Supplementary Information for:

Recombination Resolves the Cost of Horizontal Gene Transfer in Experimental Populations of *Helicobacter pylori*

Authors: An N.T Nguyen[1†], Laura C, Woods[1†], Rebecca Gorrell[2,3,4], Shamitraa Ramanan[1], Terry Kwok[2,3,4], Michael J McDonald[1*]

Affiliations:

1. School of Biological Sciences, Monash University: Clayton, Wellington Road, Clayton, Vic300, Australia
2. Department of Biochemistry and Molecular Biology, Monash University: Wellington Road, Clayton, Vic300, Australia
3. Department of Microbiology, Monash University: Wellington Road, Clayton, Vic300, Australia
4. Biomedical Discovery Institute, Monash University: Wellington Road, Clayton, Vic300, Australia

*Corresponding author. Email: mike.mcdonald@monash.edu

This file contains Supplementary Figures S1-10, and an Appendix: Validation of the HGT variant calling pipeline.

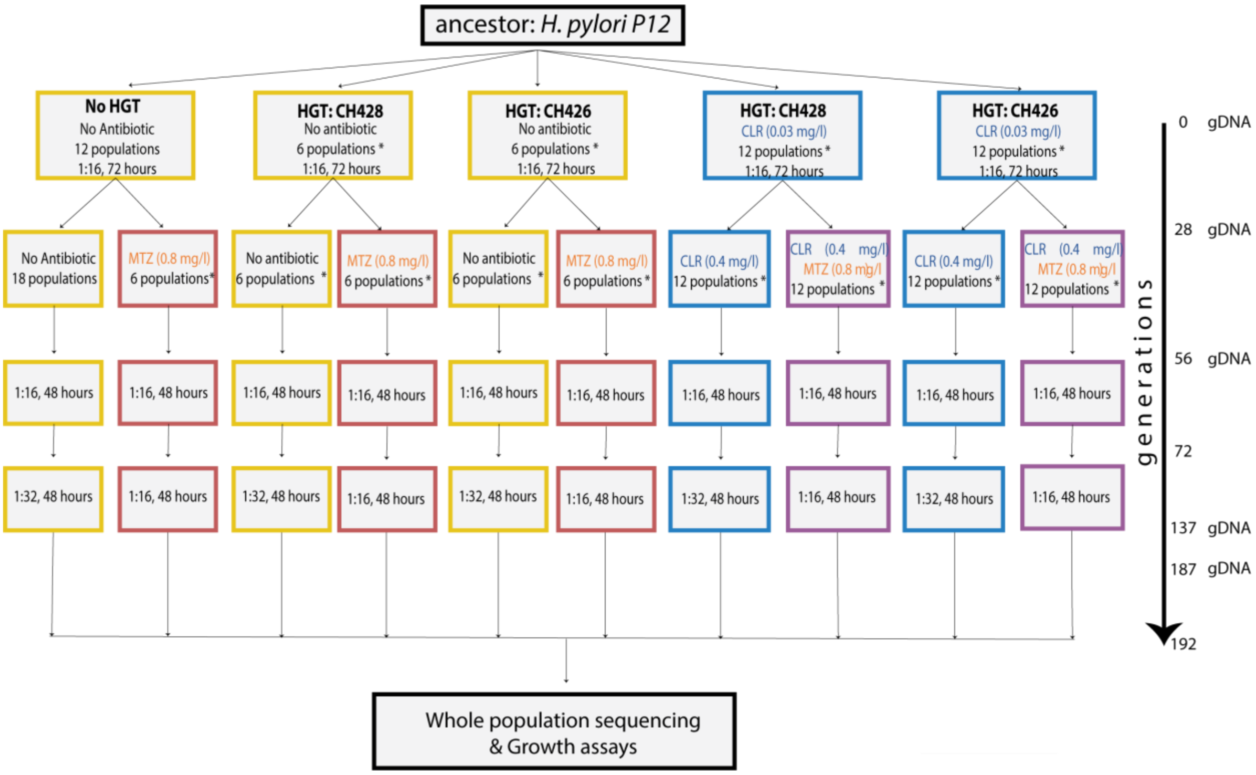


Fig. S1.
Flow diagram showing dilution rates and antibiotic concentration increments during experimental evolution treatments. To ensure that experimental cultures were not over-diluted (leading to extinction) or under-diluted (slowing the chronological rate of adaptation), the rates of dilution and concentrations of antibiotics were modified during the course of the experiment. While all non-HGT treatments evolved in clarithromycin or clarithromycin + metronidazole conditions went extinct, two of the non-HGT treatments propagated in metronidazole survived and evolved resistance to metronidazole. The generations where dilution rates or antibiotic concentrations were varied are indicated on the right. The generations where donor genomic DNA (gDNA) was added are also shown. The dilution rates are indicated by ratios. For example, 1:16, 48 hours indicates a 16-fold dilution once every 48 hours. * indicates each donor strain.

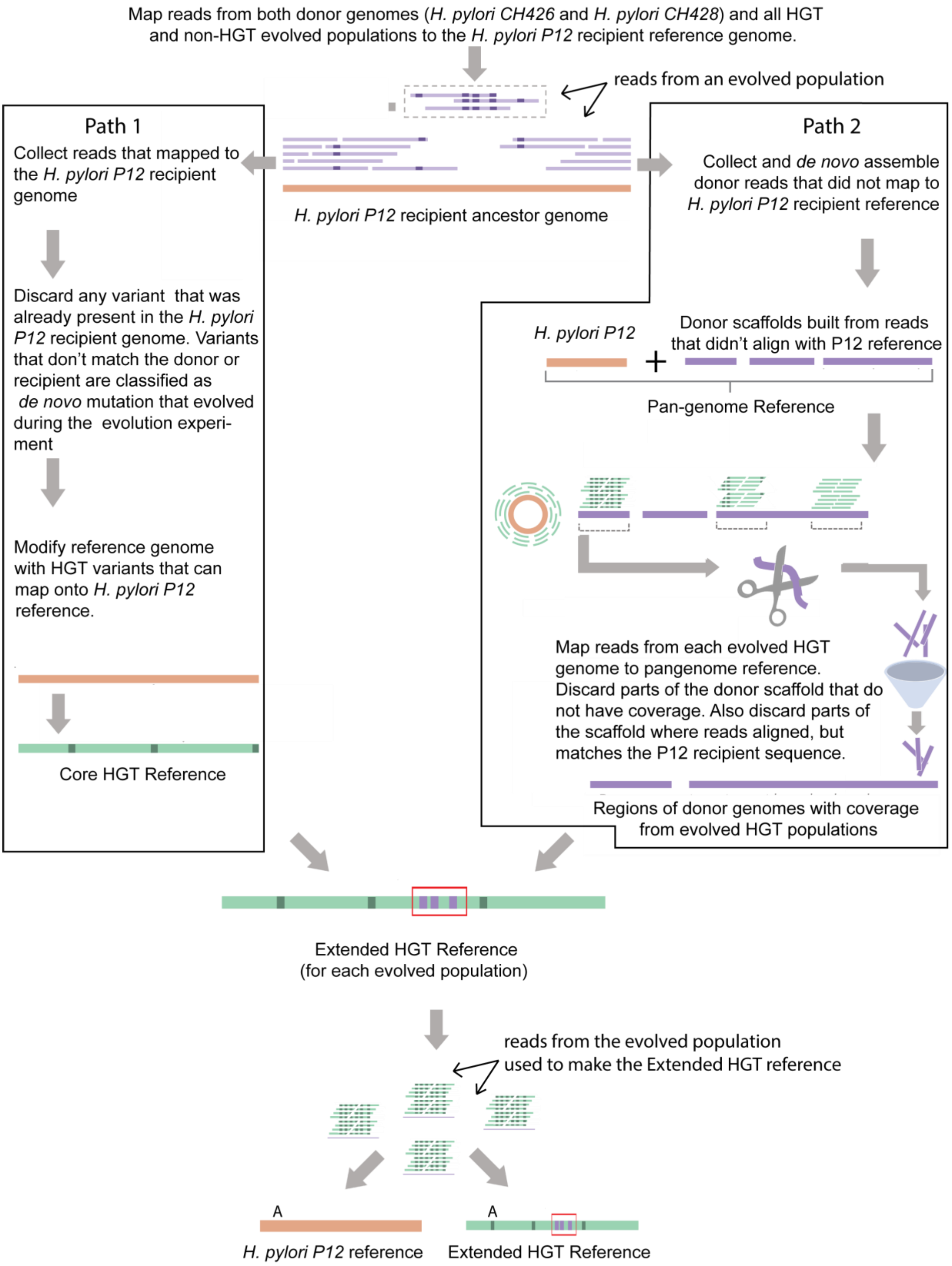

Map reads from both donor genomes (*H. pylori CH426* and *H. pylori CH428*) and all HGT and non-HGT evolved populations to the *H. pylori P12* recipient reference genome.
reads from an evolved population
H. pylori P12 recipient ancestor genome
Path 1
Collect reads that mapped to the *H. pylori P12* recipient genome
Discard any variant that was already present in the *H. pylori P12* recipient genome. Variants that don't match the donor or recipient are classified as *de novo* mutation that evolved during the evolution experiment
Modify reference genome with HGT variants that can map onto *H. pylori P12* reference.
Core HGT Reference
Path 2
Collect and *de novo* assemble donor reads that did not map to *H. pylori P12* recipient reference
H. pylori P12
Donor scaffolds built from reads that didn't align with P12 reference
Pan-genome Reference
Map reads from each evolved HGT genome to pangenome reference. Discard parts of the donor scaffold that do not have coverage. Also discard parts of the scaffold where reads aligned, but matches the P12 recipient sequence.
Regions of donor genomes with coverage from evolved HGT populations
Extended HGT Reference (for each evolved population)
reads from the evolved population used to make the Extended HGT reference
A
A
H. pylori P12 reference
Extended HGT Reference

Fig. S2. Generating donor-recipient hybrid reference genomes and calling variants for each evolved HGT treatment population. Path 1 shows how a core genome reference was

constructed by incorporating all possible variants from the donor genomes that can map onto the *H. pylori P12* genome. This is referred to as the “core genome reference” since this reference contains genes shared by the recipient and donor strains. Path 2 includes regions of the donor genomes that are highly diverged from the *H. pylori P12* genome, including accessory genes that do not have orthologues in the *H. pylori P12* genome. Every sequenced HGT genome has its own Extended HGT reference. To call variants that have evolved in that particular population, reads from evolved population were aligned to its extended HGT reference and the *H. pylori P12* reference. Each read should map to one reference genome but not the other. The frequency of the HGT variant were determined by comparing coverage at the wt. equivalent site on the reference genome. For example, if an HGT originated variant at site A on the HGT genome has 300-fold coverage, and the wt. equivalent variant at site x also has 300-fold coverage, the HGT allele frequency is 0.5, or 50%.

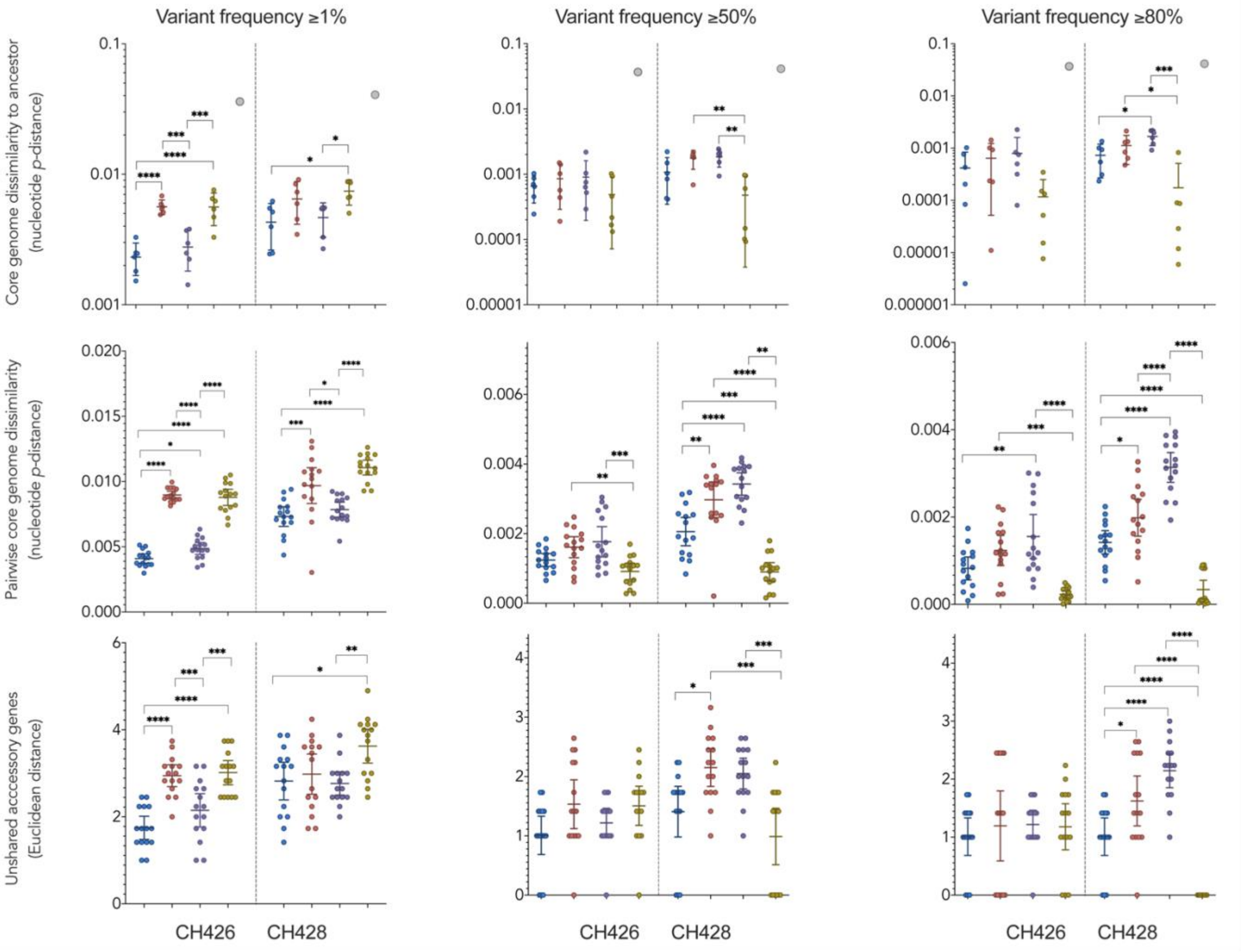


Fig. S3. The divergence of evolved populations from *H. pylori P12*.

Top row: Nucleotide *p*-distance between each replicate population and the *H. pylori* P12 recipient. We include only core-genome differences attributed to horizontally transferred variants at or above 1% frequency (**a**), 50% frequency (**b**), and 80% frequency (**c**). Donor p-distances to the ancestor are also indicated for comparison (grey circles). Note that the core genome is comprised of open reading frames (ORFs) that are shared by the two donor strains and *H. pylori* P12. ORFs are considered as “shared” if they have a 90% BLASTP identity or above. Middle row: Group pairwise (number of pairs given six replicates per treatment = 15) nucleotide *p*-distance (per-base difference) due to horizontally transferred variants at or above 1% frequency (**d**), 50% frequency (**e**), and 80% frequency (**f**) in the core genome. Bottom row: Group pairwise number of pairs given six replicates per treatment = 15) unshared accessory gene content (Euclidean distance, where each accessory gene is either present or absent) due to horizontally transferred variants at or above 1% frequency (**g**), 50% frequency (**h**), and 80% frequency (**i**). Accessory genes are those which are not shared between all donors, all evolved populations, and the ancestor according to 90% BLASTP cutoff.

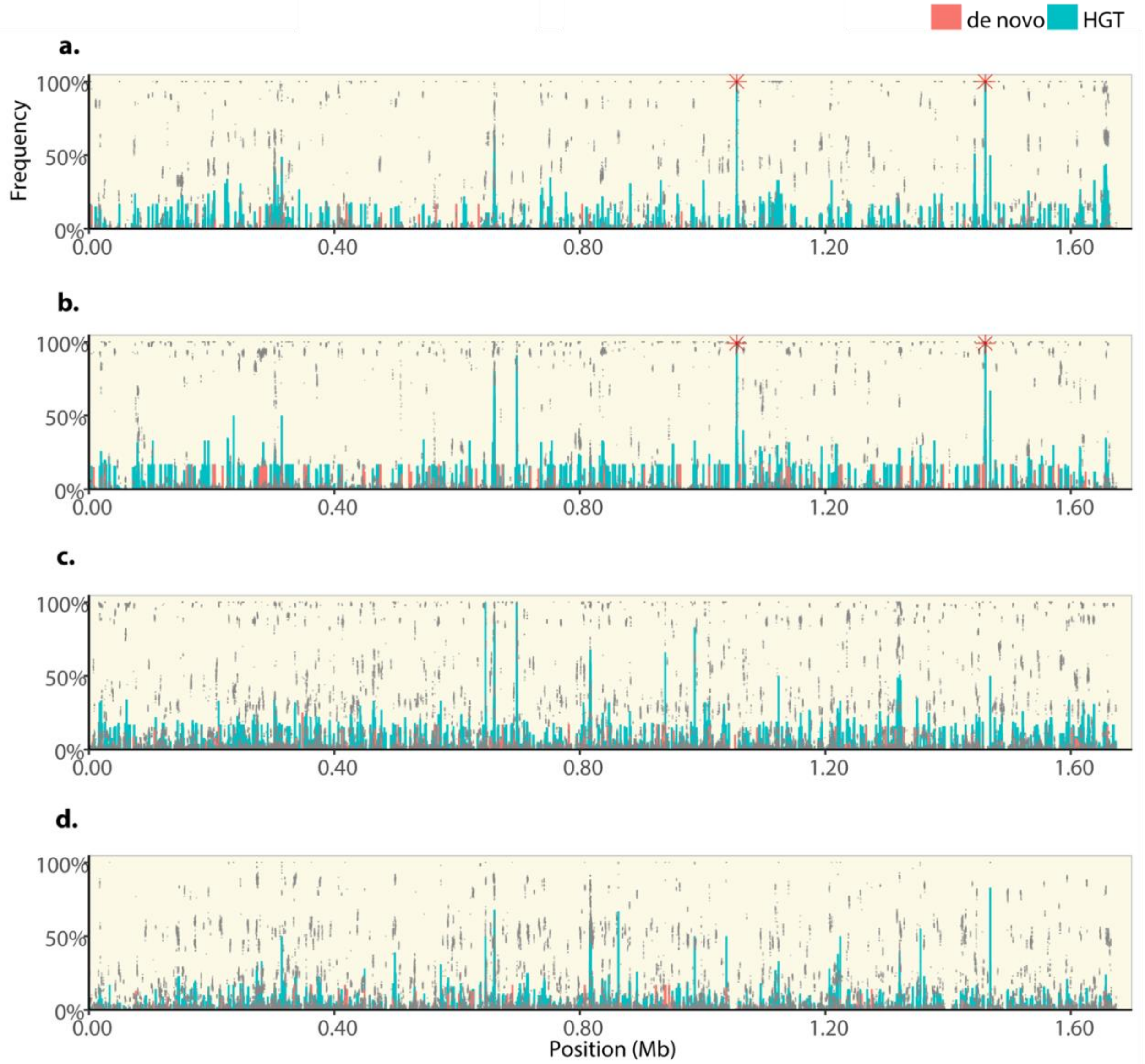


Fig. S4. The distribution of HGT events across the *H. pylori* genome populations evolving with HGT from CH426. Each panel shows the frequency of genetic variants along the length of the *H. pylori* 12 genome. Populations evolving with HGT from the CH426 donor, in growth media with clarithromycin and metronidazole (a), clarithromycin (b),

metronidazole (c) or in growth media without antibiotic (d). Each panel shows sequencing data from 6 replicate populations, including data points from each individual population (grey dots) and the average frequency of each variant for HGT (green bars) and *de novo* (red bars) genetic variants. Red stars indicate 23S mutations.

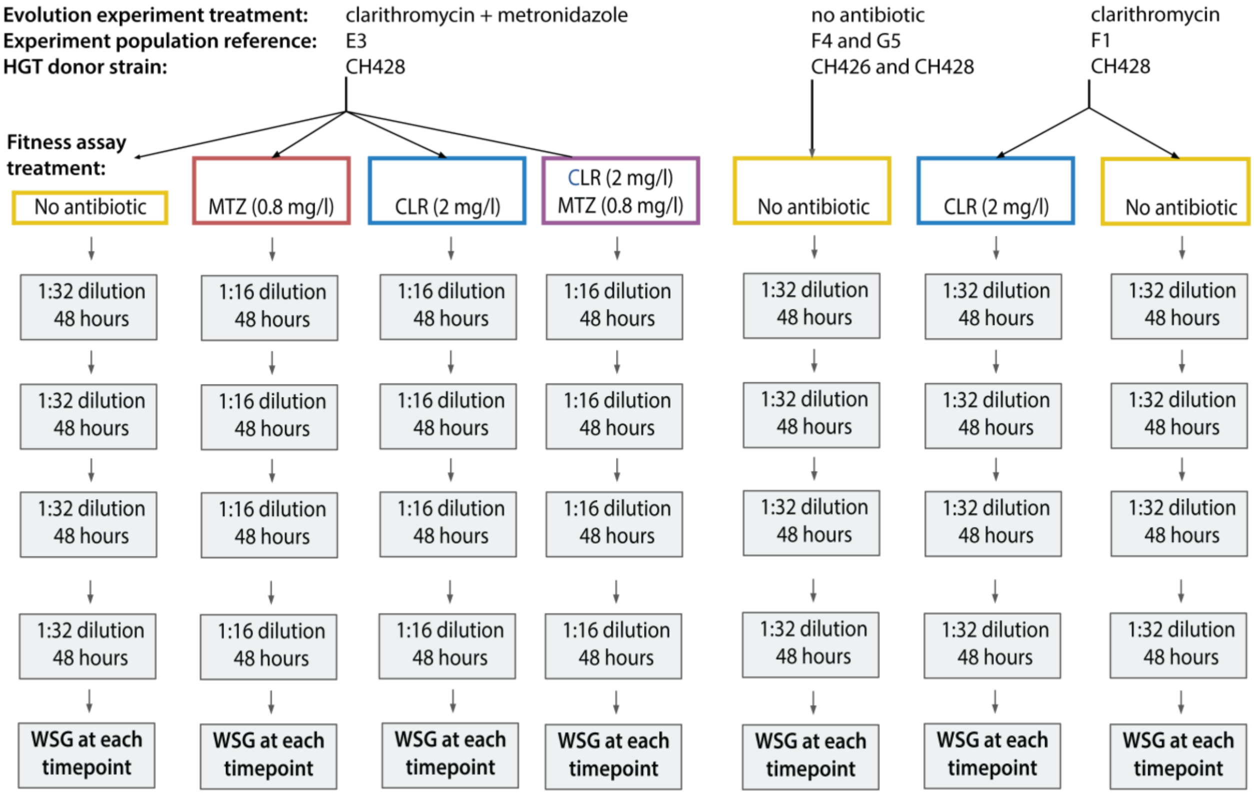
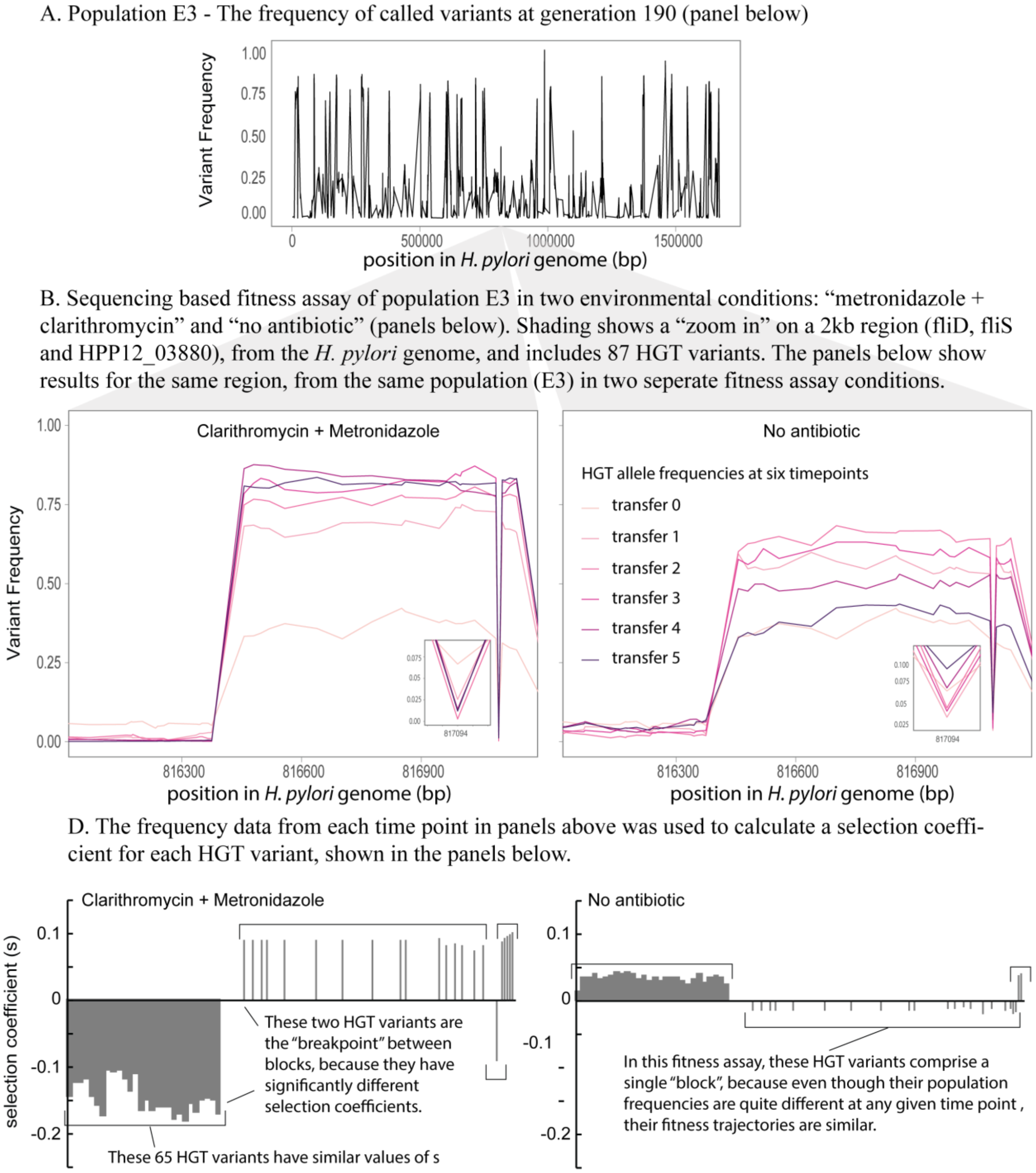


Fig. S5 Flow diagram showing an overview of the sequencing-based fitness assay's populations and experimental treatments.


A. Population E3 - The frequency of called variants at generation 190 (panel below)

B. Sequencing based fitness assay of population E3 in two environmental conditions: “metronidazole + clarithromycin” and “no antibiotic” (panels below). Shading shows a “zoom in” on a 2kb region (fliD, fliS and HPP12_03880), from the *H. pylori* genome, and includes 87 HGT variants. The panels below show results for the same region, from the same population (E3) in two seperate fitness assay conditions.

D. The frequency data from each time point in panels above was used to calculate a selection coefficient for each HGT variant, shown in the panels below.

These two HGT variants are the “breakpoint” between blocks, because they have significantly different selection coefficients.

These 65 HGT variants have similar values of s

In this fitness assay, these HGT variants comprise a single “block”, because even though their population frequencies are quite different at any given time point , their fitness trajectories are similar.

E. Variants that are next to each other and do not have sigificantly different “s” are assigned into the same block. Bracketed regions show seperate blocks. Note that in the clarithromycin+metronidazole fitness assay the HGT variants were assigned into four blocks, while in the fitness assay without antibiotic, the HGT variants were reassigned into three different blocks. These two panels correspond to Figure 6d & 6e in the main text..

Fig. S6. Schematic of converting whole population sequencing at multiple time points into fitness data, and “blocks”.

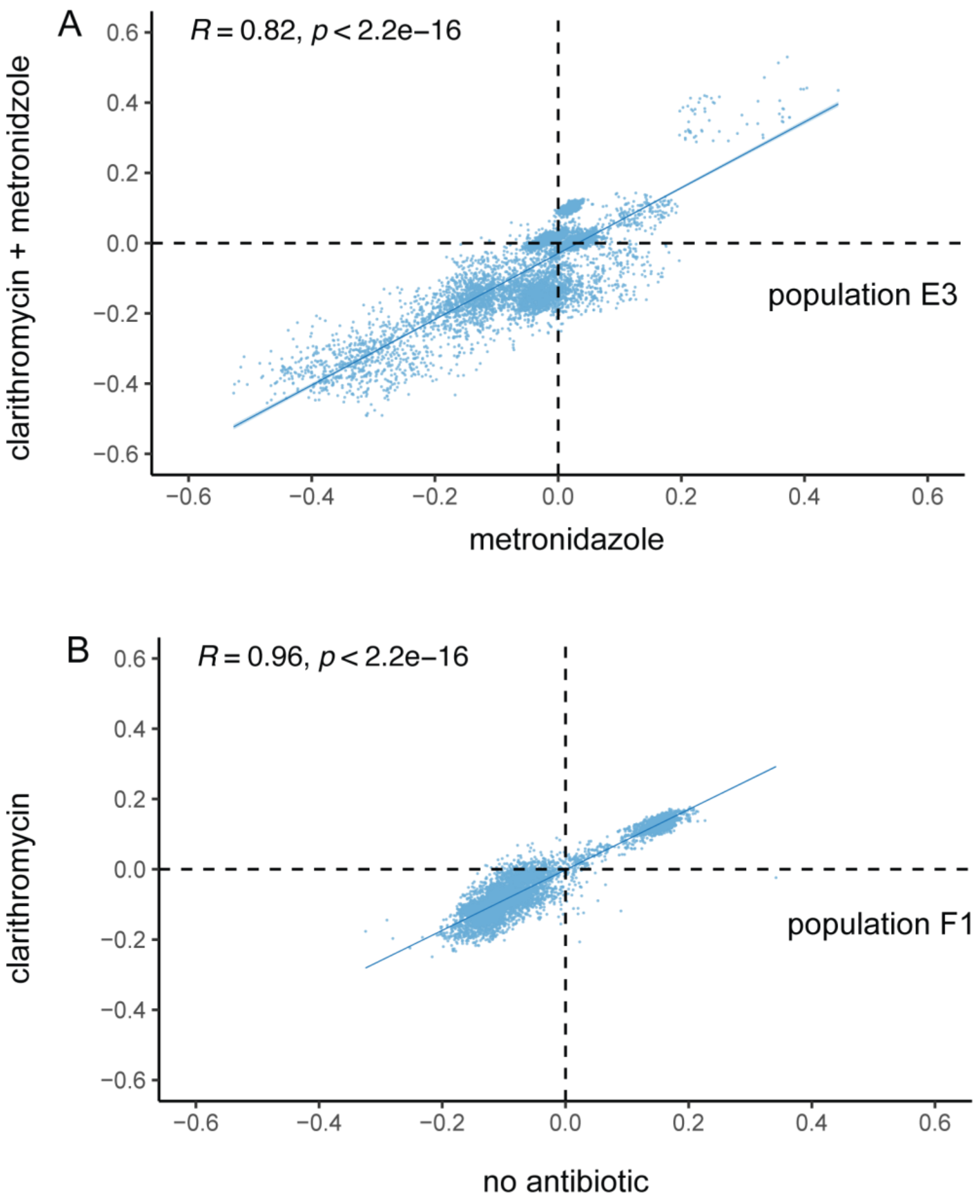


Fig. S7 Correlation of fitness measurements for the same sets of genetic variants measured in two growth conditions. Panel A shows fitness measurements (selection coefficient, s) for HGT treatment population E3 which evolved in growth media with clarithromycin and metronidazole. The fitness in clarithromycin and metronidazole is shown on the y axis, and the measurement in metronidazole are shown on the x-axis. Panel B shows fitness measurements (s) for HGT treatment population F1 which evolved in growth media with clarithromycin. The fitness in clarithromycin is shown on the y-axis, and the

measurement in growth media without antibiotic are shown on the x-axis. R values calculated using Pearson's correlation coefficient.

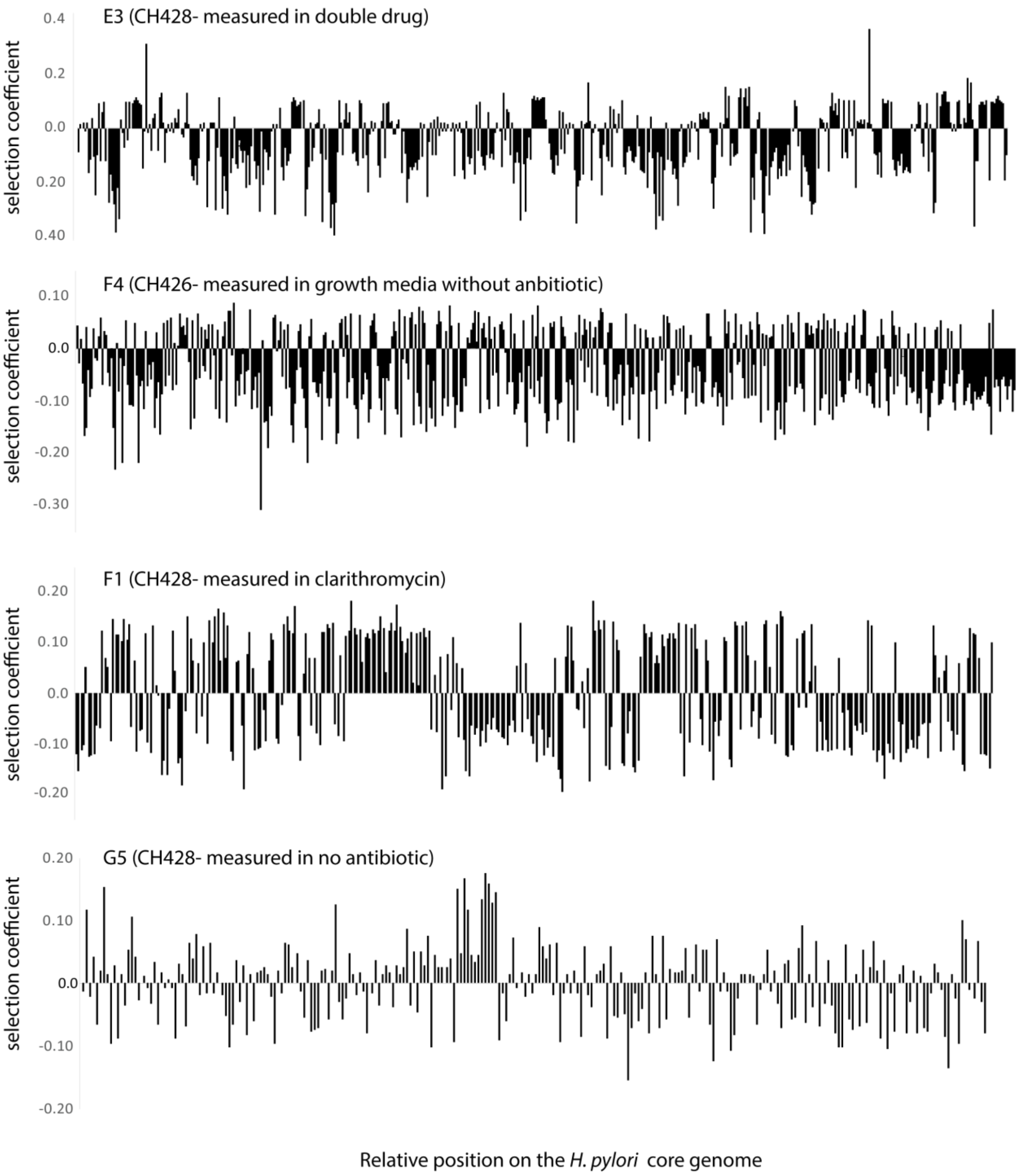


Fig. S8 Gene blocks in evolved HGT treatment populations carry a wide range of selective effects. Gene blocks were determined using a custom script that identified linked groups of horizontally transferred variants (methods). The fitness effect of each block was taken as the average fitness of each variant within the block, and the average fitness effects of all blocks is shown by black horizontal lines. Plotting these blocks shows that strongly gene blocks with strongly deleterious fitness effects are interspersed by gene blocks with strongly beneficial fitness effects.

Appendix.

Validation of HGT variant calling pipeline.

In this study, we describe a bioinformatics pipeline for calling SNP's in population sequence data for populations of *H. pylori* that have evolved with Horizontal Gene Transfer. In part I of this appendix, we describe computer simulations that were designed to compare the performance of our variant calling pipeline to a perfectly understood *in silico* population of *H. pylori* P12 that had received simulated (biologically plausible) horizontal gene transfer from the donor *H. pylori* CH426. Then, in part II, we describe the iterative process that we used to improve the capacity to call the frequency of HGT variants in the sequencing based fitness assay, and confirm that the extended HGT genome references (Fig. S2) were capturing a significant proportion of HGT variants in each experimental population.

Part I: *In-silico* simulations of evolved *H. pylori P12* populations with horizontal gene transfer from *H. pylori CH426*.

Approach: We simulated populations of *H. pylori* P12 that had undergone horizontal gene transfer from the *H. pylori* CH426 donor. The aim was to have a continuous distribution of HGT variant frequencies, around 500-fold coverage and recombination breakpoints that mapped to homologous regions between the two genomes. We started with a pangenome analysis that described the gene content differences between the CH426 donor and P12 recipient from Roary to determine the most likely sites of recombination. The aim was to avoid making genomes with recombination of identical fragments, but wanted enough homology that recombination was possible as well as including some accessory gene transfer.

Methods: 1. Identify all core genes (genes shared by both recipient and donor genomes with a 95% identify cut-off) where the length of the donor gene was unequal to the length of the recipient's version of the gene. This ensured that there was enough homology to simulate a reasonable site of HGT integrations, without having the HGT event result in the transfer of identical an DNA sequence.

2. Group the two donor genes immediately upstream and downstream of this core gene to make a trio. These other genes can include identical genes, or accessory genes (genes found in the donor but not the recipient).

3. For trios that overlap, collect into a single block. This produced around 180 blocks. In our real evolution experiment data, we found many more blocks (~500) of smaller size. The method used to ensure that each simulated block includes at least one equal core gene as well as genes of other classifications (identical, accessory) meant that the blocks were quite large and only so many were possible given the small size of the *H. pylori* genome. This could be a drawback, though sequence divergence of the in-silico genomes could still be high.

4. Next, we randomly selected of these 150 blocks for integration into the *H. pylori P12* recipient genome.

5. To do this, the recipient genes that had corresponding donor genes (shared core genes) included in these blocks, were replaced with the donor gene. The integration site was determined by the gene order in the recipient.

6. To find integration sites for accessory donor genes, we located block members and replaced the accessory *H. pylori* P12 gene, if there is one, while retaining the gene order within the parent genome. For example, donor gene 5 is an accessory gene and therefore doesn't have a corresponding recipient gene. To figure out where this gene should be inserted by homologous recombination (if it were taken up by HGT), we would find out where either donor gene 4 or 6 is in the recipient genome by looking for their orthologs. One of the constraints built into the definition of "blocks" is that at least one of the genes will not be an accessory gene and will be locatable in the recipient genome. An example of a common

result would be to see donor gene 7, donor gene 6, recipient gene 117, donor gene 4, so simply replace recipient gene 117 with donor gene 5. Another possibility is that there is no *H. pylori P12* accessory gene but there is an obvious insertion point. An example would be the insertion of accessory donor gene 1001 in a block that includes donor gene 999, donor gene 1000, donor gene 1002 and donor gene 1003. Donor accessory genes that do not have an obvious "fit" within the genome are excluded. This incorporates the biologically realistic requirement for sequence homology for DNA integration via homologous recombination. This results in some refining of the block definitions and a final list of genes representing HGT events from the donor to the recipient genome.

7. Next, we wanted to make a population representing different combinations of these blocks. The list of genes representing the HGT genome were divided into several parts, where blocks were redefined as the group of donor genes as well as all recipient genes upstream of them. The corresponding blocks were identified in the recipient genome as well, so that each block had both an HGT (mixed donor and recipient) and ancestral version (only recipient) version, giving two possible groups of genes for each of "n" blocks. These gene lists can be found in Table 1 of Dataset 3. A custom script was written that would randomly select either the ancestral or HGT version of the gene for each block, with the probability favouring the selection of the ancestral block (70% to 30%) to get a low-intermediate amount of HGT. This was repeated five times to produce five lists of genes representing five genomes, excluding intergenic regions (Supplementary Dataset 3). Because HGT blocks are selected randomly, blocks are expected to overlap between genomes. An example of the full gene arrangements from a simulation run is provided in Dataset 3.

8. The gene lists were converted into nucleotide sequences. The resultant FASTA sequences are supplied as Supplementary Dataset 4.

9. Next, we generated reads with 100x depth for each of the five HGT genomes using randomreads from the BBMap suite of tools (default settings, with no additional mutations, default Illumnia sequencing error rate, 150 bp, average Phred quality ≥ 30). Each of the five simulated genomes therefore represented 20% of the population, but because the blocks overlap between genomes and are selected randomly, HGT block frequencies will be either 0, 20, 40, 60, 80 or 100%. Since block selection was weighted toward ancestral blocks, most HGT blocks will be of a lower frequency. See Supplementary Dataset 3 for block overlaps between the genomes.

10. Next we carried out a Mauve alignment between the ancestor (*H. pylori P12* chromosome with no intergenic sequences) and each HGT genome to get the SNP and indel calls. These are given in Dataset 3, with the sheets marking each output. Note that Mauve reports indels as gaps with reference to each of the sequences.

11. Then, we perform the HGT identification protocol for the generated reads, starting with alignment to the ancestral (*H. pylori* P12 chromosome with no intergenic sequences) sequence.

12. Compare the variants identified using our pipeline to known frequency of variants generated in step 9.

SNP discovery	
Total SNPs found in in silico population (observed)	12631
Total # SNPs from combined genome (expected)	18651
% SNPs	0.677
False SNPs	400
False SNPs at >= 1%	377
% false SNPs	0.031
False SNPs >= 1%	0.029
SNP frequency paired t test (observed vs expected	
P value	<0.0001
P value summary	****
Significantly different (P < 0.05)?	Yes

One- or two-tailed P value?	Two-tailed
t, df	t=4.177, df=12230
Number of pairs	12231
Mean of differences (B - A)	-0.00152
SD of differences	0.0404
SEM of differences	0.000365
95% confidence interval	-0.002242 to -0.00081
R squared (partial eta squared)	0.00142

Table S1. Performance of HGT SNP identification pipeline compared to known frequency of HGT SNPs in the *in silico* generated genomes with HGT from *H. pylori* CH426.

Indel discovery	
Total indels found in in silico population (observed)	83
Total # indels from combined genome (expected)	585
% indels found	0.141
False indels	36
False indels at >= 1%	31
% false indels	0.433
False indels >= 1%	0.373
Indel frequency paired t test (observed vs expected)	
P value	0.616
P value summary	ns
Significantly different (P < 0.05)?	No
One- or two-tailed P value?	Two-tailed
t, df	t=0.504, df=46
Number of pairs	47
Mean of differences (B - A)	-0.0101
SD of differences	0.138
SEM of differences	0.02013
95% confidence interval	-0.0506 to 0.0303
R squared (partial eta squared)	0.00549

Table S2. Performance of HGT indel identification pipeline compared to known frequency of HGT indels in the *in silico* generated *H. pylori* P12 genomes with HGT from *H. pylori* CH426.

Conclusions: We found 67% of SNPs were identified with the first iteration. Our false discovery rate and incorrect annotation rates for SNPs were both <0.05. While the observed frequency of SNPs was significantly less that the expected frequency, the magnitude of this difference was very small: we underestimated the frequency of SNPs by -0.001526. Our capacity to identify indels was quite poor relative to SNPs, although the frequency of identified indels was not significantly different from expectations, we underestimated indel frequency by -0.01015. Indel identification is known to be difficult with short read data and higher error rates are expected compared to SNPs.

Part II: Iterative test of the extended HGT Reference Genome to refine estimates of HGT variant frequencies in population sequence data.

Approach: Part of the process of accurately identifying HGT variants in the evolved populations is to build a hybrid reference genome comprised of the *H. pylori P12* reference sequence and those parts of the donor genome that had been identified as having coverage in an evolved HGT treatment population (Fig. S2). The creation of this reference facilitated more accurate calls for individual HGT variant frequencies. This is because variants near donor-recipient junctions (integration sites for HGT events) are more likely to be underestimated since the reads that cover those regions are more likely to be hybrid reads containing regions that align with both the donor and recipient genomes. We noticed that the bam files following the final alignment of the evolved short reads to the ancestral+plasmid+maxHGT genome sequences (Fig. S2) revealed previous undetected HGT variants discovered around the HGT junction regions. We ruled out that these were evolved reads corresponding to the ancestral sequence that were aligning to the HGT events, by finding that these reads were matching perfectly at the HGT event positions. However, we found that the variants were occurring at sequences that had not been incorporated into the maxHGT genome, so they were actually indicative of further HGT that had not been discovered. To maximise the number of detected HGT variants and to accurately measure the frequency of variants, we used these newly discovered HGT variants to improve the extended HGT reference genome, and then repeated the alignment of the evolved reads. To determine how many iterations of this process we should carry out, we determined whether successive iterations facilitated new HGT variant discovery and whether doing so would lead to changes in estimation of block size and the estimation of fitness from sequence data taken at multiple time points.

Methods

1. First, we selected one of the populations that had been used in the sequencing based fitness assays—in this case, 426F4.
2. We aligned the short reads for the appropriate donor (CH426) against the P12+plasmid+426F4maxHGT reference that had been generated (see path 2 in Fig. S2). This was also done for the ancestral reads as a control—the expectation is that reads from the ancestral P12 genome should only align to P12+plasmid, but not to the 426F4maxHGT genome.
3. The variants identified in these runs were compared to those identified when the population short reads had been run against this reference set (this had been completed previously as part of the workflow). Variants in common with the donor appearing on the maxHGT genome would indicate HGT events, while those that were in common with the ancestor would indicate artefacts from read mismapping. The HGT events identified at this stage were then applied to the maxHGT genome to update the P12+plasmid+maxHGT genome reference set.
4. Short reads from 426F4, the donor (CH426) and the ancestor (P12, including its plasmid) were aligned using breseq against the new reference set.
5. Steps 3 and 4 were repeated until the newly identified HGT variants were fewer than 10 (6 rounds).
6. The short reads for the 42F4 fitness assay timepoints (T0,T1,T3,T5) were aligned to the final reference set.

7. The HGT events for all 426F4 sequences were quantified as in the end of the general HGT identification workflow (ie using Mauve to identify reciprocal nucleotide changes between the ancestor and maxHGT genome and find the read coverage based on the bam file from the breseq output).
8. The selection coefficients for all HGT variants were calculated and blocks determined as previously described. The ancestral and de novo variants would not have changed across iterations (because the P12 sequence wasn't altered), so the ancestral and *de novo* variants were kept the same.

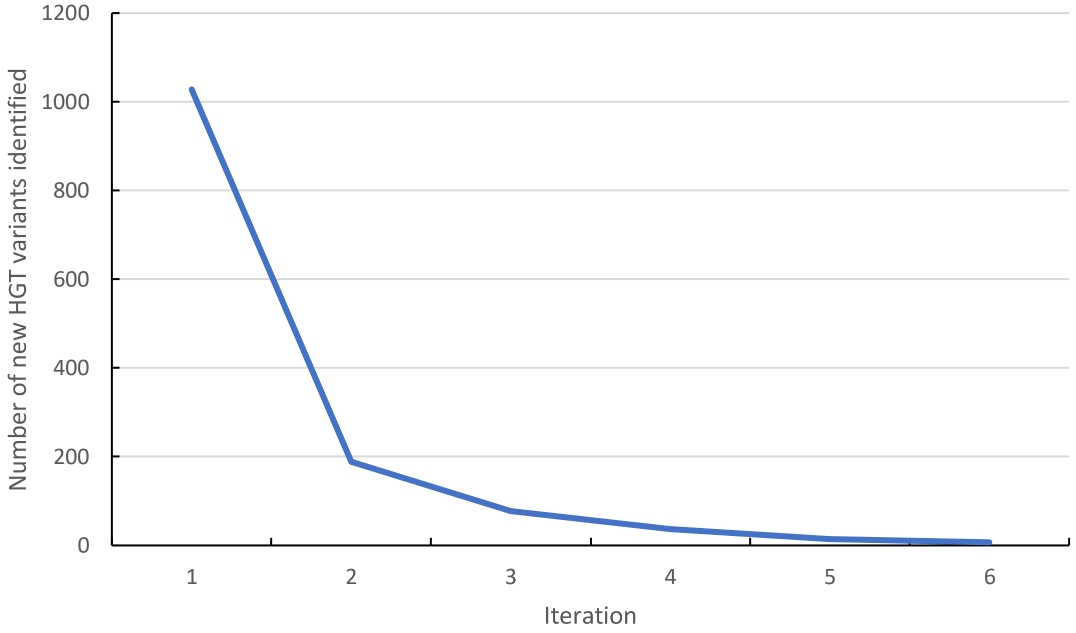


Figure S9. Number of HGT variants identified with each iteration.

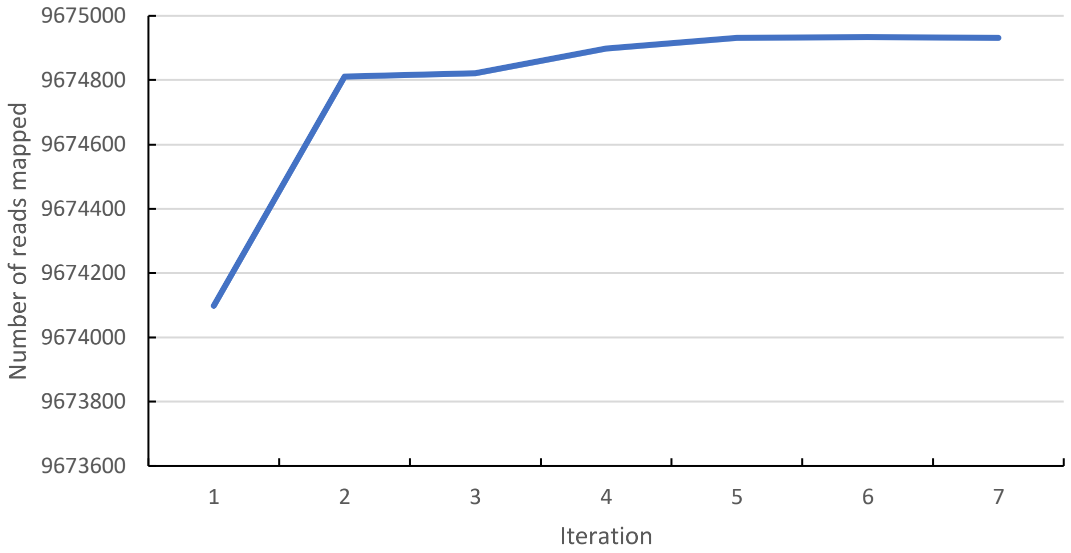


Figure S10. Total number of evolved population reads that mapped to the HGT reference genome at each iteration. Note that the y-axis does not start at 0.

Conclusions: We found diminishing returns with each iteration of improving the HGT reference genome (Fig. S9), with each round resulting in the discovery of fewer HGT events (Fig. S10). At each iteration, a control experiment was carried out by aligning reads from the *H. pylori* P12 ancestor. No variants identified by read mapping were shared between either the evolved population or donor, and the P12 ancestor. This supports that the iterations were discovering HGT variants, at each iteration and not haplotypes from the *H. pylori* P12 reference genome. Moreover, we found that 76% of SNPs that could be discovered after 7 iterations were discovered in the first iteration. Altogether, these data support that while the max HGT genome increased the number of discovered HGT variants, two iterations of this process were sufficient to discover ~ 90% of all possible HGT snps. In the original fitness assay analysis, we had included additional variants beyond those identified in the evolutionary endpoints found in the time-point populations when they were aligned against the corresponding P12+plasmid+maxHGT genome reference set with breseq as “misc”. These suspected HGT “misc” variants were included for downstream analysis of the recombinative blocks based on selection coefficient. Therefore, most of the HGT events discovered via the iterative method overlapped with “misc” variants discovered later upon alignment of the populations representing fitness assay time-points to the corresponding reference, and simply led to these variants being reclassified, but as all variant types must be considered in the generation of the selection coefficient blocks, this reclassification had no impact on the distribution of fitness effects of the size distribution of blocks.