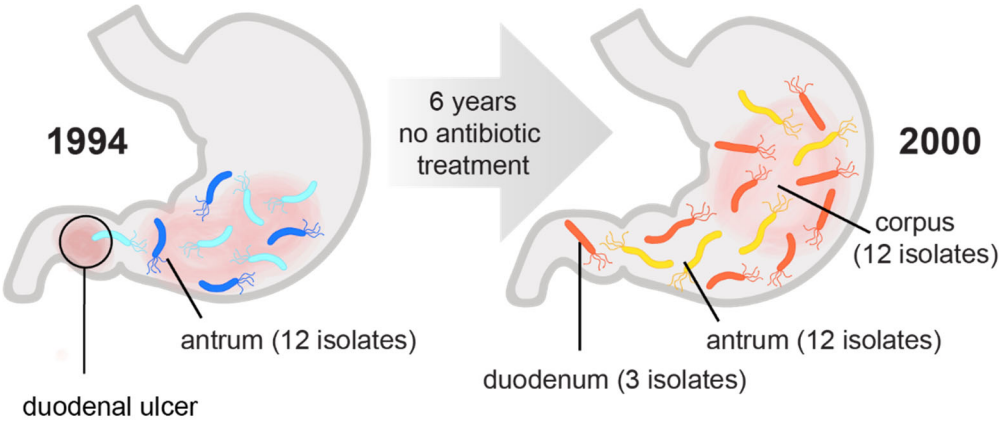


Figure 1. Sampling utilized to characterize genetic and phenotypic diversification of infecting population over six years.

This study leverages a collection of *H. pylori* isolates obtained from a treatment-naïve subject initially presenting with a duodenal ulcer at two different time points over a 6-year period of infection. A total of 12 isolates were analyzed from a single antral biopsy in 1994, and a total of 27 isolates were analyzed from single corpus, antrum, and duodenum biopsies collected in 2000 as indicated. In 2000 the subject displayed corpus atrophic gastritis and elevated stomach pH.

In order to measure the genetic diversity of *H. pylori* populations both within each time point and between time points, we performed whole genome sequencing using Illumina MiSeq. Sequences were aligned using the published sequence of J99 as the reference (AE001439). All isolates shared 99.99% average nucleotide sequence identity (ANI) to the reference strain J99. By comparison, J99 shares 92% ANI with strain 26695, originating from a distinct individual and geographic region [46]. High ANI among the isolates in the collection is consistent with a single diversifying strain population rather than mixed infection with genetically distinct strains. Unique SNPs, and insertion

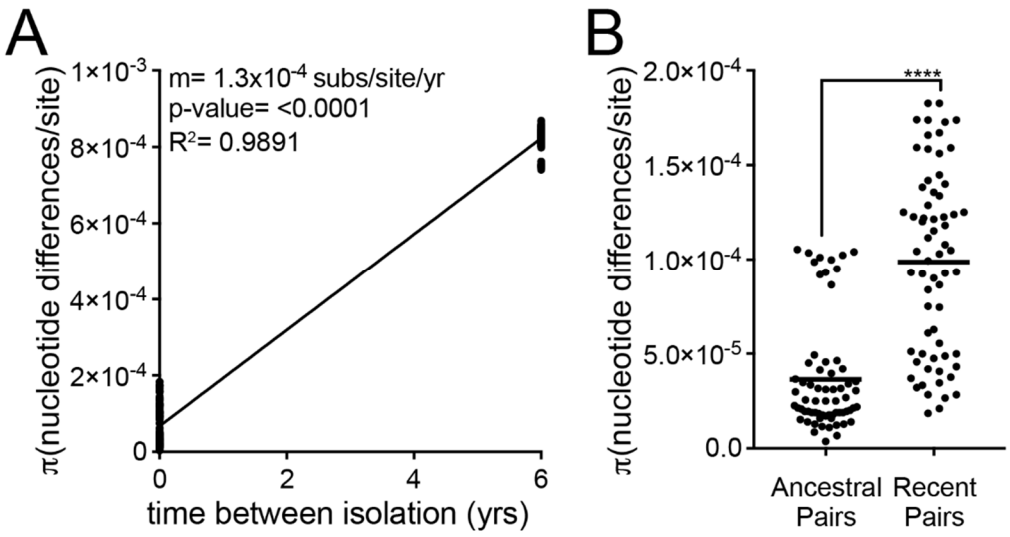
and deletion (indel) events detected in the collection are reported in Table 1. In total, 2,232 SNPs and 573 indels were identified (Table 1, Table S1a-b). This sequence variation represents changes introduced by both de-novo mutation and recombination. SNPs were distributed proportionally across coding and intergenic regions. By contrast, indels were biased towards intergenic regions (chi-squared, p-value<0.0001). Depletion of indels within coding regions, likely reflects purifying selection due to high potential of indels to introduce frameshifts, disrupting gene function. Additionally, the ratio of unique nonsynonymous SNPs (nSNPs) to synonymous SNPs (sSNPs) detected is close to one (0.98) despite higher number of synonymous sites within the genome. Of the total unique SNPs and indels detected (n= 2,805), 791 were shared between ancestral and recent populations, while 263 and 1,751 were exclusively found within the ancestral and recent populations, respectively. The high number of mutations unique to recent isolates demonstrates the substantial population divergence that occurred in this patient over time.

Table 1. Summary of unique SNPs, Indels detected by WGS among all the isolates (n=39) and in the subset of recent isolates (n=27, yr 2000)

	Total[a]	Coding[a]	nS[b]	S[b]	Intergenic[a]
Total SNPs	2,232	2,058	1,018	1,040	174
Recent SNPs	1,379	1,270	536	734	109
	Total	**Coding[a]**			**Intergenic[a]**
Total Indels	573	359			214
Recent Indels	372	231			141

[a]Unique events are labeled as either coding or intergenic. [b]Events within coding regions are further subdivided into nonsynonymous (nS) or synonymous (S) categories.

196 To assess extent of genetic diversity within the collection, we calculated the
197 average pairwise genetic distance (π) for unique pairwise comparisons of isolates from
198 the same time point (within time point) to isolates from different time points (between
199 time point). For between time point comparisons, only antral isolates (n=24) were used
200 to reduce potential confounding effects introduced from comparing isolates from
201 different anatomical locations. The average genetic distances (π, nucleotide
202 differences/site) of within time point pairs was 6.75×10^{-5}, while π of between time point
203 pairs was 8.23×10^{-4}, indicating within host evolution with an average molecular clock
204 rate of 1.3×10^{-4} substitutions/site/year (Fig. 2a). Overall, recent antral isolates have
205 increased diversity ($\pi = 9.9 \times 10^{-5}$) compared to the ancestral isolates ($\pi = 3.6 \times 10^{-5}$),
206 demonstrating accumulation of genetic diversity during chronic infection (Fig. 2b).



207

208 **Figure 2. Pairwise comparisons show within host diversification with increasing**
209 **diversity of infecting populations over time.** (**A**) Plot shows genetic distance for all
210 pairwise comparisons of antral samples isolated within the same time point (n=132, time
211 between isolation=0 yrs) and samples isolated from different time points (n=144, time

between isolation=6 yrs). Each point is a unique pairwise comparison (n=276). Linear regression with the slope (m) as the estimation of the molecular clock rate, p-value derived from F-test, and correlation coefficient (R^2) shown. (**B**) Each point represents a pairwise comparison between antral isolates within the ancestral population (yr 1994, n=66) or recent population (yr 2000, n=66). The average values between all pairwise comparisons in the population (π statistic) is shown with a black bar. Significance was determined using a Student's t-test (****, $p<0.0001$).

Identification of genomic regions enriched for within-host genetic variation

To identify regions of the genome that accumulate within host variation, we examined enrichment of nonsynonymous SNPs (nSNPs) in specific genes and functional classes assigned by the microbial genome database (MGDB) add reference. Out of the 1,495 genes in the reference sequence, 931(62.2%) are annotated with a functional class (Fig. 3a). Enrichment or depletion was determined by comparing the distribution of nSNPs among MGDB classes to expected values based on a normal distribution (Fig. 3, Table S2). We observed cell envelope genes, including OMPs, accumulated a disproportionate number of nSNPs in both the total dataset of unique nSNPs and the subset of nSNPs unique to recent group of isolates (Fig. 3b-c, Table S2). These results are similar to what others studying within-host variation have found. Accordingly, cell envelope diversification may serve a selective advantage in both the acute phase of infection as a mechanism of adaptation to a specific host and in the chronic phase of infection as a mechanism to persist in changing host environments.

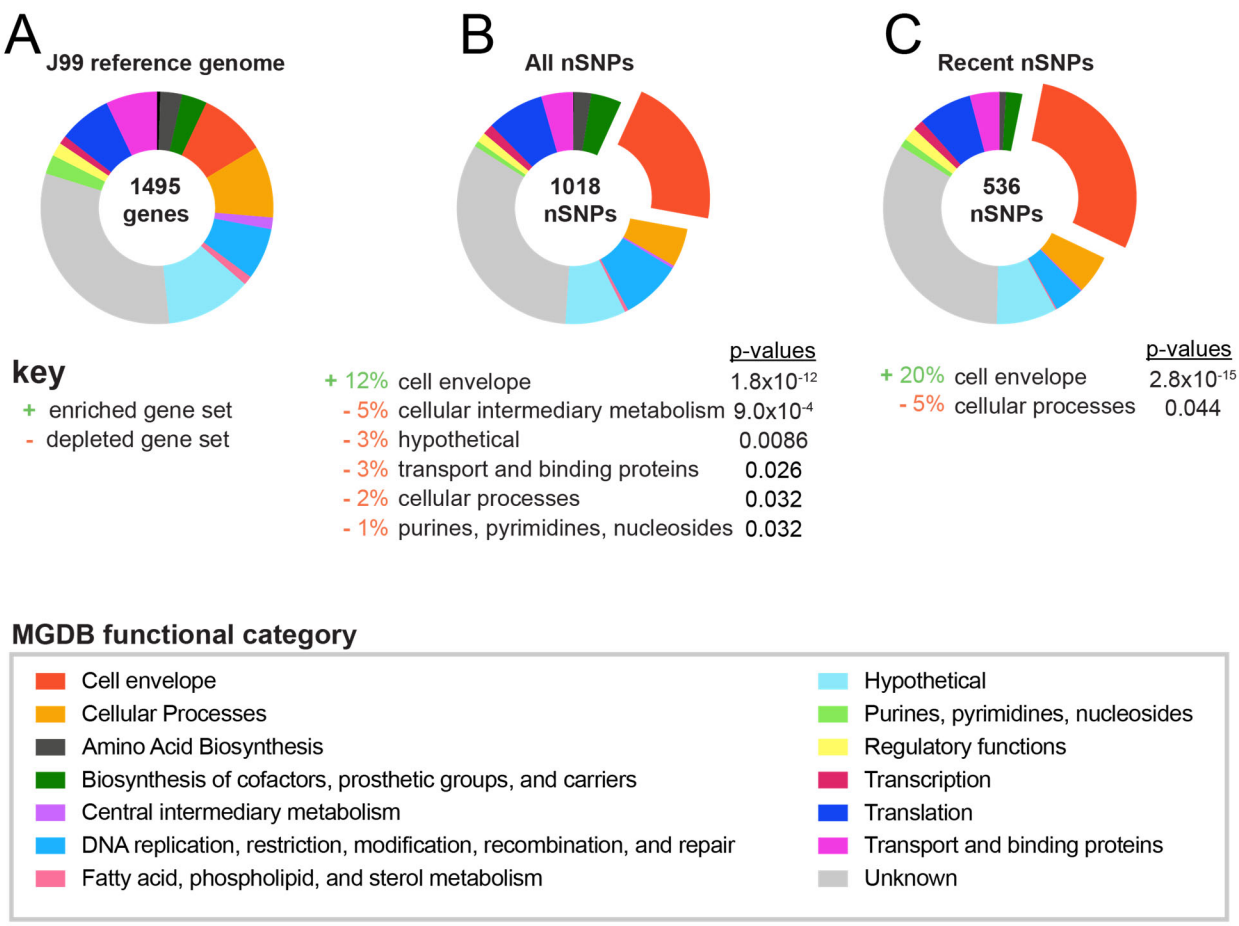


Figure 3. Cell envelope genes accumulate genetic variation during chronic infection. (**A**) Proportion of genes within the reference genome (J99) comprising each of 13 functional classes identified in the Microbial Genome Database and color-coded according to key [48]. Percentage of genes with unknown function are labeled in gray. (**B-C**) Proportion of nSNPs from (**B**) the entire dataset (all nSNPS) and (**C**) from the subset unique to recent isolates (recent nSNPs) that fall within each functional class. Categories with statistically significant enrichment or depletion are listed below each chart with associated percentages and p-values. Fisher's exact tests were used to determine significance and corrected for multiple testing using Benjamini and Hochberg false discovery rate methods.

Next, individual genes acquiring the most genetic variation over the six year period were identified. The number of nSNPs unique to the recent isolates detected for each gene were counted and weighted according to the gene length. The genes most highly enriched are listed in Table 2 (Table S3). Many of the genes identified encode OMPs (*babA*, *sabA*, *sabB*, and *hopQ*) that play roles in adhesion and exhibit variation between and within hosts [49,50].

Table 2. Genes with excess accumulation of nSNPs during chronic infection

Gene ID	Annotation	MGDB function	Total nSNPs	Z-scores[a]
jhp1300		Unknown	18	21.52
jhp1103	*hopQ*	Outer membrane protein	45	14.89
jhp1068	*birA*	Biotin protein ligase	11	10.96
jhp0659	*sabB*	Outer membrane protein	27	8.94
jhp0303		Hypothetical	3	7.90
jhp1096	*glnP_1*	Glutamine ABC transporter permease	8	7.73
jhp1409		Unknown	44	7.39
jhp0302	*argS*	Arginine-tRNA ligase	18	6.98
jhp0336		Unknown	19	5.07
jhp1097	*glnP_2*	Glutamine ABC transporter permease	5	4.63
jhp0634		Unknown	7	4.29
jhp1102		Guanine permease	9	4.26
jhp0833	*babA*	Outer membrane protein	15	4.16
jhp0662	*sabA*	Outer membrane protein	13	4.12
jhp0929		Unknown	3	4.10

[a]The top 15 annotated genes with Z-scores > 4 for number of nSNPs uniquely acquired in the yr 2000 group of isolates within single genes are shown. [b]Z-scores were calculated using number of counts per gene normalized according to gene length.

The MGDB does not specifically analyze antibiotic resistance genes. While this subject had no known history of antibiotic treatment, the presence of antibiotic resistance to metronizidole, ampicillin, clarithromycin, was previously tested. Four isolates, three antral and one duodenal, were resistant to clarithromycin due to a

mutation in the 23S rRNA gene, but all the other isolates were sensitive to all three [34]. To validate, we queried our sequence data for mutations known to confer antibiotic resistance, but no mutations indicative of additional antibiotic resistance were discovered.

Genomic diversity within the recent population is not driven by stomach region specific adaptation

To display the genetic relatedness of all the isolates in the collection, a similarity dendrogram was generated with the SNP data using the Nextstrain platform (Fig. 4a, [51]). This depiction represents a dendrogram rather than a phylogeny as SNPs included in these analyses could arise from recombination and therefore are not necessarily clonally derived [52]. Isolates collected from the two separate time points (1994 and 2000) cluster into distinct groups on the dendrogram with a long branch representing an average divergence of 7.8×10^{-4} substitutions/site between the two populations.

In addition to genetic divergence of *H. pylori* populations between time points, we observed substantial diversity within time points (Fig. 2b). Isolates collected from the most recent time point originated from biopsy samples from distinct stomach regions, allowing us to examine if region specific adaptation drives subgroup formation. To assess this, π for all the pairs isolated from the same source biopsy (within region) was compared to π from all the pairs from different source biopsies (between region). Although we hypothesized that isolates from the same source biopsy would be more similar, we instead found between region pairs have the same level of diversity as within

region pairs (Fig. 4b). We also did not find any specific SNPs or indels associated with isolates from either the antrum or corpus (Table S4a-d). This suggests subgroup differentiation within time points is not defined by stomach region adaptation in this patient. Consistent with this finding, nearest neighbors on the dendrogram often come from different biopsies (Fig. 4a).

We defined four distinct subgroups within the collection based on shared genetic characteristics (Fig. 4c). While the majority of isolates within the ancestral group are highly related, one isolate, SC4, is more divergent and clusters separately on the dendrogram. The average pairwise genetic distance between SC4 and each ancestral isolate is 173 nucleotide differences whereas the average pairwise genetic distance among all other unique pairs of ancestral isolates is 39 nucleotide differences. Thus, we named two subgroups of the ancestral isolates according to this divergence (1A and 1B). Within the recent group, there is additional clustering of the isolates into two subgroups, named 2A and 2B. Group 2A, is comprised of 12 total isolates with 147 unique mutations (SNPs and indels) and group 2B is comprised of 15 total isolates with 216 unique mutations. Both recent subgroups (yr 2000) contain isolates from all three biopsy locations. The formation of distinct subgroups within a population of isolates collected from a single time point, suggests the possibility of niche level adaptation, but these sub-niches must be present in all regions of the stomach sampled.

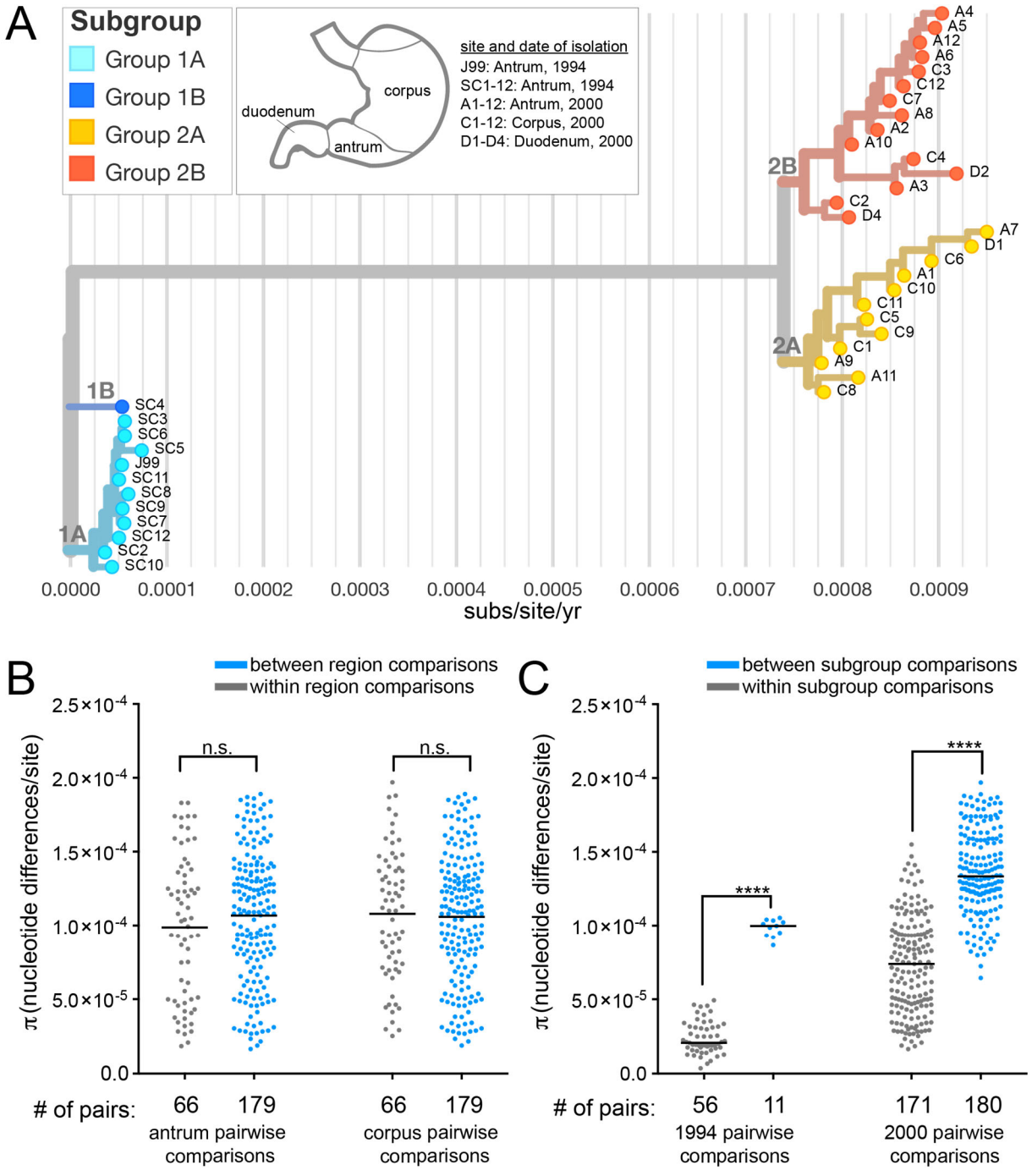
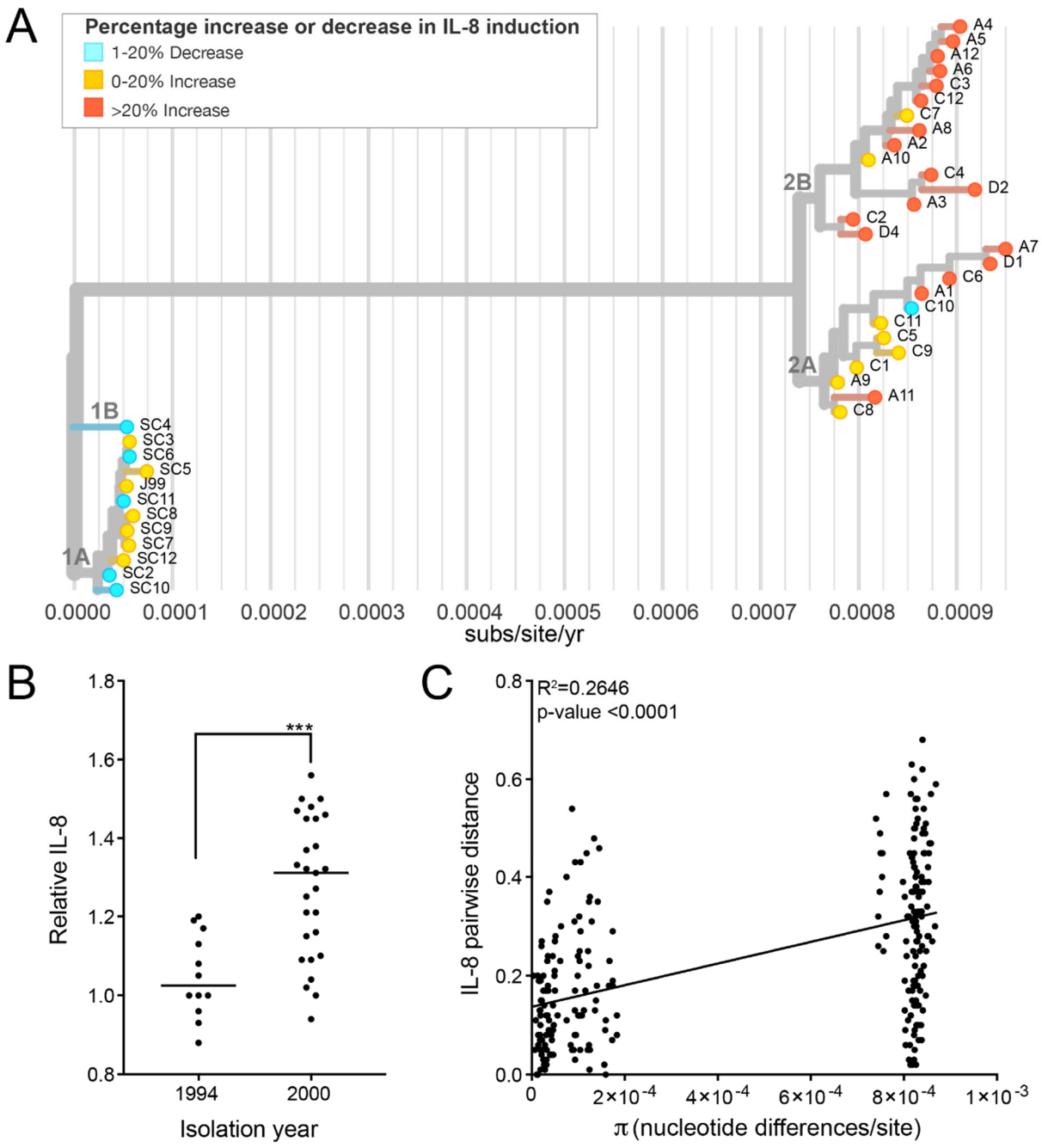


Figure 4. Clustering of strains by genetic similarity suggests distinct subgroups that do not correlate with biopsy site.

(**A**) An isolate dendrogram was generated from all SNPs in the collection with Nextstrain [51]. Isolates were named according to the anatomic region of their source biopsy as

307 annotated in the key and branch coloring was added to distinguish genetically related
308 subgroups within the collection. Light and dark blue isolates are from 1994 (1A, 1B,
309 respectively). Orange (2A) and red isolates (2B) are from 2000. The X-axis shows the
310 number of substitutions (subs)/site/year.(**B-C**) Point on the plot represents the pairwise
311 genetic distances calculated for groups of isolates described with a black bar
312 representing the mean (π). (**B**) Pairwise comparisons of recent isolates within the same
313 stomach regions (gray) have the same average genetic distance as pairwise
314 comparisons of isolates from different stomach regions (blue) in both the antrum and
315 corpus. (**C**) Pairwise comparisons of isolates within subgroups displayed on the isolate
316 dendrogram (gray) have smaller genetic distances on average than pairwise
317 comparisons of isolates from different subgroups (blue) from the same time point.
318 Significance was determined using a Student's t-test (****, $p<0.0001$).
319
320 **Recent *H. pylori* isolates have increased proinflammatory activity driven by *cagY***
321 **genetic variation.**
322 Substantial genetic divergence of *H. pylori* populations over this six year period of
323 infection, coupled with enrichment of mutations in genes related to virulence, prompted
324 exploration of pathogenic phenotypes. First, we tested ability of each strain to initiate an
325 inflammatory response. Each of the 39 isolates was co-cultured with a gastric epithelial
326 cell line (AGS) for 24hrs (MOI=10) and the release of inflammatory cytokine interleukin-
327 8 (IL-8) was measured in the supernatants. The J99 ancestral strain and J99 *ΔcagE*, a
328 mutant that blocks assembly of the Cag T4SS, were used as controls in each
329 independent experiment. All isolates were Cag PAI+ and induced IL-8 at levels above

330 J99 ⊿*cagE*. However, isolates from the most recent time point on average induced more
331 IL-8 compared to ancestral isolates (Fig. 5a-b). There was some heterogeneity in this
332 phenotype with the least inflammatory isolates inducing 12% less and the most
333 inflammatory isolates inducing 56% more IL-8 than J99 (Fig. 5b, Fig. S1). Isolates with
334 similar IL-8 phenotypes clustered together on the dendrogram; and comparison of the
335 genetic and phenotypic distances between unique pairs of antral isolates from both time
336 points (n=276) showed genetic divergence correlates with phenotypic divergence in
337 induction of IL-8 (Fig. 5c). These data show that isolates able to induce more
338 inflammation persisted, indicating a possible adaptive advantage of pro-inflammatory
339 activity during chronic infection in this patient.



340

341 **Figure 5. Isolates from recent time point have increased induction of IL-8**

342 **secretion during co-culture with gastric epithelial cells.** (**A**) Isolate dendrogram

343 from Fig. 4a overlaid with IL-8 induction phenotype of each isolate after 24 hours of co-

344 culture (MOI=10) with gastric epithelial cell line (AGS). Leaf colors represent percent

345 increased or decreased IL-8 secretion relative to ancestral isolate J99. (**B**) Each point

346 shows mean value of IL-8 detected in the supernatants of infected AGS cells relative to
347 J99 for independent isolates. The mean value was calculated from at least two
348 experiments with triplicate wells. Black line represents the mean values from each
349 subset of isolates (yr 1994, yr 2000). Significance was determined using a Student's t-
350 test (****, p<0.0001). (**C**) Comparison of genetic (π) and phenotypic (relative IL-8
351 secreted) distances between unique pairs of antral isolates from both time points
352 (n=276) is shown. Plot shows a linear regression with p-value derived from F-test and
353 correlation coefficient (R^2) reported.
354
355 To investigate the genetic basis of shared IL-8 phenotypes, we focused on
356 nSNPs that occurred within the Cag PAI. While there was no enrichment of nSNPs
357 within the Cag PAI as a whole (chi-squared, p-value>0.999), we did see enrichment in
358 two Cag PAI genes, *cagY* and *cagA* (Fig. S2a, Table S3). Recent isolates had 7 unique
359 nSNPs in *cagA*, however none were localized to known functional domains (Fig. S2b)
360 [53]. Several nSNPs were detected within the middle repeat region of *cagY*. This
361 domain contains a series of long and short direct repeat sequences that can undergo
362 recombination resulting in expansion or contraction of repeats. This can attenuate or
363 enhance Cag T4SS-dependent IL-8 secretion. In animal models of infection,
364 recombination events that diminish pro-inflammatory activity are dependent on adaptive
365 immunity [15]. Due to the difficulties in precisely mapping these recombination events
366 with short-read WGS data, we utilized restriction fragment length polymorphism (RFLP)
367 together with Sanger sequencing to identify unique alleles of *cagY* within the collection
368 (Fig. 6a-b, Fig. S2c). All of the recent isolates (Groups 2A and 2B) and ancestral isolate

SC4 (Group 1B), share the same RFLP pattern, which is distinct from the RFLP pattern shared by the other ancestral isolates (Fig. 6a, Fig. S3a). Sanger sequencing revealed that all the isolates in group 1A, including J99, share the same sequence. However the allelic variant of *cagY* in SC4 is distinct from that in the recent isolates (Fig. S3b). The SC4 *cagY* allele carries two mutations shared with recent isolates, including one that introduced a restriction site seen by RFLP, however it also harbors two unique mutations not found in any other isolates in the collection. All recent isolates (Group 2A and 2B) have 9 nSNPs total compared to the J99 *cagY* allele including the two shared with SC4 (Fig. 6b, Fig. S3c). None of the alleles had expansion or contraction of the number of repeats, but likely arose from gene conversion from sequences within other repeats (Fig. S3c). In order to test for a functional link between the variation in *cagY* and the differences in IL-8 phenotype, we performed an allelic exchange experiment, replacing the *cagY* allele in J99 ancestral strain with the two other *cagY* allelic variants (Fig. 6c). Co-culture of these engineered strains with AGS cells showed that the *cagY* allele shared by the recent isolates, confers the increase in induction of IL-8 at 24hrs. The SC4 allele in the J99 genomic context induced similar levels of IL-8 secretion as J99. Therefore, modulation of T4SS function can occur through the introduction of specific point mutations in the absence of expansion or contraction of the *cagY* repeats. The same experiment was performed with *cagA* variants, but IL-8 induction was not significantly different from J99 (Fig. S2c).

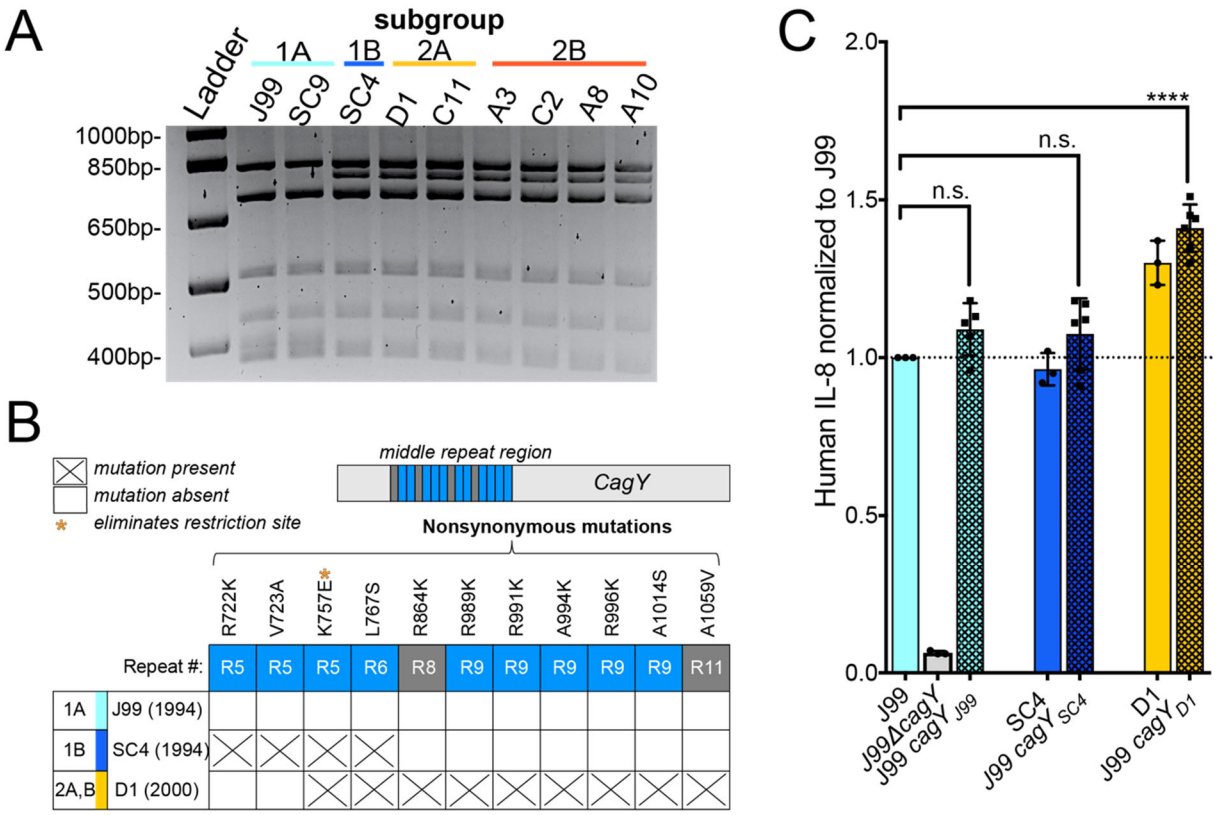


Figure 6. *cagY* genetic polymorphisms promote enhanced IL-8 secretion.

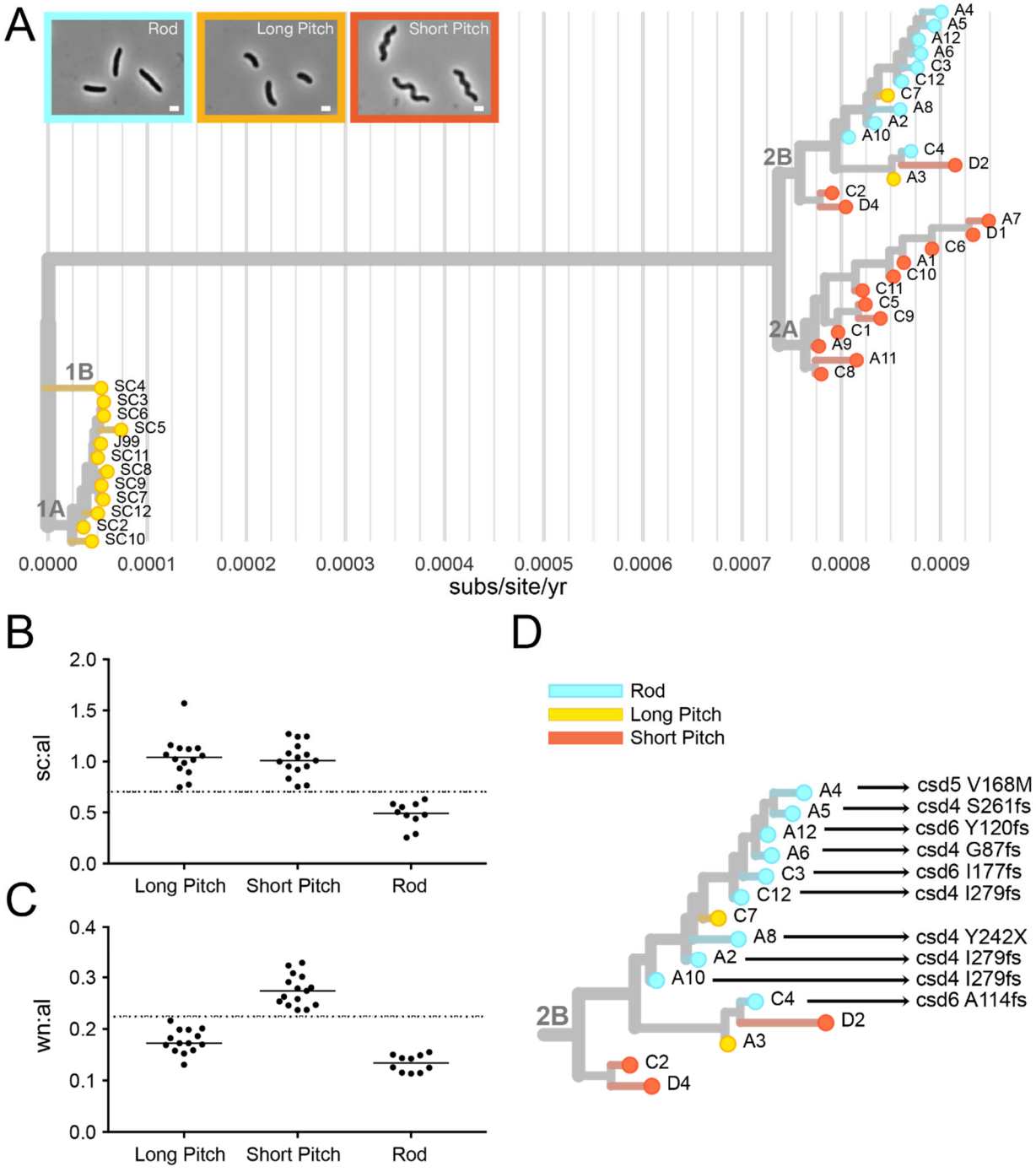
(A) RFLP analysis of amplified *cagY* repeat region from representative isolates digested with restriction enzyme DdeI reveals two distinct patterns within the populations. Isolates are colored by subgroup as in Fig. 4a. (**B**) Amino acid polymorphisms for the three different allelic variants of *cagY* in our isolate collection detected by Sanger sequencing. Isolates listed represent the three alleles (J99, SC4, D1) with date of isolation and subgroup(s) indicated. All nonsynonymous mutations detected map within the middle repeat region of *cagY* (15 total repeats; short in blue and long in gray). (**C**) Levels of IL-8 produced by *cagY* allelic exchange strains relative to J99 ancestral (y-axis line=1) 24hrs post infection of AGS cells (MOI=10). Data points represent averaged values from triplicate wells from at least 3 independent biological replicates.

Significance was determined with a one-way ANOVA with Dunnett's corrections (n.s. not significant, **** p<0.0001).

Sub-populations within the collection have distinct bacterial cell morphologies. Cell morphology has also been linked to virulence in *H. pylori* [54]. In order to measure cell morphology we used CellTool, a program which takes 2-D phase contrast images and measures quantitative cell shape parameters from cell outlines [54]. Based on these measurements, isolates were divided into three phenotypic shape categories—short pitch, long pitch, and rod. Short pitch isolates have increased wavenumber per unit centerline axis length compared to the long pitch and rod isolates. Rod isolates have decreased side curvature per unit centerline axis length compared to the short and long pitch isolates (Fig. 7b-c). Isolates with similar shape phenotypes cluster on the dendrogram (Fig. 7a, Fig. S4a-b). Interestingly, all ten rod-shaped isolates from group 2B had frameshift mutations in *H. pylori* cell shape determining (*csd*) genes known to cause rod-shape morphology when deleted. We observed four unique mutations in *csd4*, one unique mutation in *csd5*, and three unique mutations in *csd6*, one of which was shared by three isolates (Fig. S5). Thus, group 2B appears to have convergent evolution leading to straight-rod morphology (Fig. 7d).

Pairwise comparisons show that recent isolates within the same cell shape phenotype category are more genetically similar than recent isolates from different cell shape categories (Fig. 8a). Additionally, plots of the genetic and phenotypic distances between unique pairs of recent isolates from both time points (n=351) showed that genetic divergence positively correlates with phenotypic divergence both in

424 wavenumber and size curvature per unit axis length (Fig. 8b-c). This indicates a
425 signature of selection for loss in helical shape within this sub-population of recent
426 isolates.



427

Figure 7. Cell morphology varies among genetically distinct subgroups. (**A**) Cell shape phenotype clustering on isolate dendrogram. Leaves colored to indicate cell morphology phenotype of each isolate with rods in blue, short pitched in yellow, and long pitched in orange. Representative phase contrast micrographs of each morphologic class shown. (magnification=100x, scale bar = 1 μm). (**B-C**) Cell shape parameters calculated from 2-D phase images with CellTool software for isolates with indicated cell morphologies. Individual points represent mean values for measurements taken from >100 cells/isolate. Side curvature and wavenumber values were normalized by cell centerline axis length. (**B**) Mean side curvature values normalized by centerline axis length (sc:al) is decreased in rod shaped cells (<0.7, as indicated by y-axis line) and (**C**) wavenumber normalized by centerline axis length (wn:al) is increased for cells that have increased wavenumber (>0.225, as indicated by y-axis line). **(D)** Subgroup 2B labeled with amino acid mutations in cell shape determining genes (*csd4, csd5, csd6*).

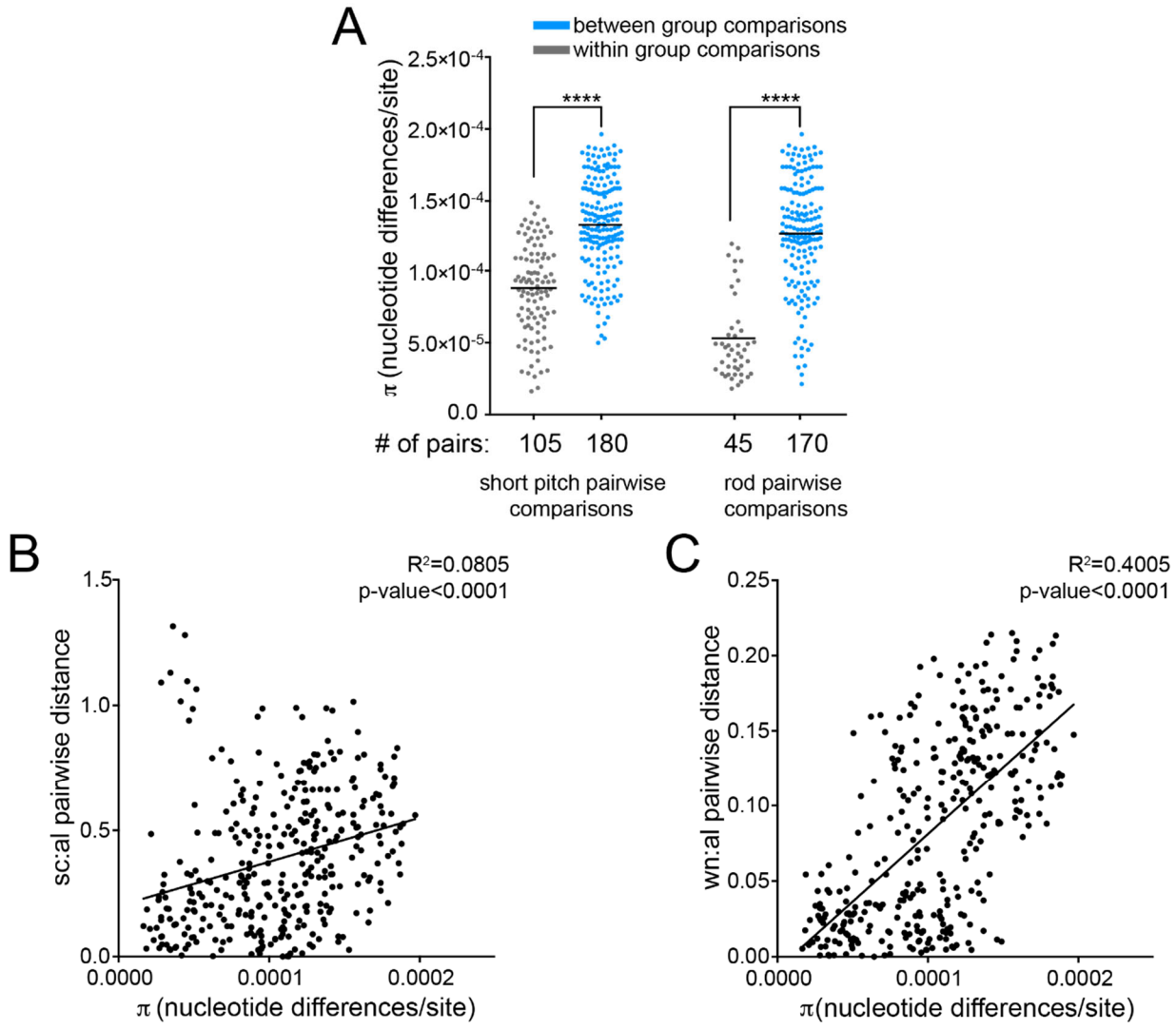


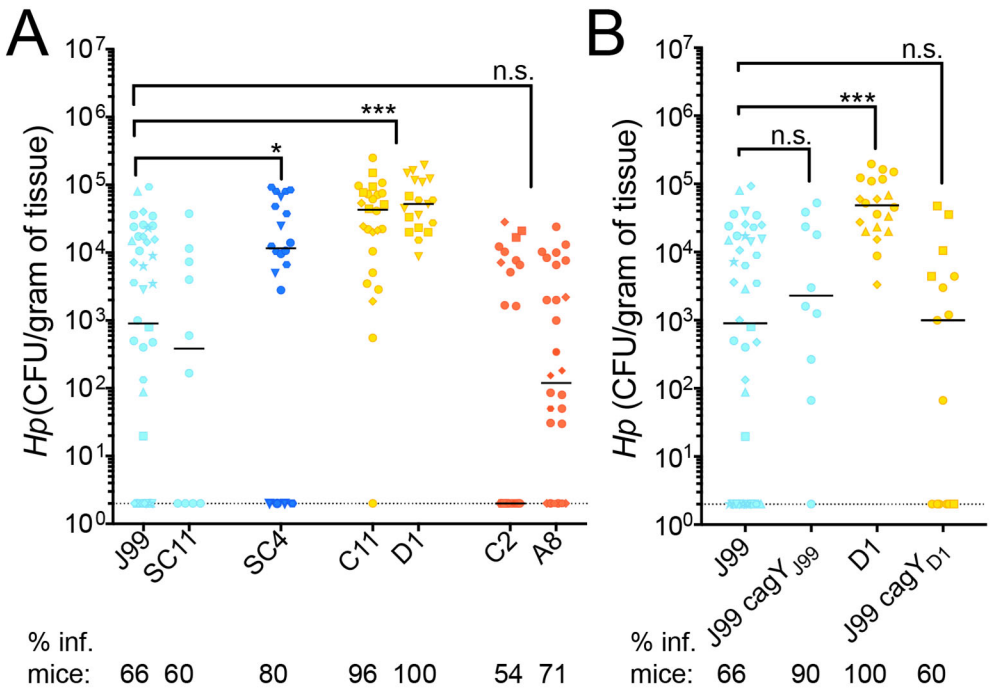
Figure 8. Cell morphology parameter divergences correlate with genetic distance during chronic infection. (**A**) Pairwise comparisons show that recent isolates within the same cell shape phenotype category (gray) are more genetically similar than recent isolates from different cell shape categories (blue). Each point represents a unique pairwise comparison between recent isolates. Midline represents the mean (π statistic). Significance was determined using Student's t-test (**** p<0.0001). (**B-C**). Correlation of genetic and phenotypic distances in (**B**) side curvature (sc) and (**C**) wavenumber (wn) per unit centerline axis length (al) between unique pairs of recent isolates (n=351) with

each point displaying a unique pairwise comparison. Plot shows a linear regression with p-value derived from F-test and correlation coefficient (R^2) shown.

Isolates differ in mouse colonization during acute infection.

Considering the observed phenotypic divergence among isolates between and within time points, we hypothesized that individual isolates may behave differently in a mouse stomach colonization model. C57BL/6 mice were infected with representative isolates from each time point and subgroup for 1 week. All isolates tested successfully colonized mice. However, the proportion of mice with detectable infection and loads (CFU/gram of stomach tissue) differed. Almost all the mice infected with the two isolates from group 2A (C11, D1) and the single isolate from the ancestral group 1B (SC4) had higher loads than representative isolates from the other groups and this increase coincided with a greater proportion of mice stably infected after one week compared to the others (Fig. 9a). Isolates within different cell shape categories and IL-8 profiles were chosen when possible. Although loss of helical shape has been shown to decrease colonization, both helical and rod-shaped isolates from clade 2B infected at lower loads (Fig. 9c). Since recombination events in *cagY* were detected in all isolates with increased loads, and these changes impacted the inflammatory response in-vitro, we tested to see if our ancestral strain (J99) with the recent variant of *cagY* (J99 $cagY_{D1}$) would also colonize at higher loads. However, J99 $cagY_{D1}$ colonized mice similarly to J99 ancestral (Fig. 9b), indicating the increased mouse colonization phenotype is not conferred by *cagY* variation. Together these results suggest that there are additional, unknown properties of these isolates contributing to colonization (Fig. 9c). However, the differences in

474 colonization between groups 2A and 2B and 1A and 1B supports the assertion that
475 subgroup differentiation has phenotypic consequences for infection.



C

Strain	Subgroup	Shape Phenotype	IL-8 Phenotype	Median CFU/gram in Mouse
J99	1A	Long Pitch	0-20% increase	9.0×10^2
SC11	1A	Long Pitch	1-20% decrease	3.8×10^2
SC4	1B	Long Pitch	1-20% decrease	1.2×10^4
C11	2A	Short Pitch	0-20% increase	4.3×10^4
D1	2A	Short Pitch	>20% decrease	5.2×10^4
C2	2B	Short Pitch	>20% decrease	2.0×10^0
A8	2B	Rod	>20% decrease	1.2×10^2

476

477 **Figure 9. Mouse colonization among isolates from distinct subgroups differs and**
478 **is not explained by *cagY* variation.** (**A-B**) Each point represents the colony forming
479 units per gram of mouse tissue homogenate from a single mouse. Two CFU/gram is the
480 limit of detection (dotted line). Biological replicates indicated by different symbol shapes
481 with the median shown. P-values were calculated from pooled experimental replicates
482 using a Mann-Whitney non-parametric test (n.s. not significant, * $p<0.05$, *** $p<0.001$).

Percentage of mice with bacterial colonization above the limit of detection (% inf. mice) is indicated below the isolate names. (**A)** Colonization in WT B6 mice 1 week post infection with representative isolates from each subgroup. Color indicates isolate dendrogram subgroup 1A (light blue), 1B (dark blue), 2A (yellow), and 2B (red). (**B**) Alleles of *cagY* from indicated strains were engineered into strain J99 at the native locus and resultant isolates used for infection experiments. Data points in blue are mice infected with J99 variant of *cagY* and points colored in yellow were infected with D1 variant of *cagY*. (**C**) Summary of representative isolate phenotypic characteristics including the subgroup, bacterial shape, IL-8 phenotype (relative IL-8), and median mouse colonization (CFU/gram).

Discussion

Within-host *H. pylori* isolates from a single individual, once thought to be homogenous, have since been shown to be genetically distinct. Next-generation sequencing provides tools to examine the breadth of diversity present, however little is known about how this diversity contributes to pathogenesis and disease progression. Our study characterized both genetic and phenotypic diversity of infecting populations from a single, chronically infected individual at two time points over a six-year period. Within host evolution of *H. pylori* is shaped both by de-novo mutation and homologous recombination events, with recombination events generating the majority of the overall diversity. Previous studies have estimated the within-host mutation rate by excluding predicted recombination sites in order to make evolutionary inferences [35,55,56]. Estimated mutation rates of serial isolates range between $6.5\text{-}0.5 \times 10^{-5}$

506 substitutions/site/year. [57] However, it remains debated whether diversity generated via
507 recombination versus mutation can be accurately identified and filtered to reconstruct
508 evolutionary relationships [56,57,58]. With our analysis, we took an agnostic approach,
509 exploring sequence level diversity acquired by both mechanisms. Our overall within-
510 host molecular clock rate ($1.3x10^{-4}$ subs/site/yr) is slightly elevated compared with other
511 published estimates, since clustered nucleotide polymorphisms (CNPs), which are
512 typically excluded from molecular clock rate calculations, were included in this analysis
513 [57]. However, the nucleotide identity of strains, falls within what has been previously
514 documented for isolates from a single individual [37].

515 We found that in this individual, with no known exposure to antibiotics, infecting
516 populations increased diversity over time and clustered into genetically distinct
517 subgroups, suggesting adaptation to specific host niches [34]. Accumulation of nSNPs
518 in OMPs supports a model of adaptation driven by interactions with the host
519 environment. In other chronic infections, such as *Pseudomonas aeruginosa* infection in
520 cystic fibrosis patients, the emergence of sub-populations is driven by region specific
521 adaptation within distinct anatomical regions of the lung [59]. Evidence of anatomical
522 stomach region specific adaptation in *H. pylori* infections is limited, but it appears to
523 occur in only a small subset of patients [37].These signatures may be obfuscated by
524 frequent population mixing and migration or deterioration of structured niches due to
525 loss of acid production and other tissue changes [42].

526 While our data do not support subgroup divergence by anatomic region in this
527 individual, the selective pressures at play appear to correspond to known pathogenicity
528 phenotypes. Nonsynonymous mutations detected within the recent population (yr 2000),

529 fell within known virulence genes, including OMPs involved in host adhesion and Cag
530 PAI-associated genes. Distinct alleles within the population were confirmed by Sanger
531 sequencing. Differences in IL-8 secretion, bacterial cell morphology, and ability to
532 colonize a mouse in an acute infection model were discovered among isolates,
533 suggesting subgroup divergence driven by tissue features that vary in the stomach
534 across multiple anatomic locations.

535 Recombination within the middle repeat region of *cagY,* resulting in expansion or
536 contraction of repeats, occurs frequently in short-term animal infections and in humans
537 [15]. We observed modulation of IL-8 induction mediated by T4SS function through
538 mutation and/or recombination without expansion or contraction of *cagY* repeats. The
539 finding that isolates at later time points were more pro-inflammatory was surprising
540 considering, inflammation is thought to limit bacterial burden. However, *H. pylori*
541 persists despite relatively high levels of inflammation, so it is possible this feature may
542 be exploited during chronic infection in order to reduce competition for host resources
543 by members of the microbiota [60].

544 Isolates within this collection also had differences in cell morphology. Morphology
545 differences have been observed among strains from different individuals [61]; here we
546 find that *H. pylori* morphologies differ among isolates from a single patient. In subgroup
547 2B, we discovered convergent loss of helical cell shape through multiple unique
548 frameshift mutations in cell shape determining (*csd*) genes. Rod-shape isolates have
549 previously been shown to have a colonization deficit manifest at early time points, but to
550 recover during 1-3 months of chronic infection in mice [62]. Due to clustering of rod
551 shapes in subgroup 2B, we suspect that helical shape, while important for early

552 infection and transmission, may be detrimental at later stages of human infection or in
553 particular stomach niches. Among isolates that retained helical shape, we detected
554 more subtle differences in helical pitch, but it is unknown what genetic determinants are
555 responsible or if these differences have direct impacts on colonization.

556 The observed differences in mouse colonization between isolates from each sub-
557 population supports our initial hypothesis that there are functional consequences of sub-
558 population divergence. Typically, clinical isolates infect mice poorly as mice are not
559 natural hosts for *H. pylori*. However, a few clinical isolates have the intrinsic ability to
560 colonize and can become more robust via serial passage in the mouse stomach [63].
561 Bacterial properties, including chemotaxis, cell shape, and activity of the Cag PAI,
562 impact mouse colonization and are likely important in establishing human infections
563 [64]. In our collection, there was heterogeneity in mouse colonization among isolates
564 that corresponded to sub-group defined by the isolate dendrogram. Robust colonizers
565 may be more likely to be involved in person-to-person transmission in humans, but it is
566 also possible that these strains may behave differently in other animal models or human
567 hosts. Increases in colonization potential of representative isolates within 1B and 2B
568 sub-populations does not correlate with differences in morphology or IL-8 phenotypes,
569 indicating an additional unknown factor or combination of factors is responsible for
570 conferring a colonization advantage. Further exploration of the genetic basis for mouse
571 colonization advantage using the subgroup specific variants defined in this study may
572 give new clues to the complex selective forces operant during chronic stomach
573 colonization by *H. pylori*.
574

Materials and Methods

Growth and isolation of *H. pylori*

In the initial sampling, a total of 43 *H. pylori* isolates (13 (antral, 1994), 5 (duodenum, 2000), 12 (corpus, 2000), 1 (cardia), and 12 (antral, 2000)) were collected from biopsy samples from two separate upper gastrointestinal endoscopies performed in a single 48-yr old Caucasian male (1994) residing in Tennessee and treated at the Nashville VA Medical Center. Only 39 isolates with sufficient sequence coverage (30x) were analyzed in this study (Fig. 1). *H. pylori* isolates were grown on solid media, horse blood agar (HB agar) or shaking liquid cultures. HB agar plates contain 4% Columbia agar base (Oxoid, Hampshire, UK), 5% defibrinated horse blood (Hemostat Labs, Dixon, CA), 10 mg/ml vancomycin (Thermo Fisher Scientific, Waltham, MA), 2.5 U/ml polymyxin B (Sigma-Aldrich, St.Louis, MO), 8 mg/ml amphotericin B (Sigma-Aldrich), and 0.2% β-cylodextrin (Thermo Fisher). For HB agar plates used to grow *H. pylori* from homogenized mouse stomach, 5 mg/ml cefsulodin (Thermo Fisher), 5 mg/ml trimethoprim (Sigma) and 0.2mg/uL of Bacitracin (Acros Organics, Fisher) are added to prevent outgrowth of mouse microflora. Shaking liquid cultures were grown in brucella broth (Thermo Fisher Scientific, Waltham, MA) supplemented with 10% heat inactivated FBS (Gemini BioProducts, West Sacramento, CA). Plate and flasks were grown at 37°C under micro-aerobic conditions in 10% CO_2, 10% O_2, 80% N_2, as previously described [65]. For resistance marker selection, HB agar plates were supplemented with 15 μg/ml chloramphenicol, or 30 mg/ml sucrose, as appropriate.

DNA extraction, genome sequencing

Genomic DNA from each isolate to be sequenced was purified using the Wizard Genomic DNA Purification Kit (Promega, Fitchburg, WI) and libraries were constructed and indexed using Nextera^{RTM} DNA Library Prep Kit (Illumina, San Diego, CA) and Nexterna^{RTM} Index Kit (Illumina). All cultured isolates (n=43) were sequenced on an Illumina MiSeq instrument in the Fred Hutchinson Genomics Shared Resource. Four isolates with average coverage below 30x were dropped from the analysis. Short read fastq sequence files from the remaining 39 isolates in this study are publicly available on NCBI SRA database (BioProject accession: PRJNA633860, <https://www.ncbi.nlm.nih.gov/sra/PRJNA622860>).

Enrichment of nonsynonymous mutations in genes and functional gene classes

Gene annotations were made using the available Genbank file available for J99 (AE001439) with some manually added annotations of OMPs. All annotation files are available at <https://github.com/salama-lab/Hp_J99>. For identification of genes with excess accumulation of nSNPS, Z-scores were calculated using number of counts per gene normalized according to gene length. Genes with nSNP accumulation greater or equal to four standard deviations from the mean are listed in Table 2. Z-scores for all genes are listed in Table S3. To identify enrichment of nSNPs within functional gene-sets, each of the 1495 genes were annotated with designations in the Microbial Genome Database (MGDB, http://mbgd.genome.ad.jp/). A Fisher's exact test was used to identify MGDB gene class categories with enrichment or depletion of nSNPs. The number of nSNPs falling within certain MGDB categories were compared to expected values based on a normal distribution and p-values were corrected for multiple testing

621 using Benjamini and Hochberg false discovery rate methods [66]. Adjusted p-values
622 <0.05 were considered statistically significant.
623
624 **Bioinformatic analysis**
625 Using J99 ancestral as the reference strain, variants were called from raw paired end
626 reads using the Breseq v0.35.0 software with default parameters and SNPs were further
627 validated using default Samtools software suite [67]**.** The number of nucleotide
628 differences per site (genetic distance) between pairs of isolates was calculated using
629 PopGenome (R) (nucleotide diversity, π, [68], PopGenome, [69]). All sites that did not
630 align to the reference genome, J99, or had read depth <5 were excluded from the
631 analysis. The unique number of shared sites for each pair was calculated using the
632 BEDtools intersect function [70], reported in Table S5 and the total number of nucleotide
633 differences were derived from the list of SNPs detected (Table 1a) with low quality sites
634 filtered according to the read depth parameters above. All detected indels also were
635 excluded from this analysis to avoid inflation of genetic distance due to alignment errors
636 within highly repetitive regions [71]. The statistical significance of differences between
637 groups was assessed using Student's t-test as indicated in figure legends. Isolate
638 dendrogram was created using Nextstrain v 1.8.1. All datasets, config files,
639 documentation, and scripts used in this analysis or to generate figures are available
640 publicly at <https://nextstrain.org/community/salama-lab/Hp-J99>.
641
642 **Sequencing and PCR-RFLP of cagY middle repeat region**

643 The *cagY* sequences were determined using Sanger sequencing using primers listed in
644 Table S6 and PCR-RFLP as previously described [15]. Flanking primers were used to
645 amplify the c*agY* repeat region from every isolate. Amplicons were purified with
646 QIAquick PCR purification kit according to the instructions from the manufacturer
647 (Qiagen, MD) and digested with restriction enzyme DdeI (New England Biolabs,
648 Ipswich, MA). Digested amplicons were run on a 3% agarose for visualization after
649 ethidium bromide staining.

650
651 **Construction of *H. pylori* mutants**
652
653 Six J99 mutants were constructed (J99Δ*cagY*, J99 *cagY*$_{D1}$, J99 *cagY*$_{SC4}$, J99 Δ*cagA*,
654 J99 *cagA*$_{D1}$, J99 Δ*cagE*) and are listed in Table S7. Isogenic knockout mutants, J99
655 Δ*cagY* and J99 Δ*cagA*, were constructed using a vector-free allelic replacement
656 strategy. Upstream and downstream genomic regions flanking the gene were amplified
657 and ligated to a *catsacB* cassette, which confers mutants both chloramphenicol
658 resistant (*cat*) and sucrose sensitivity (*sacB*). Positive clones were selected with 15
659 μg/ml chloramphenicol, as previously described [72,73]. We integrated variant alleles of
660 the deleted gene at the native locus using sucrose counter selection. All mutants were
661 validated via diagnostic PCR and Sanger sequence. Primers used for generating *H.*
662 *pylori* mutants are listed in Table S6 in the supplemental material.

663

664 ***H. pylori* co-culture experiments and IL-8 Detection**
665 AGS cells, from a human gastric adenocarcinoma cell line (ATCC CRL-1739), were
666 grown in Dulbecco's modified Eagle's medium (DMEM) (Thermo-Fisher) supplemented
667 with 10% heat-inactivated FBS (Gemini-Benchmark). For co-culture with *H. pylori*, AGS

cells were seeded at 1 × 10^5 cells/well in 24-well plates 16h prior to infection. The day of infection, medium was removed from AGS cells and mid-log-phase (optical density at 600 nm (OD) 0.3-0.6) *H. pylori* resuspended in DMEM–10% FBS–20% Brucella broth was added at multiplicity of infection of 10:1. Supernatants from triplicate wells of each condition were collected at 24 hrs and assayed for the IL-8 concentration using a human IL-8 enzyme-linked immunosorbent assay (ELISA) kit according to the instructions of the manufacturer (BioLegend, San Diego, CA). IL-8 values were reported as normalized values defined as a proportion increased or decreased compared to values obtained for J99, which was included in each experimental replicate. P-values were calculated from pooled replicates from at least two independent experiments using a one-way ANOVA with Dunnett's corrections.

Analysis of Cell Morphology

Phase contrast microscopy and quantitative analysis using CellTool software package was performed as previously described [54]. Bacterial cell masks were generated through thresholding function in ImageJ. Average side curvature, wavenumber, and centerline axis length were derived from thresholded images of bacteria (>100 cells/strain) using the CellTool software package. Average parameters were then used to calculate side curvature or wavenumber to centerline axis length ratios for each isolate.

Mouse colonization

Female C57BL/6 mice 24–28 days old were obtained from Jackson Laboratories and certified free of endogenous *Helicobacter* by the vendor. The mice were housed in

692 sterilized microisolator cages with irradiated rodent chow, autoclaved corn cob bedding,
693 and acidified, reverse-osmosis purified water. All mouse colonization experiments were
694 performed exactly as described [74]. The inoculum for each infection was 5×10^7 cells.
695 After excision, the forestomach was removed and opened along the lesser curvature.
696 Stomachs were divided in equal halves containing both antral and corpus regions and
697 half stomachs were place in 0.5 mL of sterile BB10 media, weighed, and homogenized.
698 Serial homogenate dilutions were plated on nonselective HB plates. After 5-9 days in tri-
699 gas incubator, colony forming units (CFU) were enumerated and reported as CFU per
700 gram of stomach tissue. P-values were calculated from pooled experimental replicates
701 using a Mann-Whitney non-parametric test.

702 **Statistical analysis**

703 Statistical analyses were performed according to test specified above and in each figure
704 legend using Prism v7 software (GraphPad) or R v3.2.1. P-values greater than or equal
705 to 0.05 were considered statistically significant and are marked with asterisks (*,
706 p<0.05, **,p<0.01; ***, p<0.001; ****, p<0.0001; n.s., not significant).
707

708 **Ethics Statement**

709 All procedures were approved by Vanderbilt University and Nashville Department of
710 Veterans Affairs institutional review boards. All mouse experiments were performed in
711 accordance with the recommendations in the National Institutes of Health Guide for the
712 Care and Use of Laboratory Animals. The Fred Hutchinson Cancer Research Center is
713 fully accredited by the Association for Assessment and Accreditation of Laboratory
714 Animal Care and complies with the United States Department of Agriculture, Public

715 Health Service, Washington State, and local area animal welfare regulations.
716 Experiments were approved by the Fred Hutch Institutional Animal Care and Use
717 Committee, protocol number 1531.
718

719 **Acknowledgements**

720 The authors would like to acknowledge Rick Peek Jr lab members for collecting,
721 processing, and sharing the samples and isolates used in this study, Katherine Xue for
722 assistance with Breseq and Samtools variant calling, Jesse Domingo and Sherwin
723 Shabdar for their contributions to the CellTool collection of 2D images. This research
724 was supported by R01 AI054423 (NRS), T32 CA009657(LKJ), P30 DK056465-16S1,
725 R35 GM119774 (TB), and the Genomic & Bioinformatics and Comparative Medicine
726 Shared Resources of the Fred Hutch/University of Washington Cancer Consortium (P30
727 CA015704). TB is a Pew Biomedical Scholar.
728

729 **References**

730 1. Parsonnet J, Friedman G, Vandersteen D. Helicobacter pylori infection and the
731 risk of gastric carcinoma. N Engl J Med. 1991;325(16):1127–31.
732 2. Ohata H, Kitauchi S, Yoshimura N, Mugitani K. Progression of chronic atrophic
733 gastritis associated with Helicobacter pylori infection increases risk of gastric
734 cancer. Int J Cancer. 2004;109:138–43.
735 3. Blaser MJ, Atherton JC. Helicobacter pylori persistence : biology and disease. Sci
736 Med. 2004;113(3):321–33.
737 4. Blaser MJ, Berg DE. Helicobacter pylori genetic diversity and risk of human

738 disease. J Clin Invest. 2001;107(7):767–73.
739 5. Salama NR, Gonzalez-Valencia G, Deatherage B, Aviles-Jimenez F, Atherton JC,
740 Graham DY, et al. Genetic Analysis of Helicobacter pylori Strain Populations
741 Colonizing the Stomach at Different Times Postinfection †. J Bacteriol.
742 2007;189(10):3834–45.
743 6. Suerbaum S, Josenhans C. Helicobacter pylori evolution and phenotypic
744 diversification in a changing host. Nat Rev Microbiol. 2007;5(6):441–52.
745 7. Linz B, Windsor HM, McGraw JJ, Hansen LM, Gajewski JP, Tomsho LP, et al. A
746 mutation burst during the acute phase of Helicobacter pylori infection in humans
747 and rhesus macaques. Nat Commun [Internet]. 2014;5(May):1–8. Available from:
748 http://www.nature.com/doifinder/10.1038/ncomms5165
749 8. Dang BN, Graham DY. Helicobacter pylori infection and antibiotic resistance: a
750 WHO high priority? Nat Publ Gr [Internet]. 2017;14(7):383–4. Available from:
751 http://dx.doi.org/10.1038/nrgastro.2017.57
752 9. Sun L, Talarico S, Yao L, He L, Self S, You Y, et al. Droplet digital PCR-based
753 detection of clarithromycin resistance in Helicobacter pylori isolates reveals
754 frequent heteroresistance. J Clin Microbiol. 2018;56(9):1–11.
755 10. Blaser MJ, Perez-perez GI, Kleanthous H, Cover TL, Peek RM, Chyou PH, et al.
756 Infection with Helicobacter pylori Strains Possessing cagA Is Associated with an
757 Increased Risk of Developing Adenocarcinoma of the Stomach. Cancer Res.
758 1995;55:2111–6.
759 11. Cover TL. Helicobacter pylori diversity and gastric cancer risk. MBio. 2016;7(1):1–
760 9.

761 12. Olbermann P, Josenhans C, Moodley Y, Uhr M, Stamer C, Vauterin M, et al. A
762 global overview of the genetic and functional diversity in the helicobacter pylori
763 cag pathogenicity island. PLoS Genet. 2010;6(8).
764 13. Backert S, Brandt S, Kwok T, Hartig R, Ko W. NF- kB activation and potentiation
765 of proinflammatory responses by the Helicobacter pylori CagA protein. Proc Natl
766 Acad Sci. 2005;102(26):9300–5.
767 14. Barrozo RM, Cooke CL, Hansen LM, Lam AM, Gaddy JA, Johnson EM, et al.
768 Functional Plasticity in the Type IV Secretion System of Helicobacter pylori. PLoS
769 Pathog. 2013;9(2).
770 15. Barrozo RM, Hansen LM, Lam AM, Skoog EC, Martin ME, Cai LP, et al. CagY is
771 an Immune-Sensitive Regulator of the Helicobacter pylori Type IV Secretion
772 System. Gastroenterology [Internet]. 2016;(October):1–12. Available from:
773 http://linkinghub.elsevier.com/retrieve/pii/S001650851634954X%0Ahttp://www.nc
774 bi.nlm.nih.gov/pubmed/27569724
775 16. Skoog EC, Morikis VA, Martin ME, Foster GA, Cai LP, Hansen LM, et al. CagY-
776 dependent regulation of type IV secretion in helicobacter pylori is associated with
777 alterations in integrin binding. MBio. 2018;9(3):1–16.
778 17. Brandt S, Kwok T, Hartig R, König W, Backert S. NF-kappaB activation and
779 potentiation of proinflammatory responses by the Helicobacter pylori CagA
780 protein. Proc Natl Acad Sci U S A. 2005;102(26):9300–5.
781 18. Lamb A, Yang XD, Tsang YHN, Li JD, Higashi H, Hatakeyama M, et al.
782 Helicobacter pylori CagA activates NF-κB by targeting TAK1 for TRAF6-mediated
783 Lys 63 ubiquitination. EMBO Rep [Internet]. 2009;10(11):1242–9. Available from:

http://dx.doi.org/10.1038/embor.2009.210

19. Salama NR, Hartung ML, Müller A. Life in the human stomach: persistence strategies of the bacterial pathogen Helicobacter pylori. Nat Rev Microbiol [Internet]. 2013;11(6):385–99. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3733401&tool=pmcentrez&rendertype=abstract

20. Palframan SL, Kwok T, Gabriel K. Vacuolating cytotoxin A (VacA), a key toxin for Helicobacter pylori pathogenesis. Front Cell Infect Microbiol. 2012;2(July):92.

21. Javaheri A, Kruse T, Moonens K, Mejías-luque R, Debraekeleer A, Asche CI, et al. <italic>Helicobacter pylori</italic> adhesin HopQ engages in a virulence-enhancing interaction with human CEACAMs. Nat Microbiol [Internet]. 2016;2(1):1–12. Available from: http://dx.doi.org/10.1038/nmicrobiol.2016.189

22. Solnick J V., Hansen LM, Salama NR, Boonjakuakul JK, Syvanen M. Modification of Helicobacter pylori outer membrane protein expression during experimental infection of rhesus macaques. Proc Natl Acad Sci U S A. 2004;101(7):2106–11.

23. Danielli A, Amore G, Scarlato V. Built shallow to maintain homeostasis and persistent infection: Insight into the transcriptional regulatory network of the gastric human pathogen Helicobacter pylori. PLoS Pathog. 2010;6(6).

24. Wang G, Ge Z, Rasko DA, Taylor DE. Lewis antigens in Helicobacter pylori: Biosynthesis and phase variation. Mol Microbiol. 2000;36(6):1187–96.

25. Bergman M, Prete G Del, Kooyk Y Van, Appelmelk B. Helicobacter pylori phase variation , immune modulation and gastric autoimmunity. Nat Rev Microbiol. 2006;4(February):1194–7.

807 26. Björkholm B, Sjölund M, Falk PG, Berg OG, Engstrand L, Andersson DI. Mutation
808 frequency and biological cost of antibiotic resistance in Helicobacter pylori. Proc
809 Natl Acad Sci U S A. 2001;98(25):14607–12.
810 27. Wang GE, Wilson TJM, Jiang QIN, Taylor DE, Al WET. Spontaneous Mutations
811 That Confer Antibiotic Resistance in Helicobacter pylori. Antimicrob Agents
812 Chemother. 2001;45(3):727–33.
813 28. Lee H, Popodi E, Tang H, Foster PL. Rate and molecular spectrum of
814 spontaneous mutations in the bacterium Escherichia coli as determined by whole-
815 genome sequencing. Proc Natl Acad Sci U S A. 2012;109(41).
816 29. García-Ortíz MV, Marsin S, Arana ME, Gasparutto D, Guérois R, Kunkel TA, et al.
817 Unexpected role for helicobacter pylori dna polymerase i as a source of genetic
818 variability. PLoS Genetics. 2011;7(6):1–8.
819 30. Eutsey R, Wang G, Maier RJ. Role of a MutY DNA glycosylase in combating
820 oxidative DNA damage in Helicobacter pylori. DNA Repair (Amst). 2007;6(1):19–
821 26.
822 31. Dorer MS, Cohen IE, Sessler TH, Fero J, Salama NR. Natural competence
823 promotes Helicobacter pylori chronic infection. Infect Immun. 2013;81(1):209–15.
824 32. Baltrus DA, Guillemin K, Phillips PC. Natural transformation increases the rate of
825 adaptation in the human pathogen Helicobacter pylori. Evolution (N Y).
826 2008;62(1):39–49.
827 33. Linz B, Balloux F, Moodley Y, Manica A, Liu H, Roumagnac P, et al. An African
828 origin for the intimate association between humans and Helicobacter pylori.
829 Nature. 2007;445(7130):915–8.

830 34. Israel DA, Salama N, Krishna U, Rieger UM, Atherton JC, Falkow S, et al.
831 Helicobacter pylori genetic diversity within the gastric niche of a single human
832 host. Proc Natl Acad Sci U S A [Internet]. 2001;98(25):14625–30. Available from:
833 http://www.pnas.org/content/98/25/14625.abstract%5Cnhttp://www.pnas.org/cont
834 ent/98/25/14625.short
835 35. Morelli G, Didelot X, Kusecek B, Schwarz S, Bahlawane C, Falush D, et al.
836 Microevolution of Helicobacter pylori during prolonged infection of single hosts
837 and within families. PLoS Genet. 2010;6(7):1–12.
838 36. Didelot X, Walker AS, Peto TE, Crook DW, Wilson DJ. Within-host evolution of
839 bacterial pathogens. Nat Rev Microbiol. 2016;14(3):150–62.
840 37. Ailloud F, Didelot X, Woltemate S, Pfaffinger G, Overmann J, Bader RC, et al.
841 Within-host evolution of Helicobacter pylori shaped by niche-specific adaptation,
842 intragastric migrations and selective sweeps. Nat Commun. 2019;10(1).
843 38. Dunne C, Dolan B, Clyne M. Factors that mediate colonization of the human
844 stomach by Helicobacter pylori. World J Gastroenterol. 2014;20(19):5610–24.
845 39. Howitt MR, Lee JY, Lertsethtakarn P, Vogelmann R, Joubert L-M, Ottemann KM,
846 et al. ChePep Controls Helicobacter pylori Infection of the Gastric Glands. MBio.
847 2011;2(4):1–10.
848 40. Keilberg D, Ottemann KM. How Helicobacter pylori senses, targets and interacts
849 with the gastric epithelium. Environ Microbiol. 2016;18(3):791–806.
850 41. Amieva M, Jr RMP. Pathobiology of Helicobacter pylori – Induced Gastric Cancer.
851 Gastroenterology [Internet]. 2016;150(1):64–78. Available from:
852 http://dx.doi.org/10.1053/j.gastro.2015.09.004

853 42. Blanca Piazuelo PC. The Gastric Precancerous Cascade. J Clin Exp Pathol
854 [Internet]. 2013;03(03):2–9. Available from: https://www.omicsonline.org/the-
855 gastric-precancerous-cascade-2161-0681-3-147.php?aid=20275
856 43. Nilsson C, Bjo B, Skoglund A, Ba HK, Engstrand L. A Changing Gastric
857 Environment Leads to Adaptation of Lipopolysaccharide Variants in Helicobacter
858 pylori Populations during Colonization. PLoS One. 2009;4(6):1–9.
859 44. Salama NR, Hartung ML, Müller A. Life in the human stomach: Persistence
860 strategies of the bacterial pathogen Helicobacter pylori. Nat Rev Microbiol
861 [Internet]. 2013;11(6):385–99. Available from:
862 http://dx.doi.org/10.1038/nrmicro3016
863 45. Falush D, Kraft C, Taylor NS, Correa P, Fox JG, Achtman M, et al. Recombination
864 and mutation during long-term gastric colonization by Helicobacter pylori:
865 Estimates of clock rates, recombination size, and minimal age. Proc Natl Acad Sci
866 [Internet]. 2001;98(26):15056–61. Available from:
867 http://www.pnas.org/cgi/doi/10.1073/pnas.251396098
868 46. Alm RA, Ling LL, Moir DT, King BL, Brown ED, Doig PC, et al. Genomic-
869 sequence comparison of two unrelated isolates of the human gastric pathogen.
870 Nature. 1999;397(February):176–80.
871 47. Krishna U, Romero-Gallo J, Suarez G, Azah A, Krezel AM, Varga MG, et al.
872 Genetic evolution of a helicobacter pylori acid-sensing histidine kinase and gastric
873 disease. J Infect Dis. 2016;214(4):644–8.
874 48. Uchiyama I, Mihara M, Nishide H, Chiba H, Kato M. MBGD update 2018:
875 Microbial genome database based on hierarchical orthology relations covering

876 closely related and distantly related comparisons. Nucleic Acids Res.
877 2019;47(D1):D382–9.
878 49. Solnick J V. Dynamic Expression of the BabA Adhesin and its BabB paralog
879 during Helicobacter pylori infection in Rhesus Macaques. Infect Immun.
880 2017;(April).
881 50. Åberg A, Gideonsson P, Vallström A, Olofsson A, Öhman C, Rakhimova L, et al.
882 A Repetitive DNA Element Regulates Expression of the Helicobacter pylori Sialic
883 Acid Binding Adhesin by a Rheostat-like Mechanism. PLoS Pathog. 2014;10(7).
884 51. Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, Callender C, et al.
885 Nextstrain: real-time tracking of pathogen evolution. Bioinformatics [Internet].
886 2018;2–6. Available from: https://academic.oup.com/bioinformatics/advance-
887 article/doi/10.1093/bioinformatics/bty407/5001388
888 52. Alves JM, Prieto T, Posada D. Multiregional Tumor Trees Are Not Phylogenies.
889 Trends in Cancer [Internet]. 2017;3(8):546–50. Available from:
890 http://dx.doi.org/10.1016/j.trecan.2017.06.004
891 53. Kaplan-türköz B, Jiménez-soto LF, Dian C, Ertl C, Remaut H, Louche A.
892 Structural insights into Helicobacter pylori oncoprotein CagA interaction with β 1
893 integrin. 2012;
894 54. Sycuro LK, Pincus Z, Gutierrez KD, Biboy J, Stern CA, Vollmer W, et al.
895 Peptidoglycan crosslinking relaxation promotes helicobacter pylori's helical shape
896 and stomach colonization. Cell [Internet]. 2010;141(5):822–33. Available from:
897 http://dx.doi.org/10.1016/j.cell.2010.03.046
898 55. Didelot X, Nell S, Yang I, Woltemate S, Van Der Merwe S, Suerbaum S. Genomic

899 evolution and transmission of Helicobacter pylori in two South African families.
900 Proc Natl Acad Sci U S A. 2013;110(34):13880–5.
901 56. Didelot X, Wilson DJ. ClonalFrameML: Efficient Inference of Recombination in
902 Whole Bacterial Genomes. PLoS Comput Biol. 2015;11(2):1–18.
903 57. Kennemann L, Didelot X, Aebischer T, Kuhn S, Drescher B, Droege M, et al.
904 Helicobacter pylori genome evolution during human infection. Proc Natl Acad Sci
905 [Internet]. 2011;108(12):5033–8. Available from:
906 http://www.pnas.org/cgi/doi/10.1073/pnas.1018444108
907 58. Sakoparnig T, Field C, van Nimwegen E. Whole genome phylogenies reflect long-
908 tailed distributions of recombination rates in many bacterial species. bioRxiv
909 [Internet]. 2019;601914. Available from:
910 http://biorxiv.org/content/early/2019/04/07/601914.abstract
911 59. Jorth P, Staudinger BJ, Wu X, Bruce JE, Timothy L, Singh PK, et al. within Cystic
912 Fibrosis Lungs Regional Isolation Drives Bacterial Diversification within Cystic
913 Fibrosis Lungs. Cell Host Microbe [Internet]. 2015;18(3):307–19. Available from:
914 http://dx.doi.org/10.1016/j.chom.2015.07.006
915 60. Moyat M, Velin D. Immune responses to Helicobacter pylori infection. World J
916 Gastroenterol. 2014;20(19):5583–93.
917 61. Martinez LE, Hardcastle JM, Wang J, Pincus Z, Hoover TR, Bansil R, et al. To
918 Maintain Robust Motility in Viscous Environments. Mol Microbiol. 2017;99(1):88–
919 110.
920 62. Martínez LE, O'Brien VP, Leverich CK, Knoblaugh SE, Salamaa NR. Nonhelical
921 helicobacter pylori mutants show altered gland colonization and elicit less gastric

922 pathology than helical bacteria during chronic infection. Infect Immun.
923 2019;87(7):1–15.
924 63. Lee A, O'Rourke J, De Ungria MC, Robertson B, Daskalopoulos G, Dixon MF. A
925 standardized mouse model of Helicobacter pylori infection: Introducing the
926 Sydney strain. Gastroenterology. 1997;112(4):1386–97.
927 64. Baldwin DN, Shepherd B, Kraemer P, Hall MK, Sycuro LK, Pinto-Santini DM, et
928 al. Identification of Helicobacter pylori genes that contribute to stomach
929 colonization. Infect Immun. 2007;75(2):1005–16.
930 65. Humbert O, Salama NR. The Helicobacter pylori HpyAXII restriction-modification
931 system limits exogenous DNA uptake by targeting GTAC sites but shows
932 asymmetric conservation of the DNA methyltransferase and restriction
933 endonuclease components. Nucleic Acids Res. 2008;36(21):6893–906.
934 66. Hochberg Y. Controlling the False Discovery Rate : A Practical and Powerful
935 Approach to Multiple Testing Author (s): Yoav Benjamini and Yosef Hochberg
936 Source : Journal of the Royal Statistical Society . Series B (Methodological), Vol .
937 57 , No . 1 (1995), Publi. 1995;57(1):289–300.
938 67. Barrick JE, Colburn G, Deatherage DE, Traverse CC, Strand MD, Borges JJ, et
939 al. Identifying structural variation in haploid microbial genomes from short-read
940 resequencing data using breseq. 2014;1–17.
941 68. Nei M, Li WH. Mathematical model for studying genetic variation in terms of
942 restriction endonucleases. Proc Natl Acad Sci U S A. 1979;76(10):5269–73.
943 69. Pfeifer B, Wittelsbürger U, Ramos-Onsins SE, Lercher MJ. PopGenome: An
944 efficient swiss army knife for population genomic analyses in R. Mol Biol Evol.

2014;31(7):1929–36.

70. Quinlan AR, Hall IM. BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26(6):841–2.

71. Dixit PD, Pang TY, Studier FW, Maslov S. Recombinant transfer in the basic genome of Escherichia coli. Proc Natl Acad Sci U S A. 2015;112(29):9070–5.

72. Copass M, Grandi G, Rappuoli R. Introduction of unmarked mutations in the Helicobacter pylori vacA gene with a sucrose sensitivity marker. Infect Immun. 1997;65(5):1949–52.

73. Wang Y, Taylor DE. Chloramphenicol resistance in Campylobacter coli: nucleotide sequence, expression, and cloning vector construction. Gene. 1990;94(1):23–8.

74. Amundsen SK, Fero J, Salama NR, Smith GR. Dual nuclease and helicase activities helicobacter pylori AddAB are required for DNA repair, recombination, and mouse infectivity. J Biol Chem. 2009;284(25):16759–66.

75. Hatakeyama M. Structure and function of Helicobacter pylori CagA, the first-identified bacterial protein involved in human cancer. Proc Japan Acad Ser B. 2017;93(4):196–219.

964 **Supplementary Information**
965
966 Tables and Datasets
967
968 **Table S1. A total of 2,232 unique SNPs and 573 indels were detected in the**
969 **collection of 39 isolates**. For each of the 2,232 unique SNPs (**A**) and 573 indels (**B**),
970 the tables indicate the nucleotide (nt) position and gene ID according to the reference
971 (J99, AE001439). Unique events are labeled as either coding or intergenic. SNPs within
972 coding regions are further subdivided into nonsynonymous or synonymous categories.
973 The presence or absence of each mutation across each individual isolate is designated
974 as present (1) or absent (0) along with the total number of isolates with the mutation (n).
975
976 **Table S2a-b**. **Cell Envelope proteins have excess accumulation of nSNPs**.
977 Contingency tables of nSNPs falling within and outside each of the 15 MGDB class
978 categories compared to expected values based on a normal distribution. Significance
979 was determined using a Fisher's exact test. Raw and false discovery rate corrected p-
980 values are reported with p-values <0.05 considered significant. (**A**) Contingency table
981 values for total dataset and (**B**) values unique to recent isolates.

982 **Table S3**. **Total number and enrichment or depletion of nSNPs detected across all**
983 **genes.** All 1495 annotated genes in the J99 reference (AE001439) are reported with the
984 number of within-host nSNPs detected (n=536 across all genes). Relative z-scores
985 displayed in the table were calculated from weighted values based on gene length.

986 **Table S4a-d. Individual SNPs and indels associated with antrum or corpus were**
987 **not detected in this individual**. Nucleotide positions of SNP or indel unique to recent

isolates are reported with contingency tables showing number of isolates from corpus and antrum with and without that mutation as compared to isolates outside that region. Statistical significance was determined with a Fisher's exact test. Raw and false discovery rate corrected p-values are reported with p-values <0.05 considered statistically significant. (**A**)Table of recent SNPs (n=1,379) unique to corpus isolates (corpus, n=12) compared to isolates originating from other biopsy sites (other, n=15).(**B**) Table of recent SNPs (n=1,379) unique to antrum isolates (antrum, n=12) compared to isolates originating from other biopsy sites (other, n=15). (**C**) Table of recent indels (n=372) unique to corpus isolates (corpus, n=12) compared to isolates originating from other biopsy sites (other, n=15). (**D**) Table of recent indels (n=372) unique to antrum isolates (antrum, n=12) compared to isolates originating from other biopsy sites (other, n=15).

Table S5. Pairwise comparison data used in Figures 2, 4, 5, and 7. Genetic distance, shared sites, π values, and time between isolation for each unique pairwise comparison of isolates reported in this study. For comparisons between time point, only antral isolates were used.

Table S6. Primers used in this study. List of primer sequences used in this study for sequencing and strain construction.

Table S7. Strains used in this study.

Strains	**Description**	**Reference**
J99 ⊿cagY	J99 with catsacB cassette at *cagY* native locus	This work
J99 cagY$_{D1}$	J99 with D1 *cagY* allele at native locus	This work
J99 cagY$_{SC4}$	J99 with SC4 *cagY* allele at native locus	This work
J99 ⊿cagA	J99 with catsacB cassette at *cagA* native locus	This work
J99 cagA$_{D1}$	J99 with D1 *cagA* allele at native locus	This work
J99 ⊿cagE	J99 with catsacB cassette at *cagE* native locus	This work

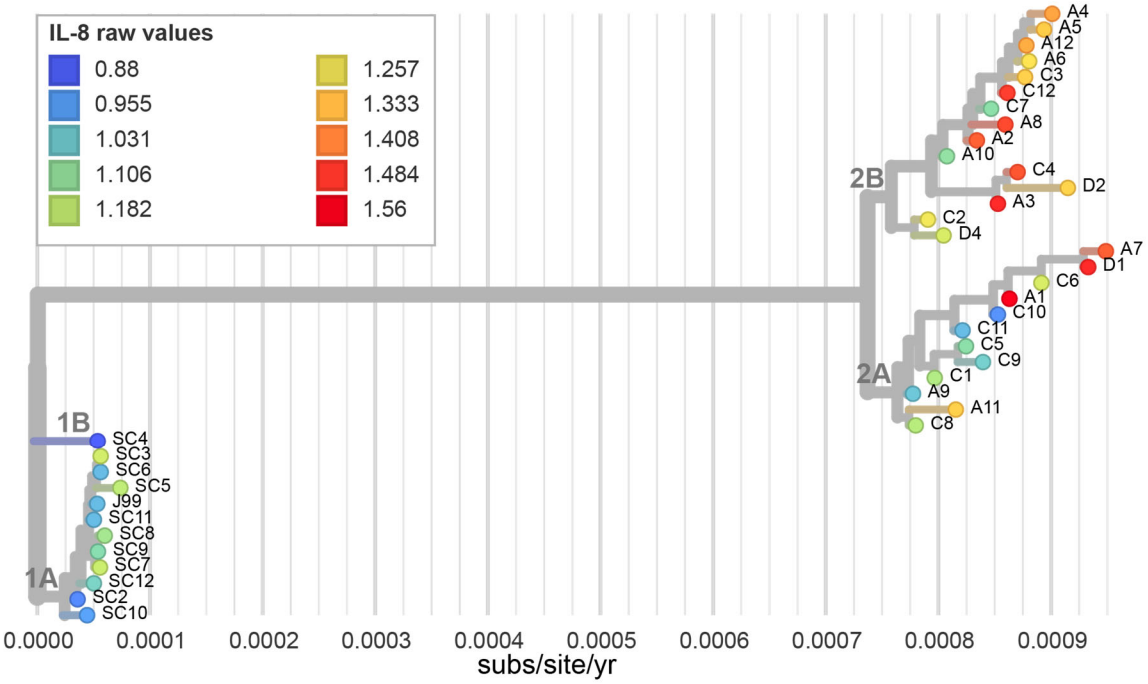


Figure S1. Proinflammatory cytokine induction during AGS cell co-culture varies between isolates. Isolate dendrogram overlaid with inflammatory cytokine, IL-8, secretion phenotype after 24 hours of co-culture (MOI=10) with gastric epithelial cell line (AGS). Leaf colors represent normalized IL-8 induction relative to ancestral isolate J99 for each isolate as shown in the figure legend.

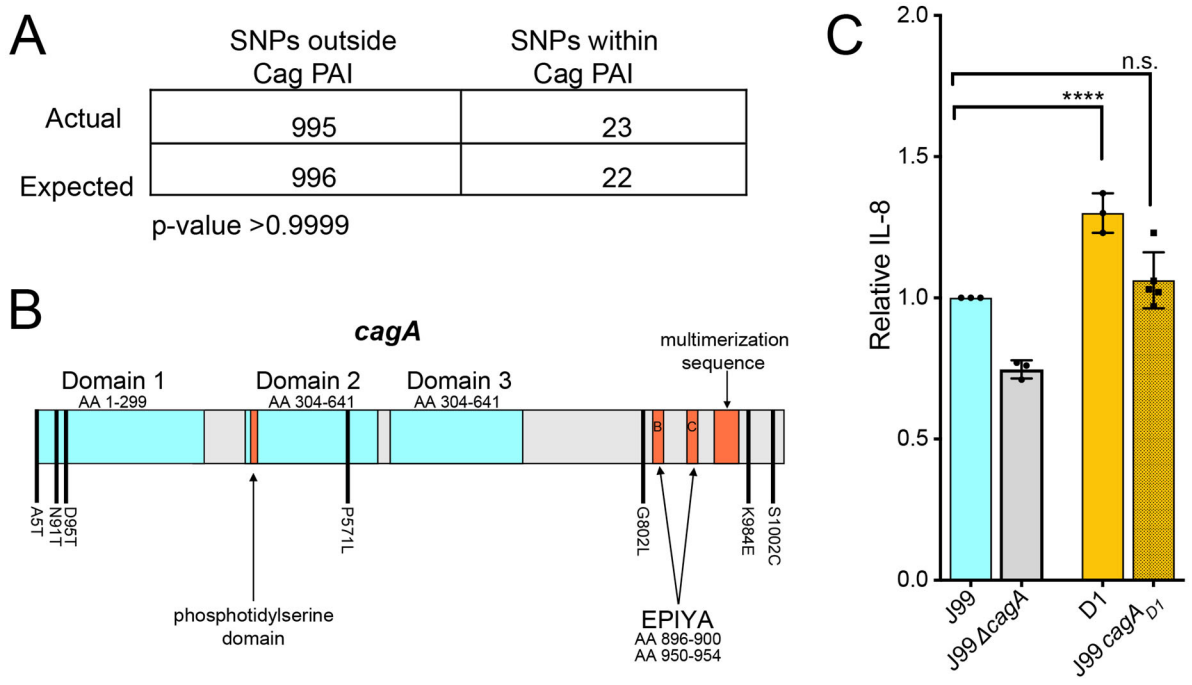
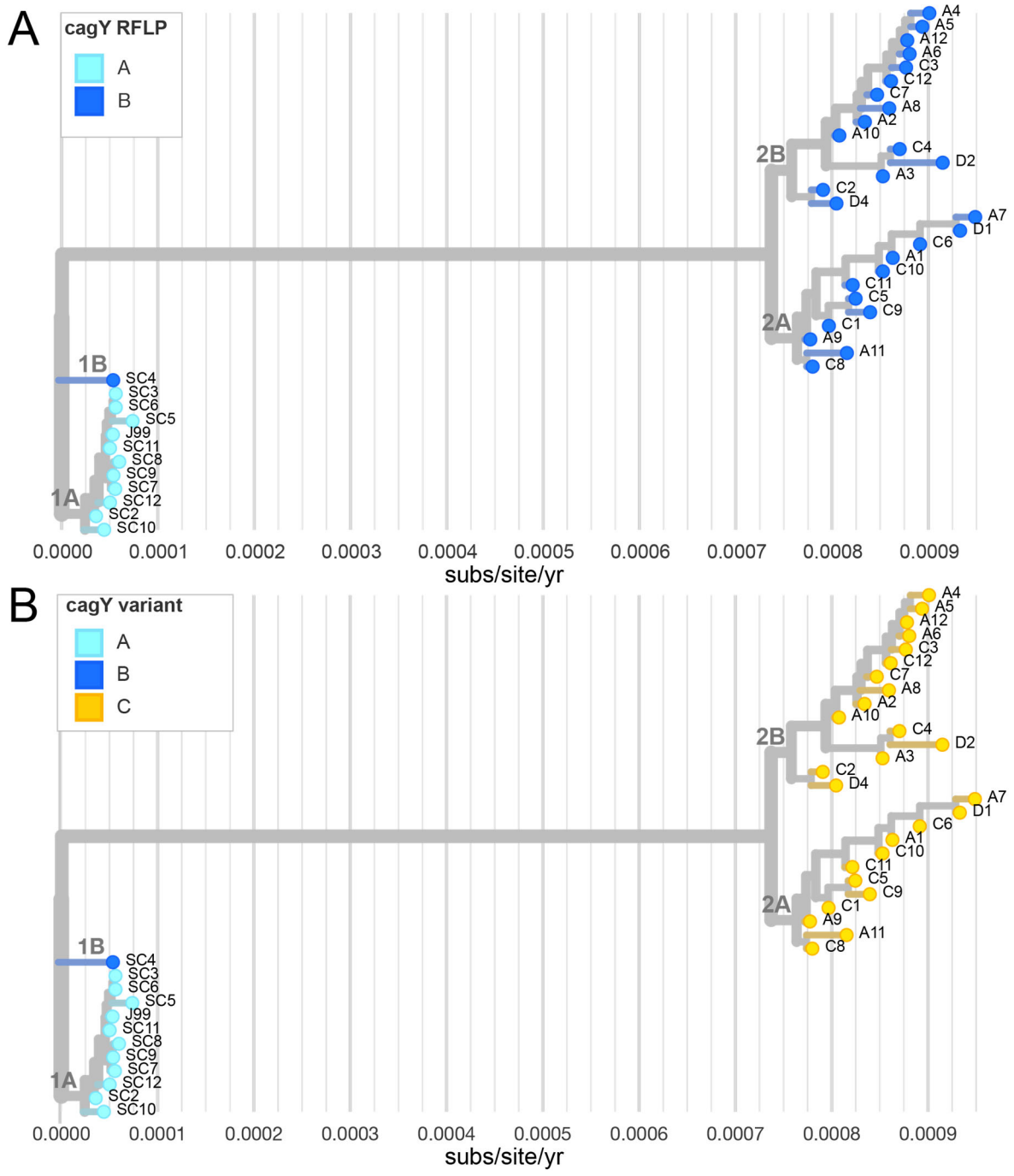


Figure S2. Genetic variation in *cagA* does not influence IL-8 induction during AGS cell co-culture. (**A**) Contingency table of nSNPs falling within and outside Cag PAI compared to expected values based on a normal distribution. Significance was determined using a Fisher's exact test.(**B**) CagA gene schematic labeled with nonsynonymous amino acid changes shared by all recent isolates (black bars). The three protein domains identified in the published crystal structure (blue), including the flexible N-terminal region (Domain I, amino acids 1-299), the anti-parallel beta sheet (Domain II, amino acids 304-641), and the (Domain III, amino acids 304-641) are labeled. Known host protein interaction motifs including the integrin binding phosphotidylserine domain, phosphotyrosine EPIYA sites, and multimerization sequence are also labeled in orange [75]. (**C**) Levels of IL-8 produced by *cagA* allelic exchange strains relative to J99 24 hrs post infection of AGS cells (MOI=10). Data points represent averaged values from triplicate wells from at least 3 independent

biological replicates. Significance was determined with a one-way ANOVA with Dunnett's corrections (n.s.,not significant; **** p<0.0001).

1040



1041

C

```
J99   630 KKECEKLLTPEAKKLEEAKKSVRAYLDCVSKAKNEAERKECEKLLTPEAKKLLENQALD
SC4   630 KKECEKLLTPEAKKLEEAKKSVRAYLDCVSKAKNEAERKECEKLLTPEAKKLLENQALD
D1    630 KKECEKLLTPEAKKLEEAKKSVRAYLDCVSKAKNEAERKECEKLLTPEAKKLLENQALD

J99   690 CLKNAKTDEERKECLKDLPKDLQKKVLAKESVRVYLDCVSKAKNEAERKECEKLLTPEAR
SC4   690 CLKNAKTDEERKECLKDLPKDLQKKVLAKESVKAYLDCVSKAKNEAERKECEKLLTPEAR
D1    690 CLKNAKTDEERKECLKDLPKDLQKKVLAKESVRVYLDCVSKAKNEAERKECEKLLTPEAR

J99   750 KLLEEAKKSVKAYKDCVLRARNEKEKQECEKLLTPEARKLLEESKKSVKAYLDCVSKAKN
SC4   750 KLLEEAKESVKAYKDCVSRARNEKEKQECEKLLTPEARKLLEESKKSVKAYLDCVSKAKN
D1    750 KLLEEAKESVKAYKDCVSRARNEKEKQECEKLLTPEARKLLEESKKSVKAYLDCVSKAKN

J99   810 EAERKECEKLLTPEARKLLEEAKESVKAYKDCVSRARNEKEKQECEKLLTPEARKLLENQ
SC4   810 EAERKECEKLLTPEARKLLEEAKESVKAYKDCVSRARNEKEKQECEKLLTPEAKKLLENQ
D1    810 EAERKECEKLLTPEARKLLEEAKESVKAYKDCVSRARNEKEKQECEKLLTPEAKKLLENQ

J99   870 ALDCLKNAKTEAEKKRCVKDLPKDLQKKVLAKESVRVYLDCVSKAKNEAERKECEKLLTP
SC4   870 ALDCLKNAKTEAEKKRCVKDLPKDLQKKVLAKESVRVYLDCVSKAKNEAERKECEKLLTP
D1    870 ALDCLKNAKTEAEKKRCVKDLPKDLQKKVLAKESVRVYLDCVSRARNEKEKKECEKLLTP

J99   930 EARKLLEEAKESVKAYKDCVSRARNEKEKQECEKLLTPEARKLLEQEVKKSVKAYLDCVS
SC4   930 EARKLLEEAKESVKAYKDCVSRARNEKEKQECEKLLTPEARKLLEQEVKKSVKAYLDCVS
D1    930 EARKLLEESKESVKAYKDCVSRARNEKEKQECEKLLTPEARKLLEQEVKKSVKVYLDCVS

J99   990 RARNEKEKQECEKLLTPEARKLLENQALDCLKNAKTEAEKKRCVKDLPKDLQKKVLAKES
SC4   990 RARNEKEKQECEKLLTPEARKLLENQALDCLKNAKTEAEKKRCVKDLPKDLQKKVLAKES
D1    990 RARNEKEKQECEKLLTPEARKLLENQALDCLKNAKTEAEKKRCVKDLPKDLQKKVLAKES
```

Figure S3. Three different *cagY* alleles distinguish isolate groups and subgroups. (**A**) Isolate dendrogram overlaid with two different *cagY* RFLP subtypes detected with restriction enzyme DdeI. RFLP subtypes, named A and B according to the figure legend, are shown in Fig. 6a. (**B**) Isolate dendrogram overlaid with unique *cagY* alleles detected with Sanger sequencing. Leaf colors correspond to each of the three unique alleles detected and reported in Fig. 6b. Group 1A shares allele A, group 1B shares allele B, and groups 2A and 2B share allele C. (**C**) Amino acid alignment of multiple repeat regions of three representative *cagY* alleles detected in the collection. J99 represents the allele found in subgroup 1A (allele A), SC4 represents the allele found in

1052 subgroup 1B (allele B), and D1 represents the allele found in 2A and 2B (allele C).

1053 Polymorphic sites are highlighted with amino acids in blue representing the reference

1054 (J99, AE001439) and red indicating a nonsynonymous substitution.

1055

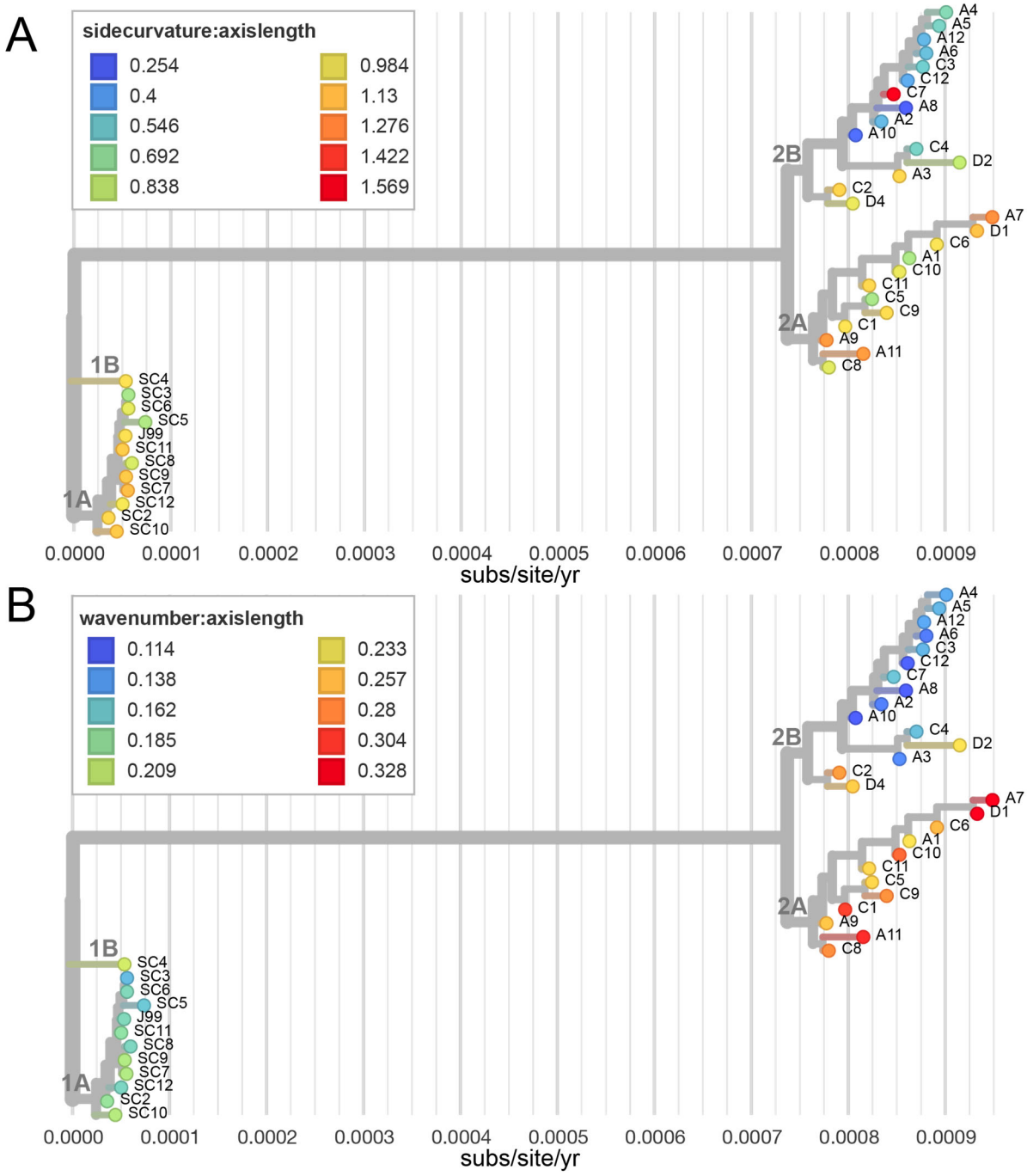


Figure S4. Cell shape parameters vary within and between isolate subgroups.

Isolate dendrogram overlaid with cell shape measurements taken from 2-D phase contrast images using CellTool. Leaf colors represent side curvature normalized by

1061 centerline axis length ratios (**A**) or wave number normalized by centerline axis length

1062 (**B**) as indicated in the figure legends.

1063

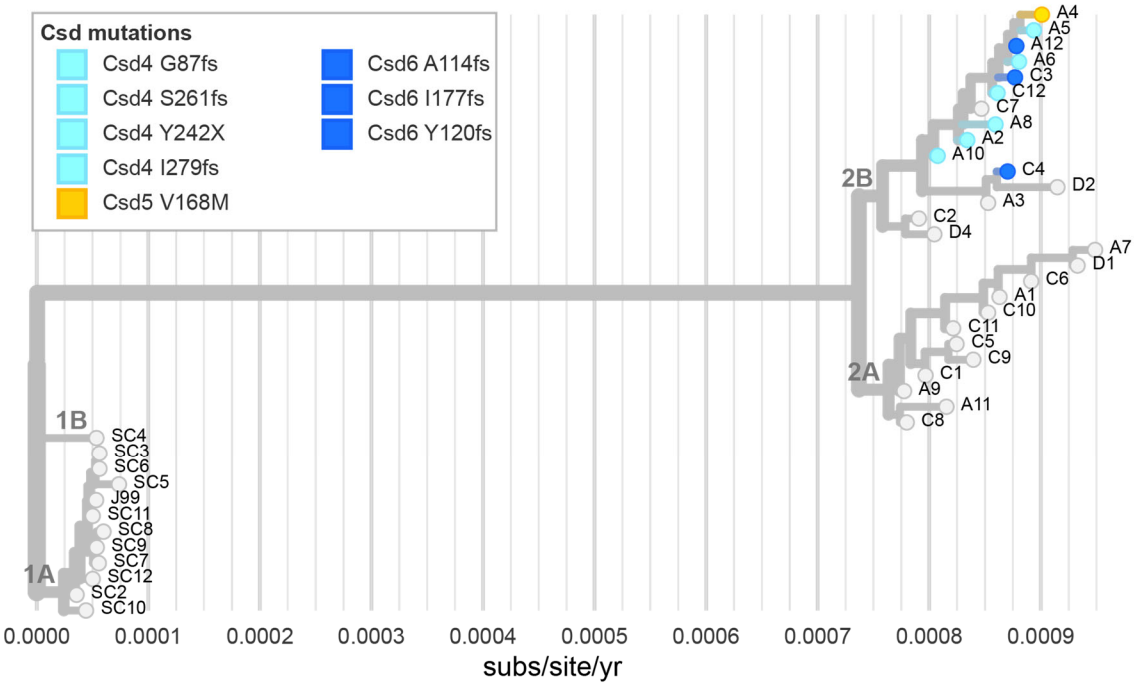


Figure S5. Mutations in *csd* genes are confined to isolate subgroup 2B. Isolate dendrogram labeled with putative loss of function mutations in cell shape determining genes (csd). Leaf colors indicate mutations in csd4 (light blue), csd5 (yellow), csd6 (dark blue) listed in the figure legend with amino acid mutations. All isolates that have retained helical shape are in gray.