Title

Recombination-independent rapid convergent evolution of the gastric pathogen Helicobacter pylori

Journal

BMC Genomics, 19(1)

ISSN

1471-2164

Authors

Chattopadhyay, Sujay
Chi, Peter B
Minin, Vladimir N
et al.

Publication Date

2018-12-01

DOI

10.1186/s12864-018-5231-7

Chattopadhyay et al. BMC Genomics (2018) 19:835
https://doi.org/10.1186/s12864-018-5231-7

BMC Genomics

RESEARCH ARTICLE Open Access

Recombination-independent rapid convergent evolution of the gastric pathogen *Helicobacter pylori*

Sujay Chattopadhyay[1*], Peter B. Chi[2], Vladimir N. Minin[3], Douglas E. Berg[4] and Evgeni V. Sokurenko[5]

Abstract

Background: *Helicobacter pylori* is a human stomach pathogen, naturally-competent for DNA uptake, and prone to homologous recombination. Extensive homoplasy (i.e., phylogenetically-unlinked identical variations) observed in *H. pylori* genes is considered a hallmark of such recombination. However, *H. pylori* also exhibits a high mutation rate. The relative adaptive role of homologous recombination and mutation in species diversity is a highly-debated issue in biology. Recombination results in homoplasy. While convergent mutation can also account for homoplasy, its contribution is thought to be minor. We demonstrate here that, contrary to dogma, convergent mutation is a key contributor to *Helicobacter pylori* homoplasy, potentially driven by adaptive evolution of proteins.

Results: Our present genome-wide analysis shows that homoplastic nonsynonymous (amino acid replacement) changes are not typically accompanied by homoplastic synonymous (silent) variations. Moreover, the majority of the codon positions with homoplastic nonsynonymous changes also contain different (i.e. non-homoplastic) nonsynonymous changes arising from mutation only. This indicates that, to a considerable extent, nonsynonymous homoplasy is due to convergent mutations. High mutation rate or limited availability of evolvable sites cannot explain this excessive convergence, as suggested by our simulation studies. Rather, the genes with convergent mutations are overrepresented in distinct functional categories, suggesting possible selective responses to conditions such as distinct micro-niches in single hosts, and to differences in host genotype, physiology, habitat and diet.

Conclusions: We propose that mutational convergence is a key player in *H. pylori*'s adaptation and extraordinary persistence in human hosts. High frequency of mutational convergence could be due to saturation of evolvable sites capable of responding to selection pressures, while the number of mutable residues is far from saturation. We anticipate a similar scenario of mutational vs. recombinational genome dynamics or plasticity for other naturally competent microbes where strong positive selection could favor frequent convergent mutations in adaptive protein evolution.

Keywords: *Helicobacter pylori*, Gastric pathogen, Protein-coding genes, Convergent mutations, Adaptive evolution

Background

Helicobacter pylori is a human-adapted bacterial species that infects about half of the world's population and is associated with chronic gastritis, stomach and duodenal ulcers, and gastric cancer [1]. It is particularly well adapted to the gastric mucosa, a highly variable environment that is hostile to virtually all other bacterial species, where it can persist for decades [2, 3]. *H. pylori*'s high adaptability to different gastric mucosal environments has been attributed to frequent homologous recombination between different strains, given that this species is highly competent for DNA uptake (transformation) [4–14]. Indeed, recombination allows for the fast emergence, spread and shuffling of different adaptive genetic changes within species. Recombination also leads to so-called genetic homoplasy – identical allelic variations in organisms with different overall ancestries. Homoplasy is extremely frequent across the genome of *H. pylori* (and other naturally competent bacteria) and is assumed to be both the result of and main evidence for rampant genetic recombination.

* Correspondence: chatsujay@gmail.com; sujayc@am.amrita.edu
[1]School of Biotechnology, Amrita Vishwa Vidyapeetham, Kollam, Kerala, India
Full list of author information is available at the end of the article

Besides homologous recombination, genetic homoplasy can also arise by convergent identical mutations occurring independently in different lineages [15–18]. We have shown previously that convergent mutations that change sequences and functions of encoded proteins contribute to adaptive evolution and often increased pathogenicity of *Escherichia coli* [19, 20], *Salmonella enterica* [20] and *Bartonella bacilliformis* [21]. However these species, unlike *H. pylori*, do not seem to be readily transformable. Untangling the contributions of recombination vs. convergent mutation in bacterial pathogens will contribute importantly to understanding the nature of homoplasy and thereby mechanisms of adaptive evolution [16, 22–25]. Previous *H. pylori* genome analysis identified many positively selected genes revealing global footprints of adaptive events [26]. However, to date, no genome-wide studies have been reported of the role of convergent mutation in genetic homoplasy and adaptive evolution of *H. pylori* or other naturally-competent bacterial species. We show here that convergent mutations accumulate at high rates across *H. pylori* genomes under strong positive selection and independently from recombination, thereby constituting the major contributor to genetic homoplasy. Importantly, we find that the homoplastic convergence is extensively overlapped by the non-homoplastic convergent mutations (repeated non-identical changes at the same position), indicating convergent mutations as a significant driver of *H. pylori* adaptive evolution.

Results

High rate of homoplastic nonsynonymous substitution in *H. pylori* core genes

We analyzed protein-coding genes from 38 *H. pylori* strains that were chosen based on multi-locus sequence typing (MLST; http://pubmlst.org/helicobacter/) profiles (Fig. 1). Two *Helicobacter cetorum* strains (MIT 00–7128 and MIT 99–5656) were used as outgroups in reconstructing the phylogeny. Of 1566 protein-coding genes annotated in the genome of reference strain 26,695, 992 genes (63%) were present in all 38 strains, based on ≥90% threshold of nucleotide sequence-identity and gene length-coverage (Table 1). The average pairwise nucleotide diversity (π) of these core *H. pylori* genes was 3.81 ± 0.03%, with about 10-fold more synonymous (silent) than nonsynonymous (amino acid replacement) substitutions on average (Table 1). Using

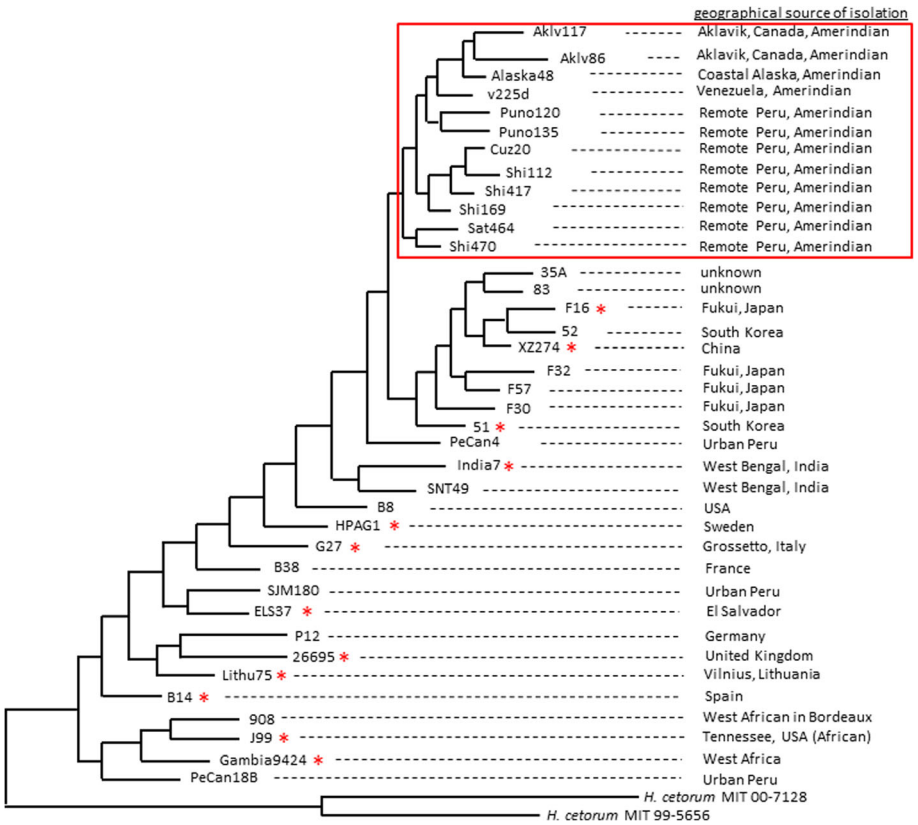

Fig. 1 Phylogram of concatenated sequences of 7 housekeeping genes (*atpA*, *efp*, *mutY*, *ppa*, *trpC*, *urel*, *yphC*) for 38 *H. pylori* strains analyzed, along with two *H. cetorum* strains used as outgroups. The red renctangle includes the strains from Amerindian population (representating the 'local' subset analyzed in Additional file 2: Figure S1). The red asterisks (*) beside the strain-names denote the representatives of the 'global' subset (analyzed in Additional file 2: Figure S1)

Table 1 Comparison of nucleotide diversity in core genes and homoplasy in the encoded proteins of *H. pylori*

	Complete set (38 strains)	'Local' (Amerindian) subset (12 strains)	'Global' subset (12 strains)
core genes analyzed	992	1106	1200
diversity, π	3.81 ± 0.03%	2.25 ± 0.02%	4.21 ± 0.03%
dN	1.44 ± 0.03%	0.91 ± 0.02%	1.74 ± 0.04%
dS	14.14 ± 0.11%	7.79 ± 0.08%	15.23 ± 0.13%
genes showing homoplasy in encoded proteins	971	963	1126
amino acid positions with repeated changes	7.85 ± 0.15%	2.33 ± 0.06%	5.10 ± 0.11%
recombinant genes	488	258	490
amino acid positions with repeated changes in encoded proteins of recombinant genes	9.03 ± 0.20%	2.83 ± 0.12%	6.02 ± 0.18%
non-recombinant genes	504	848	710
amino acid positions with repeated changes in encoded proteins of non-recombinant genes	6.66 ± 0.21%	2.15 ± 0.07%	4.40 ± 0.13%

The mean values are denoted in percentages along with their standard error values

extracted protein sequences of these 992 core genes, we determined the extent of homoplasy in amino acid sequence differences, i.e. identical changes repeatedly found in the same protein positions in different strains that were not linked by common ancestry. Almost every core gene (98%, Table 1) showed homoplasy in their encoded proteins, with 8% positions per protein on average being homoplastic.

Our dataset included 12 strains from Amerindians (both North and South America) and 26 strains from diverse other human populations worldwide. As expected, among the closely related Amerindian strains (boxed in red rectangle, Fig. 1), the average pairwise nucleotide diversity (π) of the core *H. pylori* genes was significantly lower than the diversity of the entire strains set, at both nonsynonymous and synonymous level ('local' subset, Table 1). For comparison, we selected from the remaining strains another subset of 12 non-Amerindian strains showing worldwide distribution (marked by asterisks, Fig. 1). The diversity of core genes in this subset ('global' subset, Table 1) was higher than the complete set, and almost twice that of the local dataset. Despite the difference in nucleotide diversity between the subsets, the vast majority of genes were affected by nonsynonymous homoplasy either in local or in global 12 strain subsets, and also in the complete 38 strain set. As expected, the average number of amino acid positions affected by homoplasy in the two 12 strain subsets was directly correlated with the overall nucleotide diversity level. The global subset showed two-fold more repeated identical amino acid substitutions than the local Amerindian subset. A selection of 5 additional sets of 12 strains from our non-Amerindian worldwide strain group showed an identical trend of global subset values relative to the local Amerinidian subset (Additional file 1: Table S1). However, the complete set had the highest frequency of amino acid homoplasy, in accord with its much larger size and higher diversity (Table 1). Thus, vast majority of genes in *H. pylori* is affected by high frequency nonsynonymous homoplasy, evident in either closely related or phylogenetically diverse collections, correlated with overall nucleotide diversity.

Effect of recombination on the rate of nonsynonymous homoplasy

Homoplasy could result from homologous recombination between divergent strains, a traditional interpretation, given *H. pylori*'s natural competence, and possibilities for infection by multiple strains, at least transiently [8, 27, 28]. We applied PhiPack [29] to detect intra-genic recombination events across all core genes. This detected recombination in 49.2% of the genes with nonsynonymous homoplasy in the entire dataset, 40.8% in the 12 strains from global datasets, and 23.3% in the generally more closely related local, Amerindian dataset. However, when PhiPack analysis was applied to continuous regions of 2, 3, 5 and 10 genes (rather than just to individual genes) the frequency of recombination events detected increased exponentially and, at a certain point the frequency became higher in the local set than in the global one (Additional file 2: Figure S1). But, by extending the analysis from single to multiple genes the likelihood of detecting recombination could be higher. To verify if this was the situation at least for a continuous region of 2 or 3 genes, we chose a bin of 1000-1999 bp long sequences for single gene, 2-gene and 3-gene datasets containing 347, 252 and 65 sequences respectively. Also, the nucleotide diversity for the single gene set ($\pi = 0.039 \pm 0.0004$) was almost the same for either the 2-gene dataset ($\pi = 0.039 \pm 0.0005$) or 3-gene dataset ($\pi = 0.037 \pm 0.001$). We detected 243 genes (70%) as recombinant in the single gene set of higher lengths, which

was higher than the overall frequency of genes with intra-genic recombination (49%). However, in comparison, the 2- and 3-gene datasets of similar lengths and diversity values showed 194 genes (77%) and 55 genes (85%) respectively as recombinants. Thus, intra-genic recombination appears to be less frequent than inter-genic recombination in *H. pylori*, while the members of Amerindian subpopulations might be exposed to increased opportunities for recombination owing to geographical structure as suggested previously [30].

The rates of nonsynonymous homoplasy were re-evaluated after PhiPack analysis-based separation of genes into recombinant and non-recombinant. The average number of amino acid positions affected by repeated identical changes was 1.4-fold higher in recombinant than non-recombinant genes (9.03 ± 0.20% vs. 6.66 ± 0.21% positions, respectively, $P < 0.0001$) (Table 1). The same pattern of rather modestly high rates of nonsynonymous homoplasy in recombinant vs. non-recombinant genes were also observed in the local and global datasets (Table 1). We infer that nonsynonymous homoplasy is only partially due to intra-genic recombination events.

Most repeated nonsynonymous changes are not linked to repeated synonymous changes

We used a recently developed statistic tool synDss [31] to assess linkage between repeated nonsynonymous and synonymous changes, an expected result of genetic recombination. Unlike the alignment score-based PhiPack, synDss separately computes synonymous and nonsynonymous codon distances within each sliding window for a given gene (by computing two types of phylogenetic incongruence based on synonymous-only and nonsynonymous-only substitution information) to test for recombination in intra-genic regions. Because of this tool's computationally-intensive nature, we focused on randomly selected 25 recombinant and 25 non-recombinant genes as defined by PhiPack (Additional file 1: Table S2 and Additional file 2: Figure S2).

Overall, in all 50 genes combined, synDss detected 27 genes that crossed respective 95% bootstrap significance thresholds for phylogenetic incongruence, thereby suggesting clusters of repeated phylogenetically-unlinked changes – 17 in recombinant and 10 in non-recombinant sets (Fisher's exact test two-tailed $P = 0.088$). Only 6 genes crossed the significance thresholds for both nonsynonymous and synonymous changes, and all of them represented PhiPack-defined recombinant set (Fisher's exact test two-tailed $P = 0.03$). Interestingly, 17 genes showed significance due to nonsynonymous substitutions only, which was much higher than genes showing significant values for both nonsynonymous and synonymous substitutions (6 genes; Fisher's exact test two-tailed $P = 0.016$) and, especially, for synonymous substitutions only (4 genes; Fisher's exact test two-tailed $P = 0.003$). However, between the PhiPack-based recombinant and non-recombinant gene sets, there was no statistical difference in the number of genes showing significance for nonsynonymous changes only (10 and 7 genes, respectively, Fisher's exact test two-tailed $P = 0.55$) or for synonymous changes only (1 and 3 genes, respectively, Fisher's exact test two-tailed $P = 0.61$). Multiple testing correction using the Benjamini-Hochberg method [32] showed that the significant P values fell below the 5% false discovery rate.

One possible explanation for a large number of cases showing significant phylogenetic incongruence for nonsynonymous homoplasy only could be a lack of recombination detection power in the synonymous substitution analysis. However, such an explanation is unlikely since the observed rate of synonymous substitutions is way higher than the nonsynonymous rate in *H. pylori* (Table 1). Importantly, in gene phylogenies the corresponding amino acid homoplastic sites were not clustered in any single allele-pairs, thereby ruling out the possibility that recombination happened in an extreme situation free of synonymous changes. Therefore, although synDss partially validated that some of *H. pylori*'s nonsynonymous homoplasy is of recombinant origin, the majority of repeated nonsynonymous changes are not linked to repeated synonymous changes, and thus are not likely to stem from intra-genic recombination.

Prevalence of convergent mutations in proteins with homoplastic changes

Homoplastic (i.e. repeated identical) changes constitute a type of convergent change that, taken alone, could be attributed to recombination. Another type involves repeated non-identical (i.e. non-homoplastic) changes at the same positions, and represents unambiguous evidence of convergence by mutation. We examined whether amino acid positions with homoplasy were also targeted by non-homoplastic convergence. For this further homoplasy analysis, we concentrated on the PhiPack-defined set of 504 non-recombinant genes, to reduce possible ambiguity in the results.

We found a high frequency of repeated but non-identical (i.e. non-homoplastic convergent) mutations in proteins encoded by these 504 core *H. pylori* genes (Fig. 2). Interestingly, 73% of positions with non-homoplastic convergent mutations were also targeted by homoplastic changes in our 38 genome dataset, and the frequency of homoplastic changes was higher than that of non-homoplastic ones. To determine whether the observed pattern could be explained by random mutation accumulation under no selection, for each gene we performed simulation of mutations under neutrality, based on the naturally observed mutation rate. We found that the frequency of non-homoplastic

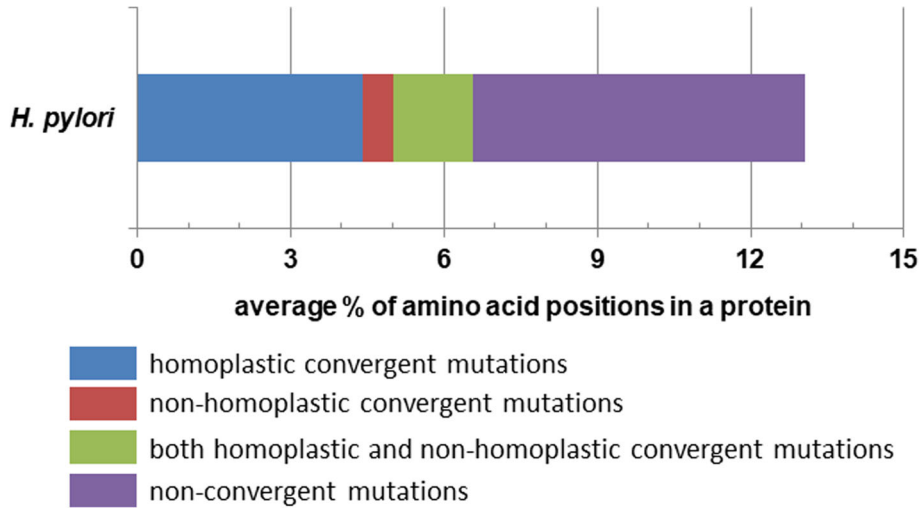

Fig. 2 Average distribution of amino acid positions with different kinds of point mutations in encoded proteins of non-recombinant core genes with convergent changes in *H. pylori*

(non-identical) amino acid convergence was > 3 times higher in the real datasets than in simulated ones; in addition, the difference in frequency of homoplastic (identical) amino acid convergence was > 50 times higher in the real datasets than in the simulated ones (Fig. 3), suggesting non-neutral accumulation of convergent changes ($P < 0.0001$). Also, as expected under neutrality, the simulated data showed higher frequency of non-homoplastic amino acid changes than the homoplastic ones (Fig. 3), which contrasted with the observed prevalence of homoplastic changes in *H. pylori* genes.

In principle, the high frequency of convergent mutations in particular positions might arise by chance alone due to *H. pylori*'s overall high mutation rate and the limited availability of mutable sites that tolerate changes without impairing function. To test this, we scored the average frequency of mutations per mutated site in encoded proteins, and found a bimodal distribution with two distinct peaks in frequencies – one with a single mutation position and the other with an average of four mutations per site, affected by multiple repeated changes of identical and/or non-identical (homoplastic or non-homoplastic convergent) nature (Fig. 4). We next analyzed the mutation frequency in 5 randomly selected replicates of 100, 50, 25 and 10 randomly chosen core genes (Additional file 2: Figure S3). We found that the bimodal mutation frequency distribution was also preserved in the 100, 50 and 25 member smaller subsets (such bimodality was not evident in the 10 gene subset due to small sample size noisiness) (Additional file 2: Figure S3). Thus, the bimodal distribution across the genome is not attributable to a skewed mutation distribution in a limited gene set.

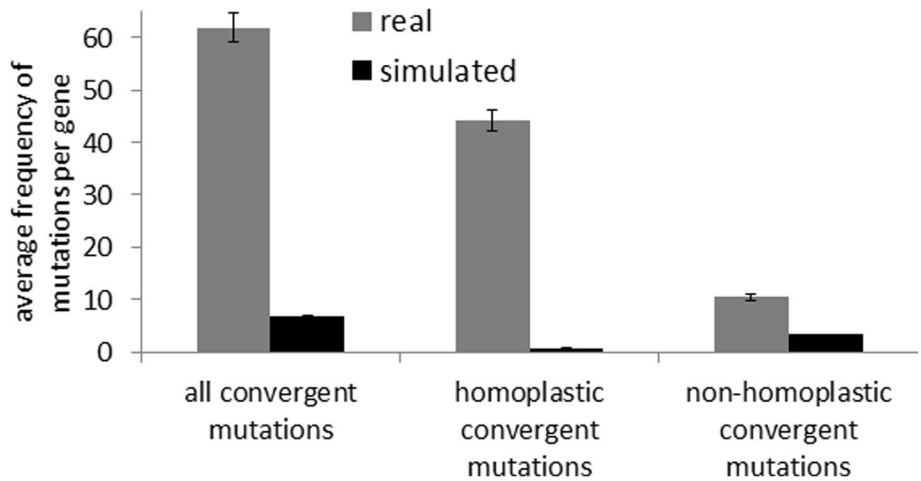

Fig. 3 Comparison of convergent mutation-frequencies in encoded proteins of real and simulated datasets of *H. pylori* non-recombinant core genes

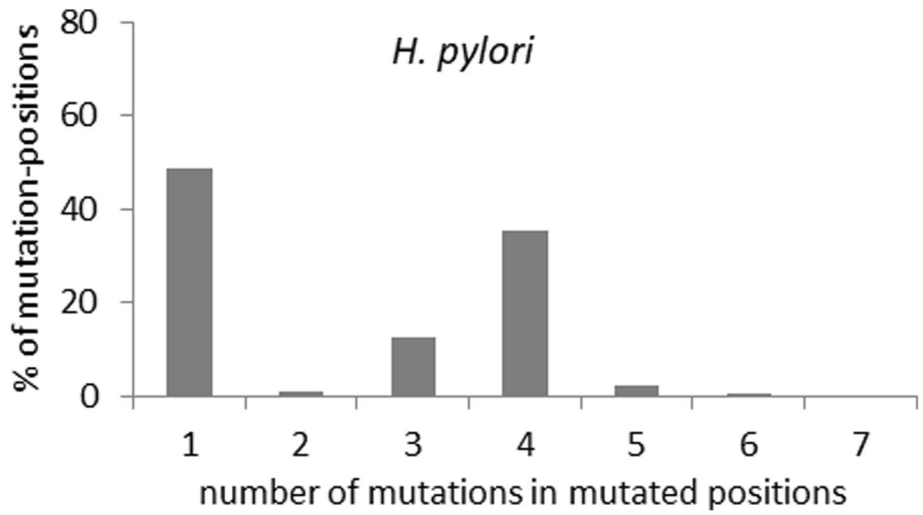

Fig. 4 Average frequency distribution of mutations in the encoded proteins showing convergence in non-recombinant core genes of *H. pylori*

We next performed a simulation analysis of one random set of 25 genes, using two sets of mutation rates and dN/dS values: (a) as observed in the *H. pylori* genes, and (b) two-fold higher than the values observed in the *H. pylori* genes. In both cases, the distributions of amino acid mutation frequencies in the simulated datasets were unimodal, not bimodal as observed for real datasets, with the probabilities of multiple mutations per site dropping progressively (Additional file 2: Figure S4). Taken together, these results demonstrate that homoplastic and non-homoplastic convergence strongly overlap, and that the occurrence of convergent nonsynonymous changes in *H. pylori* non-recombinant genes cannot be explained by neutral accumulation of mutations. This suggests underlying action of strong positive selection for convergent evolution.

Diverse array of enriched functional categories

The high rate of acquisition of convergent mutations in *H. pylori* suggests robust positive selection in the affected genes (Additional file 1: Table S3). Of the 487 non-recombinant genes with potential convergent mutations, 158 (32%) were in 16 biological process categories that were statistically enriched or over-represented ($P < 0.05$) in the dataset out of a total of 67 categories (Fig. 5). These included genes involved in protein secretion, chemotaxis, transport, virulence, flagellar assembly, RNA degradation, transcriptional regulation, and also several metabolic or biosynthetic pathways. Importantly, multiple genes in each of these pathways accumulated convergent mutations across different *H. pylori* lineages. Notably, we did not find enrichment of genes representing some of the functional categories commonly over-represented in *E. coli* or *Salmonella* [20], such as two-component systems, DNA repair, or biosynthetic pathways of cofactors, vitamins, quinones, carbohydrates, etc. Thus, *H. pylori* proteins affected by homoplastic changes that likely stem from convergent evolution are not distributed randomly among functional categories. This implies that many of the observed protein sequence changes have contributed to *H. pylori*'s adaptation to particular gastric mucosal environments.

Discussion

Here we provide a genome-wide analysis of *H. pylori* strains that shows that, independent of recombination, convergent evolution by mutation and selection contributes importantly to extensive homoplasy in the core genes of this gastric pathogen. Such a high frequency of homoplasy, we believe, would considerably impact the tree topology based on all core genes. Instead, we reconstructed phylogeny of the analyzed strains based on the MLST loci to have only an overview of clonal grouping and diversity, since a phylogenomic or phylogeographic study was not the focus of our work. The MLST loci-based phylogram, using two *Helicobacter cetorum* strains as outgroups, displayed mainstream African, East Asian and Amerindian strains in their own distinct groups. Interestingly, the clustering of an urban Peruvian strain PeCan18B with the African strains could be attributed to the history of African slavery in Peru (and much of Latin America) during colonial times, a significant contribution of genes of ethnic Black Africans to the Lima region human gene pool, and distinct self-identified Afro-Peruvian communities in Lima and elsewhere in the country [33].

Convergent mutations at specific amino acid positions are mostly selected under positive selection, for adaptive or compensatory functions [19–21, 34–40]. The possible

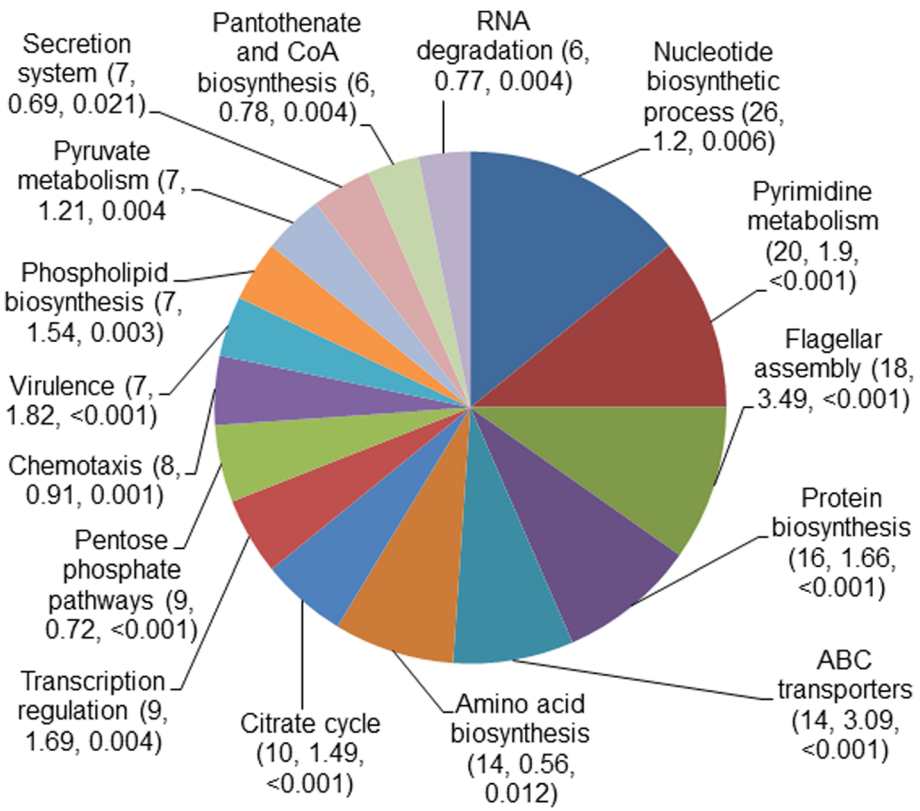

Fig. 5 Enriched (or overrepresented) functional categories of proteins encoded by non-recombinant core genes with convergent mutations

structural and functional reasons for accumulation of mutations at particular sites are manifold. The adaptive significance of convergent mutations in proteins, especially of the non-homoplastic ones, could stem from selective pressures on sets of functionally/structurally critical amino acid residues to avoid immune system epitope recognition or alter or even eliminate protein function [40]. Also, homoplastic mutations, in particular, suggest positive selection for fine-tuned directed changes in protein function [34].

Our conclusion about the non-recombinant nature of many or most homoplastic *H. pylori* gene changes is based primarily on the abundance of non-homoplastic convergent mutations and their significant overlap with homoplastic variations in *H. pylori* genes. Notably, it is not based just on a possible failure of recombination-detection tools to establish a strong link between structural homoplasy and recombination (which are probability-based, and some may need further improvement or might be biased towards assuming a priori that homoplasy is recombinant in nature). There is always a possibility that an excessively high frequency of adaptive nonsynonymous homoplasy in localized regions could be of recombinant origin, while a low frequency of synonymous homoplasy around the regions could be due to a lack of positive selection to be either introduced or maintained by recombination. Such a scenario would easily arise if the recombination tract lengths were very short, in the order of tens of nucleotides. However, in light of previously reported large lengths of imported fragments [10–12] discussed below and consistent high rate of synonymous changes (10 times higher on an average than the rate of nonsynonymous changes), nonsynonymous homoplasy without a linked synonymous homoplasy would require a large number of positionally constrained recombination events around the former type of changes. Thus, we consider the 'stand-alone' nonsynonymous homoplasy to be more plausibly (parsimoniously) explained via mutation rather than recombination, at least for a substantial part of them. Additional argument for the mutational origin of homoplasy is, if course, the overlap of both homoplastic and non-homoplastic types of change in the affected nucleotide positions.

That said, we note that our study does not diminish the adaptive importance of frequent homologous recombination in *H. pylori*. Certainly a portion of the amino acid homoplasy assigned here as likely due to mutation actually stems from recombination. Laboratory and sequence analyses of length of imported fragments in *H. pylori* genomes indicated that most recombination in *H. pylori* is inter- rather than intra-genic nature. The mean length of *H. pylori* core genes is 959 bp (95% in range of 922–996 bp). According to two recent studies, the mean

length of DNA transferred from donor to recipient in vivo by a single recombination event was estimated as about 1300 bp and 1247 bp (95% confidence intervals of 950-1850 bp and 841–1721 bp, respectively) [11, 12] . Another in vitro transformation study reported an average length of imported fragments varied from 1294 to 3853 bp [10]. Finally, an additional study [41] reported an average recombination tract length of 895 bp based on two lineages of very closely related isolates from a single patient. The above ranges of recombination tract lengths were considerably higher than the value of 417 bp detected previously among serial isolates [8]. However, a follow-up study from the same group had increased the estimated recombination length three fold [12] and attributed such a discrepancy to limited number of recombination events in the earlier study. This suggests that recombination, under selection, might be primarily responsible for exchanges of relatively large regions (likely exceeding the length of a single gene), rather than for swapping of single nucleotide polymorphisms or very small regions. Thus, if recombination would be responsible for most 'stand-alone' single nucleotide homoplasy in *H. pylori*, this would require a fairly large number of sequential, partially-overlapping recombination events within the immediate proximity a specific nucleotide position. While this is possible, a mutational origin of the widespread single nucleotide homoplasy appears to be a more parsimonious explanation. Therefore, in individual gene-based analysis, homoplastic changes owing to inter-genic recombination would most likely be phylogenetically linked and not appear as independent repeated substitutions in our analysis.

Interestingly, *H. pylori* shows a bi-modal distribution of mutation-frequencies, which could be an intrinsic characteristic of selective forces operating on proteins in this fast-evolving organism. This suggests a strong bias for multiple mutations at the same amino acid positions. But our simulations under neutrality do not support a possibility of saturation of mutable amino acid residues even at a mutation rate twice that of *H. pylori*. Although the simulation algorithm involved limitations due to simplified assumptions, our goal was to decipher if a simulation model under neutrality could at all explain the observed trend of the high frequency of homoplastic and non-homoplastic convergence, or bimodality of the mutation frequency distribution. Also, we believe that, while the lack of accuracy of the reconstructed tree topologies compromises the detection level of homoplasy events, such a scenario would equally affect the detection of nonsynonymous and synonymous homoplasy. Therefore, the findings of unlinked nature of nonsynonymous homoplasy from the synonymous ones, along with its overlap with the nonsynonymous non-homoplastic convergent mutations would mostly remain true. However, to validate these results and to decipher the detailed evolutionary history of lineages, a population-level genomic study is warranted which is out of scope of the present work. Importantly, the diversity of human gastric mucosal micro-niches for *H. pylori* is very limited, especially relative to organisms like *Escherichia coli* or *Salmonella enterica*, which have a much broader host and environmental range. Thus, the affected proteins may be constrained to only a limited set of residues that could be under strong selection-pressures to tackle highly specific environmental challenges. Interestingly, the proportion of non-convergent mutation sites is lower than the convergent ones in *H. pylori*. This indicates that the effective number of evolvable sites that respond to selection pressures might have reached a saturation point, while the frequency of positions that can carry mutations via neutral drift is possibly far from saturation.

An overall low dN/dS value with high homoplasy rates in *H. pylori* genes could indicate mutational site saturation for nonsynonymous positions and strong purifying selection as suggested previously [24]. However, we believe that single nucleotide homoplasy in *H. pylori* is positively selected that could be explained by either adaptive or compensatory nature of the convergent mutations. The adaptive mutations could provide fitness diversification in a highly variable and hostile environment encountered by *H. pylori* during colonization of the same and different individuals. This is evident from the overlap of both homoplastic (to possibly achieve a directed, fine-tuned fitness level) and non-homoplastic (to possibly offer evasion of host defenses) changes in majority of the positions with convergent mutations. On another hand, the excessively high rate of mutations could include, along with the beneficial ones, a significant amount of deleterious mutations. High rates of compensatory mutations (along with recombination) might compensate for disruptions in fitness by adjusting structure-function relationships within a single protein or between different proteins.

Maintenance of extensive genomic diversity in *H. pylori* via a high level of homoplastic and non-homoplastic convergence suggests a possible interplay of adaptive and compensatory roles of the changes to cope with extremely hostile micro-niches. For example, earlier works on *E. coli* [42] and *H. pylori* [43] strains have demonstrated that the acquisition of adaptive antibiotic resistance mutations is balanced by additional mutations that compensate for the biological cost of resistance, which thereby can lead to increased fitness and stabilization of resistant mutants in the population. To cope with broad classes of environmental conditions (ranging from within-host micro-niches to hosts diversified by geography, genotypes, dietary habits, etc.), it is likely that mutational convergence in *H. pylori* is distributed across

a wide range of proteins from different functional categories. Indeed, functional classification showed that the genes with convergent mutations represented a diverse group of functional categories. Kawai et al. [44] earlier detected seven genes where the encoded proteins accumulated positively selected amino acid changes. However, five of them were non-core in our dataset. The remaining two core genes were found to be recombinants, and therefore not considered as candidates. Recently, Montano et al. [26] demonstrated global footprints of selection in *H. pylori*, identifying an array of genes to be under positive selection in worldwide population or in different local (African, EuroAsian, East Asian, or American) sub-populations. Fifty-three of their listed selection candidates were included in the core gene dataset we analyzed here. Our PhiPack analysis identified 28 of these genes as non-recombinant. All these genes were selection candidates in our list as well owing to their accumulation of convergent mutations (marked by asterisks in Additional file 1: Table S3). Similarly, another recent work by Yahara et al. [45] offered a list of 134 *H. pylori* genes with positively selected amino acid changes. We found 23 of these genes to be non-recombinant core in our dataset, all of which were detected as candidate genes (marked by '+' in Additional file 1: Table S3). However, it was not surprising given the fact that 97% of non-recombinant genes in our dataset were selected as candidates for the accumulation of convergent mutations – homoplastic and/or non-homoplastic. Nevertheless, this listing of candidate genes under positive selection and naturally occurring convergent mutations (Additional file 1: Table S3) should encourage further studies to understand functional implications of the many mutations found here.

Conclusions

Our study highlights the genome-wide prevalence of evolutionarily convergent mutations and possible (patho-)adaptive roles in one of the most common and fast-evolving bacteria species. We hypothesize that the selection-driven convergent mutational dynamics detected in *H. pylori* is also characteristic of the evolution of other naturally competent bacterial species like *Neisseria* [46], *Streptococcus* [47], etc. Future large-scale population-wide analysis of the hundreds of sequenced *H. pylori* genomes now available will further characterize mutational footprints across isolates from both geographically diverse and localized populations, and thereby further identify sequence changes linked to within-host micro-niche adaptation, associations with distinct disease phenotypes, human genotypes, dietary habits and geographical locations [48–51].

Methods

Reconstruction of MLST phylogeny

The maximum-likelihood based phylogenetic tree of 38 completely sequenced *H. pylori* genomes was reconstructed using concatenated sequences of 7 housekeeping genes – *atpA, efp, mutY, ppa, trpC, ureI, yphC* (http://pubmlst.org/helicobacter/). The maximum-likelihood based phylogenetic trees for MLST and all individual genes in this study were reconstructed using the general time reversible (GTR) substitution model with estimated base frequencies site-specific by codon position distribution as implemented in PAUP* [52].

Extraction, phylogenetic reconstruction, homoplasy detection and diversity analysis of core genes

The TimeZone software package [53] was used to extract core protein-coding genes, excluding the annotated pseudogenes, from the genome of *H. pylori* strain 26695 used as reference. Stand-alone BLAST, implemented in TimeZone, was used to find orthologs of each gene from the genomes of other *H. pylori* strains using 90% cut-off values for both nucleotide sequence identity and sequence length coverage. For each core gene, phylogenetic analysis of its allelic sequences was performed via TimeZone to detect homoplastic events in the encoded protein. The rates of synonymous (dS) and nonsynonymous (dN) mutations using mutation-fraction method of Nei and Gojobori [54], and average pairwise nucleotide diversity (π) were also calculated via TimeZone.

Detection of recombination

We used the PhiPack software package [29] for detecting probable recombination events. This package included 3 recombination-detection statistics: pairwise homoplasy index (Phi), maximum χ^2 (MaxChi), and neighbor similarity score (NSS). These statistics sequentially compare each set of all possible combinations of three sequences in the dataset, and perform sliding window analysis for the relative distribution of nucleotide polymorphisms. Presence of segments of the sequence alignment that support significantly different sequence diversity or sequence topology than the upstream and downstream segments is considered as evidence of intra-genic recombination. A gene was considered to be recombinant if P values for all of the 3 statistics were < 0.1 [55].

The synDss analysis [31] involved detection of phylogenetic incongruence based on synonymous and nonsynonymous evolutionary distances across multiple sliding windows within each given gene. The nucleotide distances were estimated on the codon level (i.e. the expected number of substitutions per site) using the window size and step size as tuning parameters, calibrated to have approximately 100 windows across each alignment. Statistical significance of the synDss statistic

was computed by parametric bootstrapping of 500 replications under the null hypothesis of no recombination. The Monte Carlo estimate of the P value was achieved from this parametric bootstrap, and 95% bootstrap significance thresholds were shown for the synonymous and nonsynonymous plots (Additional file 2: Figure S2).

Simulations

Random simulations of mutation were performed using EvolveAGene 3 [56]. For the simulation of each gene, the sequence of reference *H. pylori* strain 26695 was considered as the root sequence to generate a random tree topology where each branch had equal probability of leading either to a terminal node or to an internal node. To assign the mutation rate, average branch lengths and average selection on amino acid replacements (i.e., dN/dS) were estimated from the real data set phylogeny of corresponding gene for the simulations shown in Fig. 4. In contrast, for the ones in Additional file 2: Figure S4 (Additional file 2), a set of 25 sequences were simulated based on two different values of average branch-lengths. Since real data sets of core genes did not have any indels, no indels were allowed in the simulated data sets. Selection over sequence, as well as over branches along the tree, was set to be constant with the default modifier value of 1. Random selection of gene subsets was performed by in-house random number generator program.

Functional enrichment analysis

DAVID software [57] was applied to perform functional annotation clustering using DAVID annotation of *H. pylori* genes as reference. For the analysis, classification stringency was set to 'medium'. To minimize redundancy, annotated clusters representing only Biological Process category of Gene Ontology (GO) were considered. The functional categories with enrichment score > 0.5 and *P* value < 0.05 were selected as the enriched or overrepresented ones.

Additional files

Additional file 1: Table S1. Comparison of nucleotide diversity in core genes and homoplasy in the encoded proteins of analyzed 'Global' subset of *H. pylori* with 5 other global subsets based on random selection of 12 non-Amerindian strains in our dataset. The mean values are denoted in percentages along with their standard error values. **Table S2.** Results of 95% significance level (based on a parametric bootstrap with *B* = 500) of synonymous and non-synonymous synDss statistics for 50 randomly selected core genes of *H. pylori* – 25 recombinant and 25 non-recombinant as defined using PhiPack analysis. **Table S3.** List of 487 non-recombinant core genes under positive selection for accumulating convergent amino acid mutations in the encoded proteins. The genes with '*' and '+' denote the positively selected genes as detected by Montano et al. [26] and Yahara et al. [45] respectively. (PDF 729 kb)

Additional file 2: Figure S1. Frequency of recombinant regions as detected by PhiPack using single (intra-genic) vs. multiple (inter-genic) genes in 'global' subset and 'local' (Amerindian) subset of strains (as denoted in Fig. 1). **Figure S2.** The synDss statistic landscapes for 50 randomly selected core genes of *H. pylori*: (A) 25 recombinant and (B) 25 non-recombinant, as designated by three recombination detection statistics in PhiPack software. **Figure S3.** Average frequency distribution of mutations in encoded proteins of randomly selected subsets of *H. pylori* non-recombinant core genes with convergent mutations. **Figure S4.** Average frequency distribution of mutations of encoded proteins in simulated datasets of 25 *H. pylori* genes (one of the randomly selected replicates from **Figure S3**) using two different mutation-rate constraints. (PDF 2545 kb)

Abbreviations

GO: Gene ontology; GTR: General time reversible; MaxChi: Maximum χ^2; MLST: Multi-locus sequence typing; NSS: Neighbor similarity score; Phi: Pairwise homoplasy index

Acknowledgements

We thank Dr Robert H Gilman and his team for isolation of all Peruvian strains and Dr Dangeruta Kersulyte for most of the complete genome sequences used in the present homoplasy mechanism analysis.

Funding

There are no funders to report for this submission.

Availability of data and materials

The datasets that are used and/or analyzed during the current study but are not included in this published article (and its supplementary information file) are available from the corresponding author on reasonable request.

Authors' contributions

Conceived and designed the experiments: SC, PBC, VNM, DEB, EVS. Performed the experiments: SC, PBC. Analyzed the data: SC, PBC, VNM, DEB, EVS. Wrote the manuscript: SC, DEB, EVS. All authors read and approved the manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Author details

[1]School of Biotechnology, Amrita Vishwa Vidyapeetham, Kollam, Kerala, India. [2]Department of Mathematics and Statistics, Villanova University, Villanova, PA, USA. [3]Department of Statistics, University of California, Irvine, California, USA. [4]Division of Infectious Diseases, Department of Medicine, University of California, San Diego, La Jolla, California, USA. [5]Department of Microbiology, University of Washington, Seattle, Washington, USA.

Received: 11 February 2018 Accepted: 7 November 2018
Published online: 21 November 2018

References

1. Feldman RA, Eccersley AJ, Hardie JM. Epidemiology of *Helicobacter pylori*: acquisition, transmission, population prevalence and disease-to-infection ratio. Br Med Bull. 1998;54:39–53.
2. Algood HMS, Cover TL. *Helicobacter pylori* persistence: an overview of interactions between *H. pylori* and host immune defenses. Clin Microbiol Rev. 2006;19:597–613.
3. Israel DA, Peek RM. The role or persistence in *Helicobacter pylori* pathogenesis. Curr Opin Gastroenterol. 2006;22:3–7.
4. Kansau I, Raymond J, Bingen E, Courcoux P, Kalach N, Bergeret M, Braimi N, Dupont C, Labigne A. Genotyping of *Helicobacter pylori* isolates by sequencing of PCR products and comparison with the RAPD technique. Res Microbiol. 1996;147:661–9.

5. Suerbaum S, Smith JM, Bapumia K, Morelli G, Smith NH, Kunstmann E, Dyrek I, Achtman M. Free recombination within *Helicobacter pylori*. Proc Natl Acad Sci U S A. 1998;95:12619–24.
6. Achtman M, Azuma T, Berg DE, Ito Y, Morelli G, Pan ZJ, Suerbaum S, Thompson SA, van der Ende A, van Doorn LJ. Recombination and clonal groupings within *Helicobacter pylori* from different geographic regions. Mol Microbiol. 1999;32:459–70.
7. Surebaum S, Achtman M. Evolution of *Helicobacter pylori*: the role of recombination. Trends Microbiol. 1999;7:182–4.
8. Falush D, Kraft C, Taylor NS, Correa P, Fox JG, Achtman M, Suerbaum S. Recombination and mutation during long-term gastric colonization by *Helicobacter pylori*: estimates of clock rates, recombination size, and minimal age. Proc Natl Acad Sci U S A. 2001;98:15056–61.
9. Kraft C, Suerbaum S. Mutation and recombination in *Helicobacter pylori*: mechanisms and role in generating strain diversity. Int J Med Microbiol. 2005;295:299–305.
10. Kullick S, Moccia C, Didelot X, Falush D, Kraft C, Suerbaum S. Mosaic DNA imports with interspersions of recipient sequence after natural transformation of *Helicobacter pylori*. PLoS One. 2008;3:e3797.
11. Lin EA, Zhang XS, Levine SM, Gill SR, Falush D, Blaser MJ. Natural transformation of *Helicobacter pylori* involves the integration of short DNA fragments interrupted by gaps of variable size. PLoS Pathog. 2009;5:e1000337.
12. Morelli G, Didelot X, Kusecek B, Schwarz S, Bahlawane C, Falush D, Suerbaum S, Achtman M. Microevolution of *Helicobacter Pylori* during prolonged infection of single hosts and within families. PLoS Genet. 2010;6:e1001036.
13. Dorer MS, Sessler TH, Salama NR. Recombination and DNA repair in *Helicobacter pylori*. Annu Rev Microbiol. 2011;65:329–48.
14. Yahara K, Kawai M, Furuta Y, Takahashi N, Handa N, Tsuru T, Oshima K, Yoshida M, Azuma T, Hattori M, Uchiyama I, Kobayashi I. Genome-wide survey of mutual homologous recombination in a highly sexual bacterial species. Genome Biol Evol. 2012;4:628–40.
15. Hughes AL, Nei M. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature. 1988;335:167–70.
16. Castoe TA, de Koning AP, Pollock DD. Adaptive molecular convergence: molecular evolution versus molecular phylogenetics. Commun Integr Biol. 2010;3:67–9.
17. Christin P-A, Weinreich DM, Besnard G. Causes and evolutionary significance of genetic convergence. Trends Genet. 2010;26:400–5.
18. Tenaillon O, Rodríguez-Verdugo A, Gaut RL, McDonald P, Bennett AF, Long AD, Gaut BS. The molecular diversity of adaptive convergence. Science. 2012;335:457–61.
19. Chattopadhyay S, Weissman SJ, Minin VN, Russo TA, Sokurenko EV. High frequency of hotspot mutations in core genes of *Escherichia coli* due to short-term positive selection. Proc Natl Acad Sci U S A. 2009;106:12412–7.
20. Chattopadhyay S, Paul S, Kisiela DI, Linardopoulou EV, Sokurenko EV. Convergent molecular evolution of genomic cores in *Salmonella enterica* and *Escherichia coli*. J Bacteriol. 2012;194:5002–11.
21. Paul S, Minnick MF, Chattopadhyay S. Mutation-driven divergence and convergence indicate adaptive evolution of the intracellular human-restricted pathogen, *Bartonella bacilliformis*. PLoS Negl Trop Dis. 2016;10:e0004712.
22. Maynard Smith J, Smith NH. Detecting recombination from gene trees. Mol Biol Evol. 1998;15:590–9.
23. Funk DJ, Helbling L, Wernegreen JJ, Moran NA. Intraspecific phylogenetic congruence among multiple symbiont genomes. Proc Biol Sci. 2000;267:2517–21.
24. Meinersmann RJ, Romero-Gallo J, Blaser MJ. Rate heterogeneity in the evolution of *Helicobacter pylori* and the behavior of homoplastic sites. Infect Genet Evol. 2008;8:593–602.
25. Caballero A, Quesada H. Homoplasy and distribution of AFLP fragments: an analysis in silico of the genome of different species. Mol Biol Evol. 2010;27:1139–51.
26. Montano V, Didelot X, Foll M, Linz B, Reinhardt R, Suerbaum S, Moodley Y, Jensen JD. Worldwide population structure, long-term demography, and local adaptation of *Helicobacter pylori*. Genetics. 2015;200:947–63.
27. Israel DA, Salama N, Krishna U, Rieger UM, Atherton JC, Falkow S, Peek RM Jr. *Helicobacter pylori* genetic diversity within the gastric niche of a single human host. Proc Natl Acad Sci U S A. 2001;98:14625–30.
28. Patra R, Chattopadhyay S, De R, Ghosh P, Ganguly M, Chowdhury A, Ramamurthy T, Nair GB, Mukhopadhyay AK. Multiple infection and microdiversity among *Helicobacter pylori* isolates in a single host in India. PLoS One. 2012;7:e43370.
29. Bruen TC, Philippe H. Bryant D. a simple and robust statistical test for detecting the presence of recombination. Genetics. 2006;172:2665–81.
30. Schwarz S, Morelli G, Kusecek B, Manica A, Balloux F, Owen RJ, Graham DY, van der Merwe S, Achtman M, Suerbaum S. Horizontal versus familial transmission of *Helicobacter pylori*. PLoS Pathog. 2008;4:e1000180.
31. Chi PB, Chattopadhyay S, Lemey P, Sokurenko EV, Minin VN. Synonymous and nonsynonymous distances help untangle convergent evolution and recombination. Stat Appl Genet Mol Biol. 2015;14:375–89.
32. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Statist Soc B. 1995;57:289–300.
33. Bowser FP. The African slave in colonial Peru, 1524–1650. Stanford, California: Stanford University Press; 1974.
34. Sokurenko EV, Feldgarden M, Trintchina E, Weissman SJ, Avagyan S, Chattopadhyay S, Johnson JR, Dykhuizen DE. Selection footprint in the FimH adhesin shows pathoadaptive niche differentiation in *Escherichia coli*. Mol Biol Evol. 2004;21:1373–83.
35. Korotkova N, Chattopadhyay S, Tabata TA, Beskhlebnaya V, Vigdorovich V, Kaiser BK, Strong RK, Dykhuizen DE, Sokurenko EV, Moseley SL. Selection for functional diversity drives accumulation of point mutations in Dr adhesins of *Escherichia coli*. Mol Microbiol. 2007;64:180–94.
36. Chattopadhyay S, Feldgarden M, Weissman SJ, Dykhuizen DE, van Belle G, Sokurenko EV. Haplotype diversity in "source-sink" dynamics of *Escherichia coli* urovirulence. J Mol Evol. 2007;64:204–14.
37. Weissman SJ, Beskhlebnaya V, Chesnokova V, Chattopadhyay S, Stamm WE, Hooton TM, Sokurenko EV. Differential stability and trade-off effects of pathoadaptive mutations in the *Escherichia coli* FimH adhesin. Infect Immun. 2007;75:3548–55.
38. Totsika M, Beatson SA, Holden N, Gally DL. Regulatory interplay between pap operons in uropathogenic *Escherichia coli*. Mol Microbiol. 2008;67:996–1011.
39. Chattopadhyay S, Tchesnokova V, McVeigh A, Kisiela DI, Dori K, Navarro A, Sokurenko EV, Savarino SJ. Adaptive evolution of class 5 fimbrial genes in enterotoxigenic *Escherichia coli* and its functional consequences. J Biol Chem. 2012;287:6150–8.
40. Kisiela DI, Chattopadhyay S, Libby SJ, Karlinsey JE, Fang FC, Tchesnokova V, Kramer JJ, Beskhlebnaya V, Samadpour M, Grzymajlo K, Ugorski M, Lankau EW, Mackie RI, Clegg S, Sokurenko EV. Convergent evolution of invasive serovars of *Salmonella enterica* via point mutations in the type 1 fimbrial adhesion FimH. PLoS Pathog. 2012;8:e1002733.
41. Cao Q, Didelot X, Wu Z, Li Z, He L, Li Y, Ni M, You Y, Lin X, Li Z, Gong Y, Zheng M, Zhang M, Liu J, Wang W, Bo X, Falush D, Wang S, Zhang J. Progressive genomic convergence of two *Helicobacter pylori* strains during mixed infection of a patient with chronic gastritis. Gut. 2015;64:554–61.
42. Schrag SJ, Perrot V, Levin BR. Adaptation to the fitness costs of antibiotic resistance in *Escherichia coli*. Proc Biol Sci. 1997;264:1287–91.
43. Björkholm B, Sjölund M, Falk PG, Berg OG, Engstrand L, Andersson DI. Mutation frequency and biological cost of antibiotic resistance in *Helicobacter pylori*. Proc Natl Acad Sci U S A. 2001;98:14607–12.
44. Kawai M, Furuta Y, Yahara K, Tsuru T, Oshima K, Handa N, Takahashi N, Yoshida M, Azuma T, Hattori M, Uchiyama I, Kobayashi I. Evolution in an oncogenic bacterial species with extreme genome plasticity: *Helicobacter pylori* east Asian genomes. BMC Microbiol. 2011;11:104.
45. Yahara K, Furuta Y, Morimoto S, Kikutake C, Komukai S, Matelska D, Dunin-Horkawicz S, Bujnicki JM, Uchiyama I, Kobayashi I. Genome-wide survey of codons under diversifying selection in a highly recombining bacterial species, *Helicobacter pylori*. DNA Res. 2016;23:135–43.
46. Hamilton HL, Dillard JP. Natural transformation of *Neisseria gonorrhoeae*: from DNA donation to homologous recombination. Mol Microbiol. 2006;59:376–85.
47. Carrolo M, Pinto FR, Melo-Cristino J, Ramirez M. Pherotype influences biofilm growth and recombination in *Streptococcus pneumoniae*. PLoS One. 2014;9:e92138.
48. Ogura M, Perez JC, Mittl PR, Lee HK, Dailide G, Tan S, Ito Y, Secka O, Dailidiene D, Putty K, Berg DE, Kalia A. *Helicobacter pylori* evolution: lineage-specific adaptations in homologs of eukaryotic Sel1-like genes. PLoS Comput Biol. 2007;3:e151.

49. Gehmert S, Velapatiño B, Herrera P, Balqui J, Santivañez L, Cok J, Vargas G, Combe J, Passaro DJ, Wen S, Meyer F, Berg DE, Gilman RH. Interleukin-1 beta single-nucleotide polymorphism's C allele is associated with elevated risk of gastric cancer in *Helicobacter pylori*-infected Peruvians. Am J Trop Med Hyg. 2009;81:804–10.
50. Atherton JC, Blaser MJ. Coadaptation of *Helicobacter pylori* and humans: ancient history, modern implications. J Clin Invest. 2009;119:2475–87.
51. Bugaytsova JA, Björnham O, Chernov YA, Gideonsson P, Henriksson S, Mendez M, Sjöström R, Mahdavi J, Shevtsova A, Ilver D, Moonens K, Quintana-Hayashi MP, Moskalenko R, Aisenbrey C, Bylund G, Schmidt A, Åberg A, Brännström K, Königer V, Vikström S, Rakhimova L, Hofer A, Ögren J, Liu H, Goldman MD, Whitmire JM, Ådén J, Younson J, Kelly CG, Gilman RH, Chowdhury A, Mukhopadhyay AK, Nair GB, Papadakos KS, Martinez-Gonzalez B, Sgouras DN, Engstrand L, Unemo M, Danielsson D, Suerbaum S, Oscarson S, Morozova-Roche LA, Olofsson A, Gröbner G, Holgersson J, Esberg A, Strömberg N, Landström M, Eldridge AM, Chromy BA, Hansen LM, Solnick JV, Lindén SK, Haas R, Dubois A, Merrell DS, Schedin S, Remaut H, Arnqvist A, Berg DE, Borén T. *Helicobacter pylori* adapts to chronic infection and gastric disease via pH-responsive BabA-mediated adherence. Cell Host Microbe. 2017;21:376–89.
52. Swofford DLPAUP. Phylogenetic analysis using parsimony and other methods. Version 4. Sunderland, MA: Sinauer Associates; 2000.
53. Chattopadhyay S, Paul S, Dykhuizen DE, Sokurenko EV. Tracking recent adaptive evolution in microbial species using TimeZone. Nat Protoc. 2013;8:652–65.
54. Nei M, Gojobori T. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol. 1986;3:418–26.
55. Chan CX, Beiko RG, Ragan MA. A two-phase strategy for detecting recombination in nucleotide sequences. South African Comp J. 2007;38:20–7.
56. Hall BG. Simulating DNA coding sequence evolution with EvolveAGene 3. Mol Biol Evol. 2008;25:688–95.
57. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2009;4:44–57.