
Supplementary information

Brief antibiotic use drives human gut bacteria towards low-cost resistance

In the format provided by the
authors and unedited

TABLE OF CONTENTS

Supplementary Figure 1. Genome quality.

Supplementary Figure 2. Diverse response trends within the *Lachnospiraceae* family.

Supplementary Figure 3. Genome crowding.

Supplementary Figure 4. Linkage group size.

Supplementary Figure 5. Variants per gene.

Supplementary Figure 6. Separating the homologs *gyrA* and *parC*.

Supplementary Figure 7. KEGG orthology table.

Supplementary Figure 8. Examples of evolving populations associated with a mutation in *gyrA*.

Supplementary Figure 9. Examples of evolving populations not associated with *gyrA*.

Supplementary Figure 10. Converging evolution controls.

Supplementary Figure 11. Pre-exposure resistant strains across subjects.

Supplementary Figure 12. *GyrA* evolvability models.

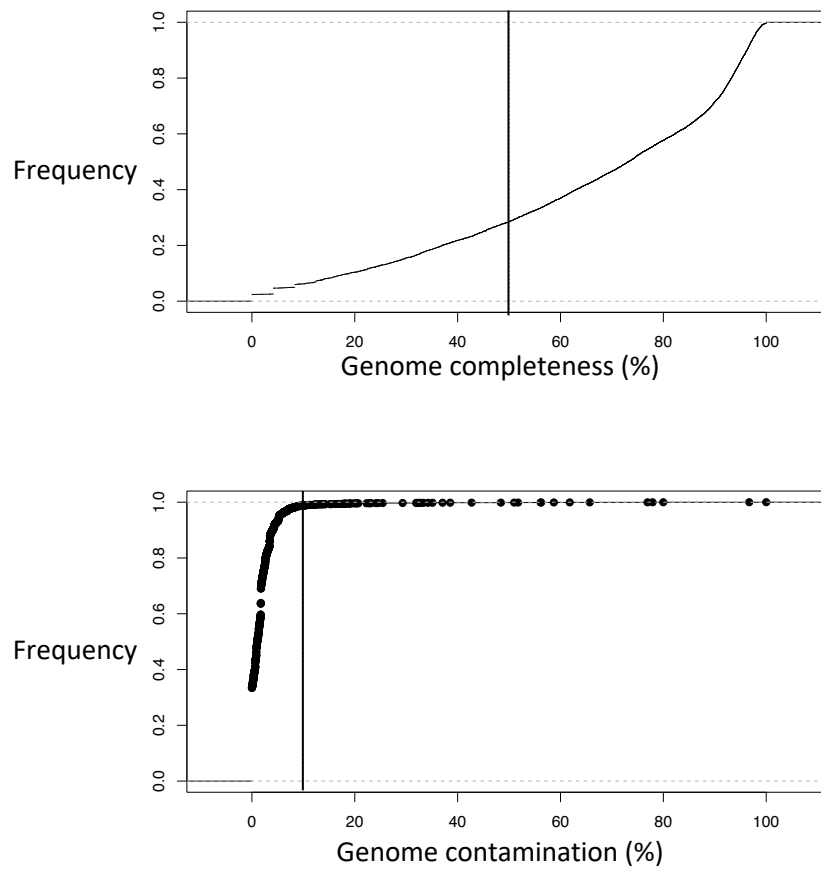
Supplementary Figure 13. *Non-GyrA* evolvability.

Supplementary Figure 14. Evolving populations across subjects.

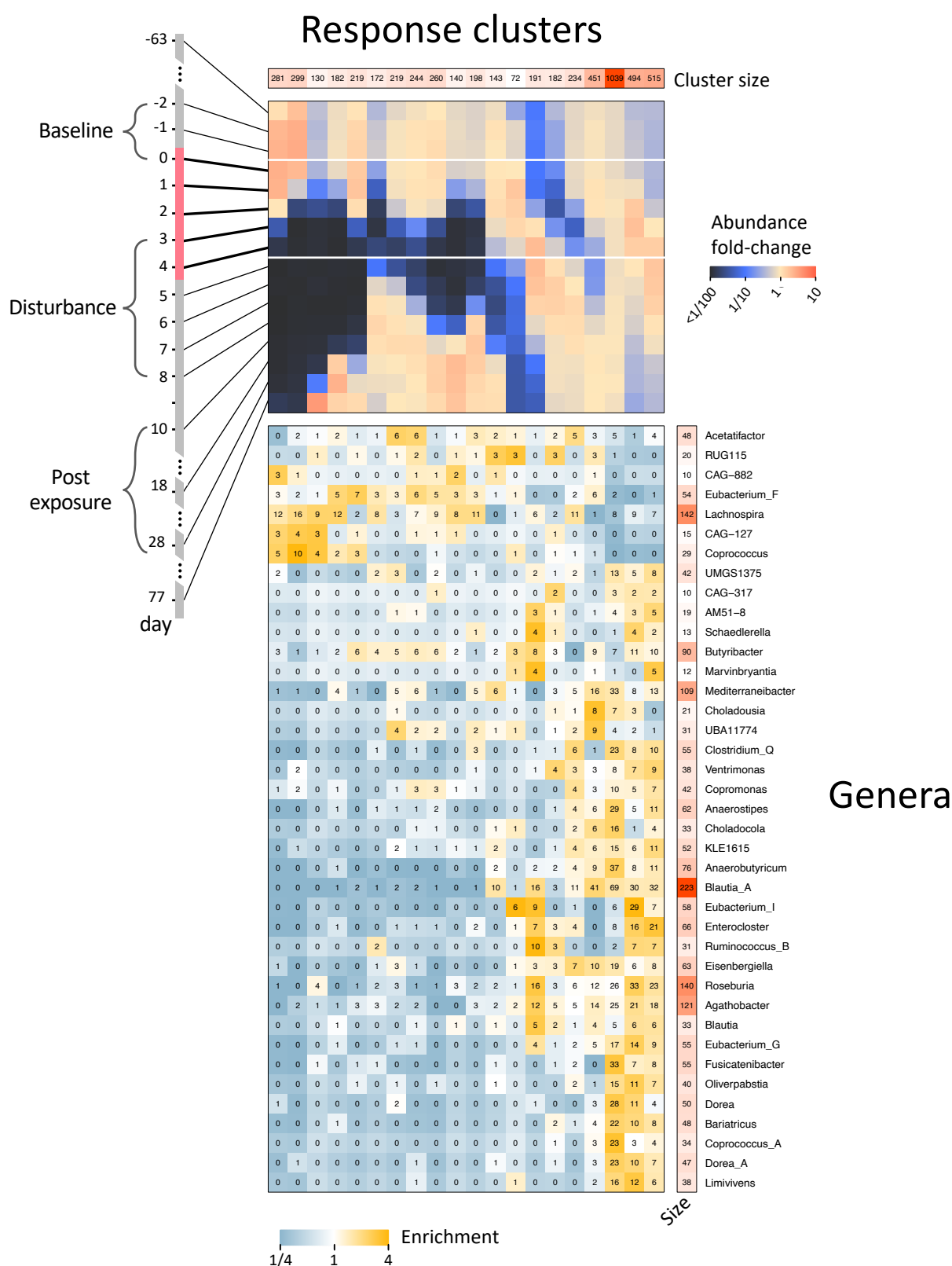
Supplementary Figure 15. Amplification of region within *gyrA* to study sweeps.

Supplementary Note 1. Removal of non-polymorphic variants.

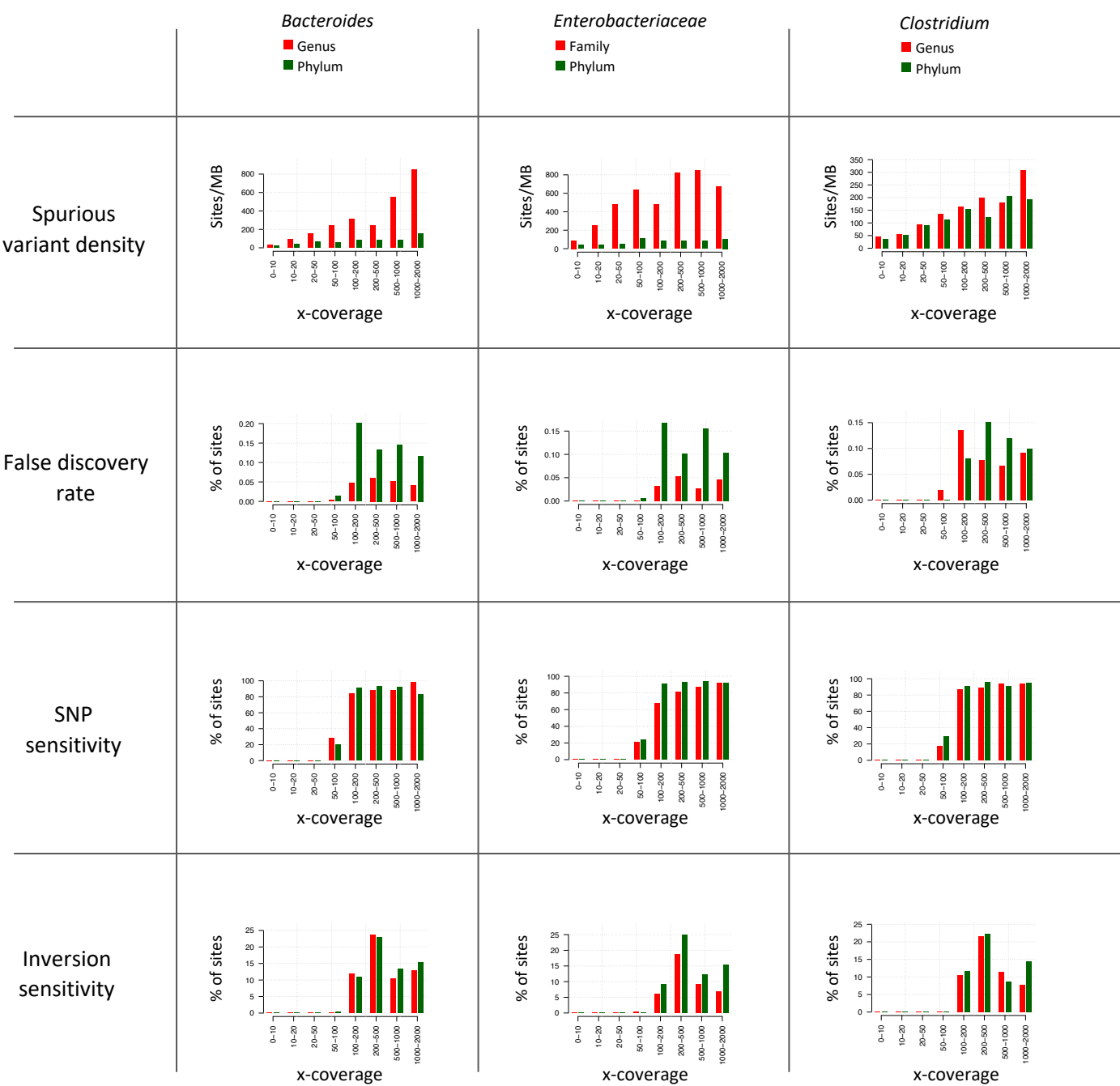
Supplementary Note 2. Calculation of a lower bound on the number of mutant cells in a population with an abundance of 0.01%.



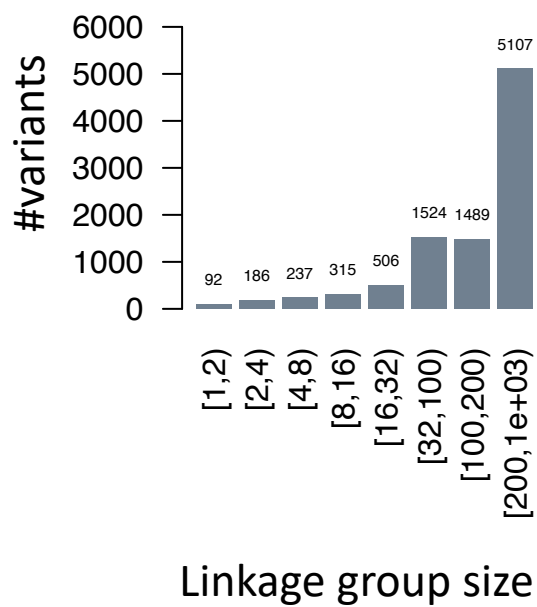
Supplementary Figure 1. Genome quality. Genome quality was assessed according to single-copy genes. An eCDF (empirical cumulative distribution function) is shown for genome completeness (top) and genome contamination (bottom). The completeness threshold (50%) and contamination threshold (10%) used to select the 5665 genomes presented in this study are depicted with black vertical lines.



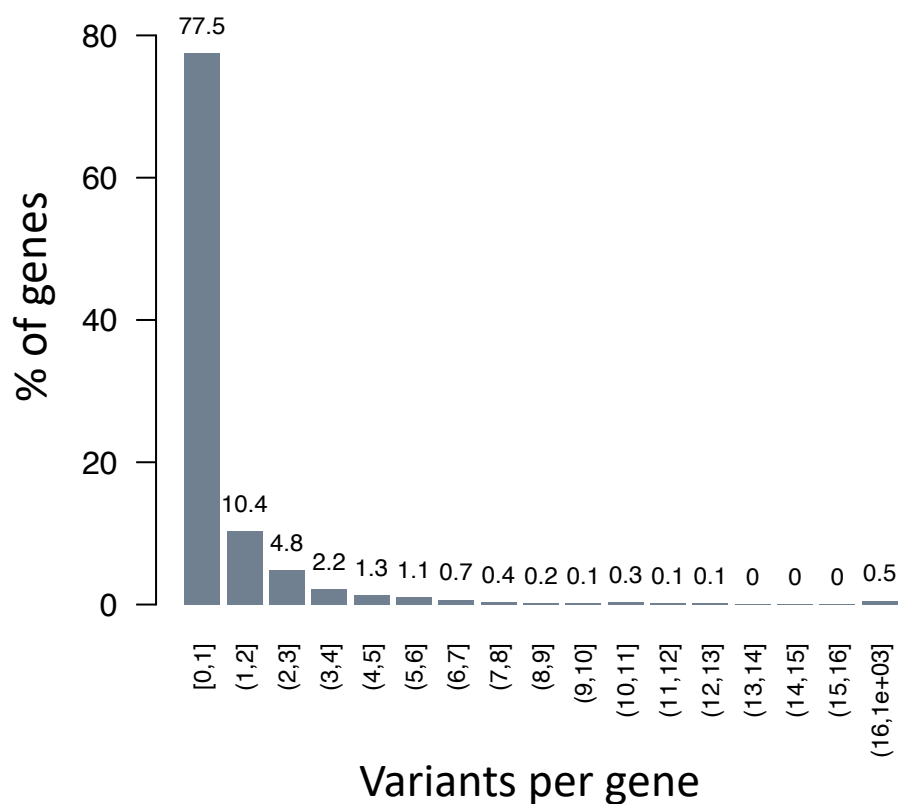
Supplementary Figure 2. Diverse response trends within the *Lachnospiraceae* family. Same as **Extended Data Figure 3**, focusing on genera within the *Lachnospiraceae* family with at least 10 populations. Genera sorted according to hierarchical clustering of genera response enrichment vectors. Genera differ in their response to the disturbance. Note for example, how *Lachnospira* collapses and *Eubacterium_I* blooms.



Supplementary Figure 3. Genome crowding. The presence of closely related genomes exacerbates the density of spurious variants and lowers sensitivity for detecting spurious variants in a taxonomy-dependent manner. We compared 10 communities composed of genomes limited to a genus or family vs. 10 communities sampled from the entire phylum for 3 cases: *Bacteroides* genus vs. the entire Bacteroidota phylum (left), the *Enterobacteriaceae* family vs. the Proteobacteria phylum (middle), and the *Clostridium* genus vs. the Firmicutes phylum (right). Generation of minor strains and simulated reads as for **Extended Data Figure 6**, while introducing 100 mutations per genome. The density of spurious variants was elevated with genetic crowding for *Bacteroides* and *Enterobacteriaceae*, but not for *Clostridium*, a diverse genus. False discovery rate was below 0.15% for cases. Sensitivity varied between clades.

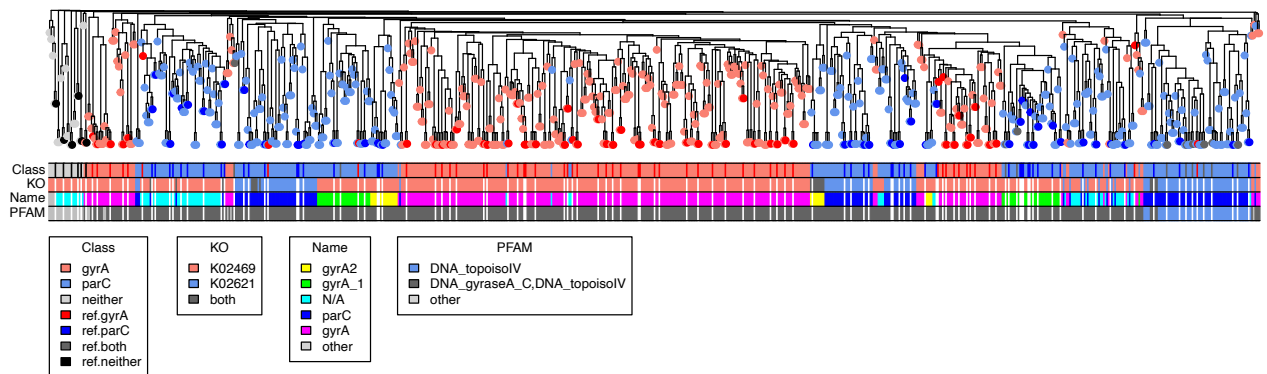


Supplementary Figure 4. Linkage group size. Histogram shows the distribution of sweeping dynamic variants by the size of their linkage group.

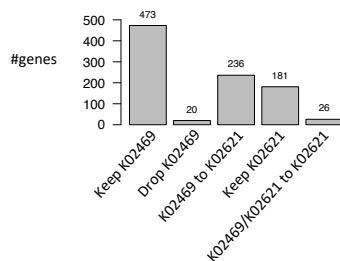


Supplementary Figure 5. Variants per gene. Histogram shows the percentage of 2016 genes that were in or adjacent to 1771 sweeping variants, according to the number of sweeping variants that were in or adjacent to them. These sweeping variants are associated with small linkage groups of no more than 100 variants, so the average density of variants across the genome is low. Nonetheless, 22.5% of genes associated with any variant were associated with more than one variant, indicating that horizontal gene transfer and recombination may have introduced multiple variants to these genes.

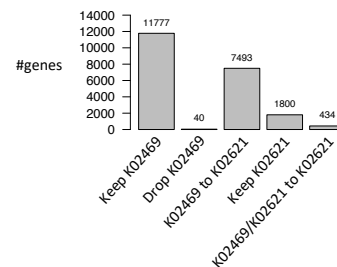
Dendrogram of 939 representative genes and 120 reference genes



Representative gene summary

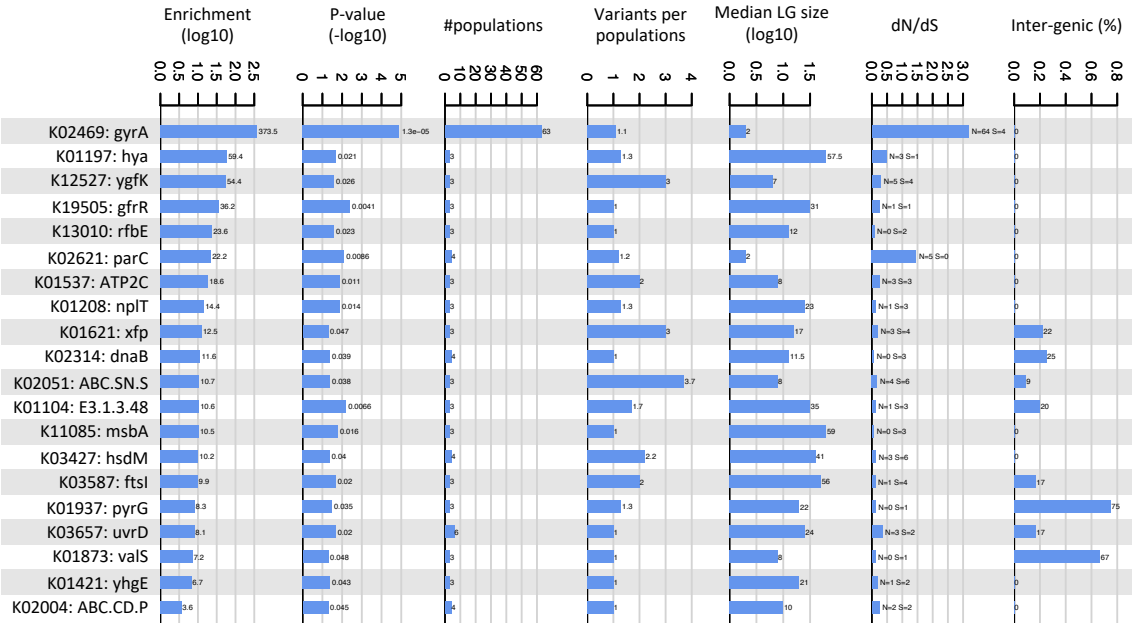


All gene summary



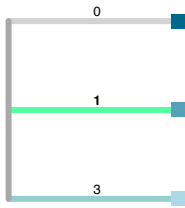
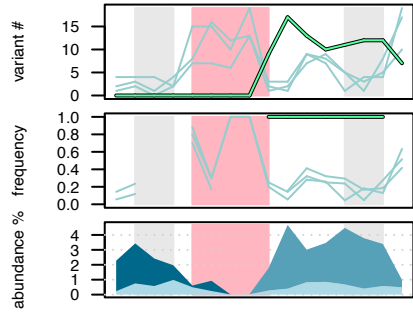
Supplementary Figure 6: Separating the homologs *gyrA* and *parC*. Genes were clustered using mmseqs2, and cluster representatives were annotated using eggNOG. There were 936 topoisomerase IIa associated gene clusters (representing 21544 genes) initially annotated as K02469 (*gyrA*) and/or K02621 (*parC*). Genes were reclassified as K02469 if their eggNOG name was '*gyrA*', the remaining genes were reclassified as K02621 if matching the PFAM entry '*DNA_topoisoIV*', and genes meeting neither criterion were dropped. Topoisomerase IIa genes were analyzed alongside 120 annotated reference genes downloaded from NCBI (Supplementary Table 5) to validate the annotation approach. Shown is a dendrogram of genes >100aa clustered based on amino acid sequence alignment identity and colored according to gene classification (reference genes shown with darker colors). Shown below, from top to bottom, are final classifications, original KO assignments, eggNOG names and the PFAM hits. There is a near-perfect correspondence between the new annotations and annotations of reference genes, validating our reclassification approach. Barplots shown below summarize the number of genes that kept or switched annotations (left: representative genes, right: all genes). After reclassification, there were 11777 genes annotated as *gyrA* with K02469 and 9727 annotated as *parC* with K02621.

KEGG Orthology

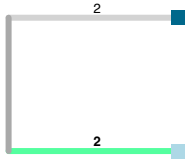
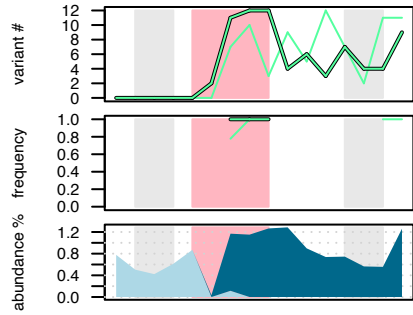


Supplementary Figure 7. KEGG orthology table. All 20 KEGG orthology entries (KOs) significantly enriched in sweeping variants are shown. Columns describe attributes of the variants associated with the KO, from left to right: Observed over expected enrichment ratios (background model generated through shuffling of variants within their respective genomes). Enrichment P-values. Number of populations with one or more KO-associated variants. Average number of KO variants per population. Median size of linkage group. dN/dS ratios of assigned variants, calculated from intragenic substitution variants. Percent of intergenic variants. For visualization purposes dN/dS ratios were calculated using a pseudo count of 1: $dN/dS = [(N+1)/(S+1)] / [(n+1)/(s+1)]$, where N and S are the observed number non-synonymous and synonymous, and n and s are the possible number non-synonymous and synonymous, respectively. All of the data in this figure is provided in Supplementary Table 6.

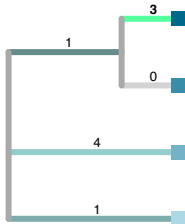
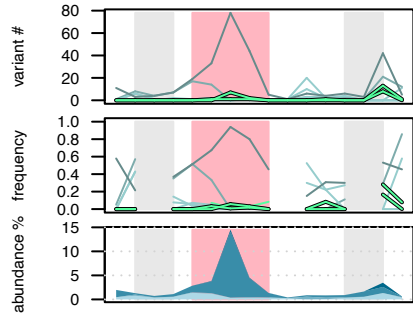
DBL_b19982_1
Lachnospira sp000437735
(Lachnospiraceae)
C=95 X=107.6 n=4



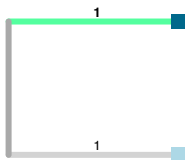
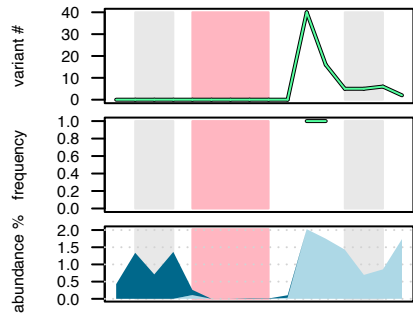
EAB_b29935
Mesosutterella multiformis
(Burkholderiaceae)
C=97 X=83.7 n=2



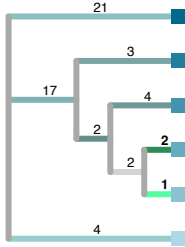
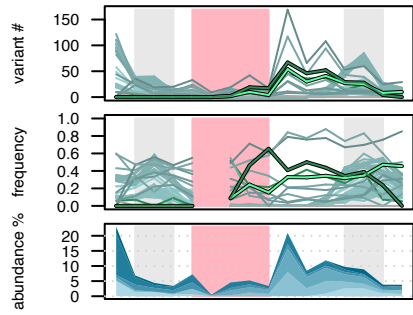
EAD_b21428
Oliverpabstia faecicola
(Lachnospiraceae)
C=94 X=309.7 n=9

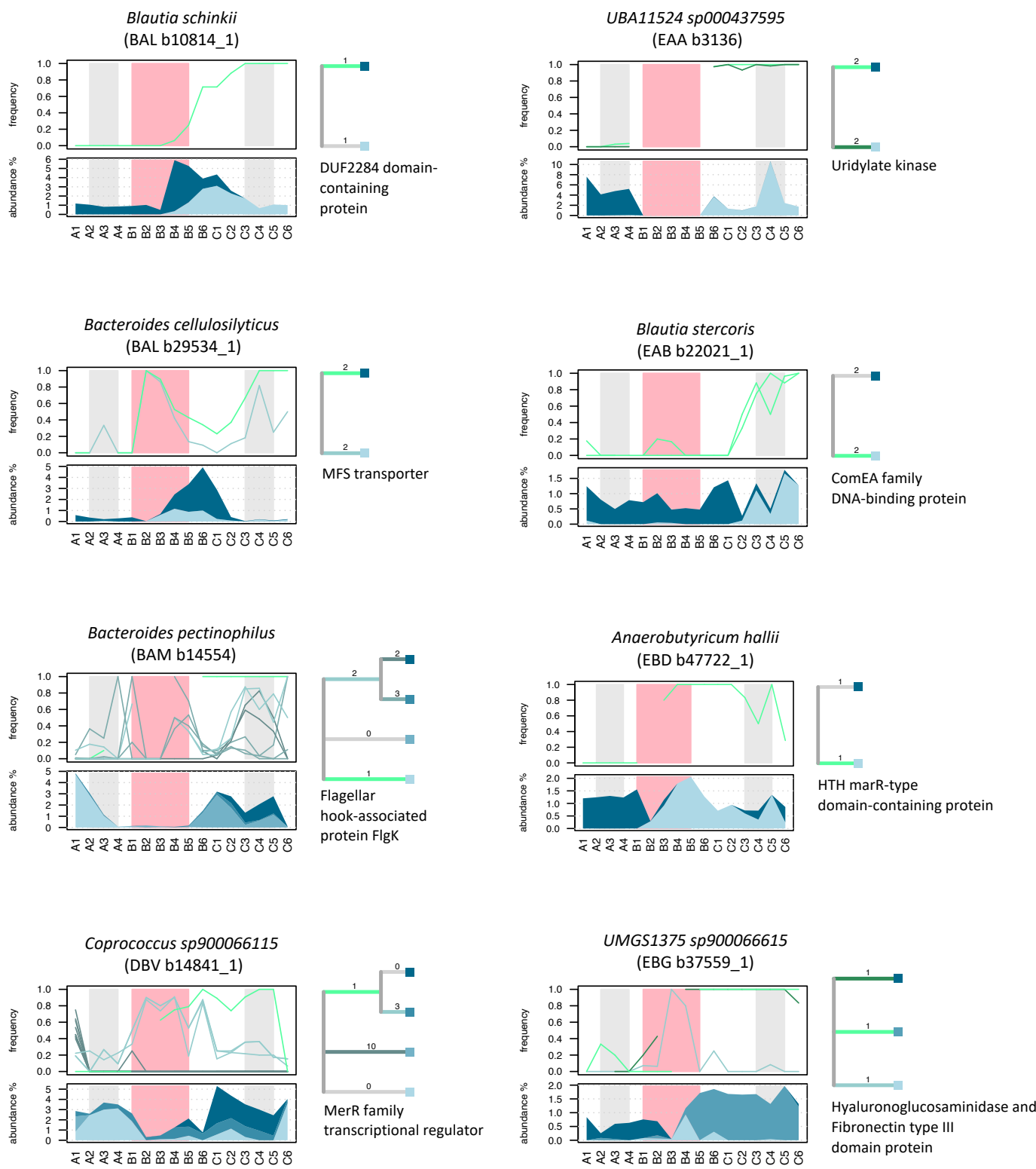


EBQ_b37216
Mesosutterella sp.
(Burkholderiaceae)
C=87 X=118.4 n=1

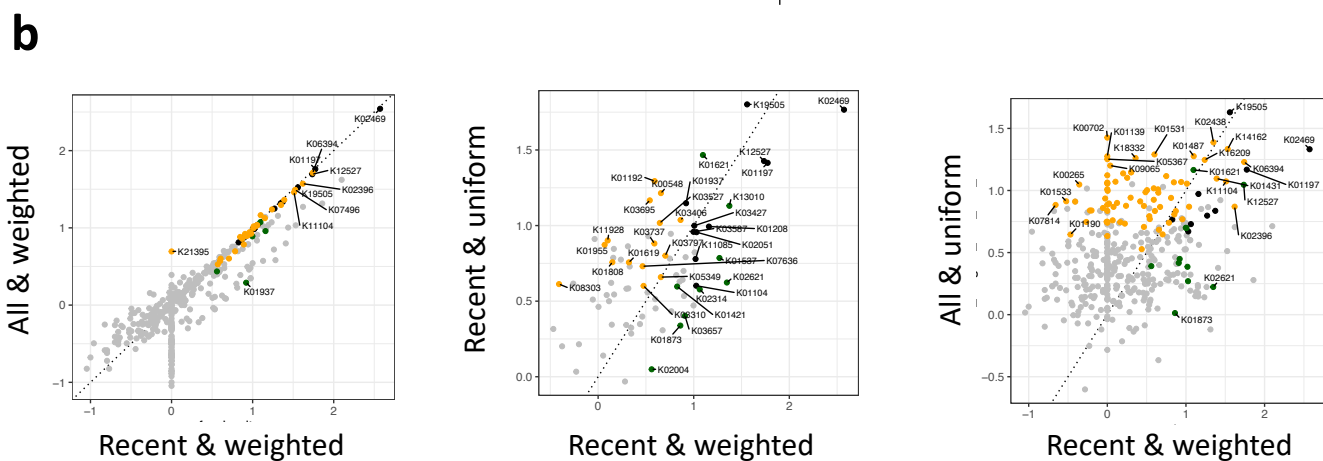
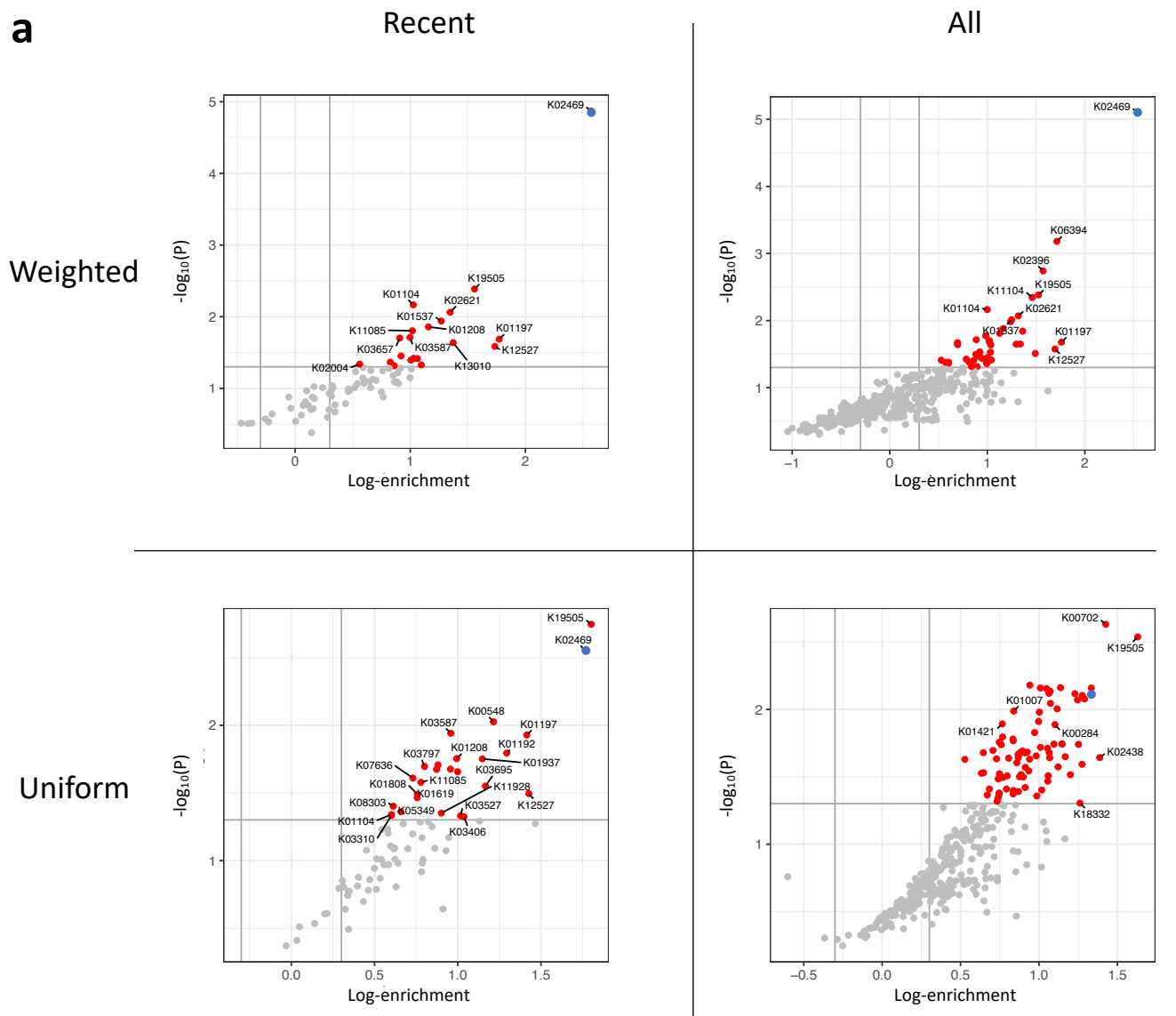


EBU_b3854_1
Prevotella copri
(Bacteroidaceae)
C=91 X=985.6 n=48

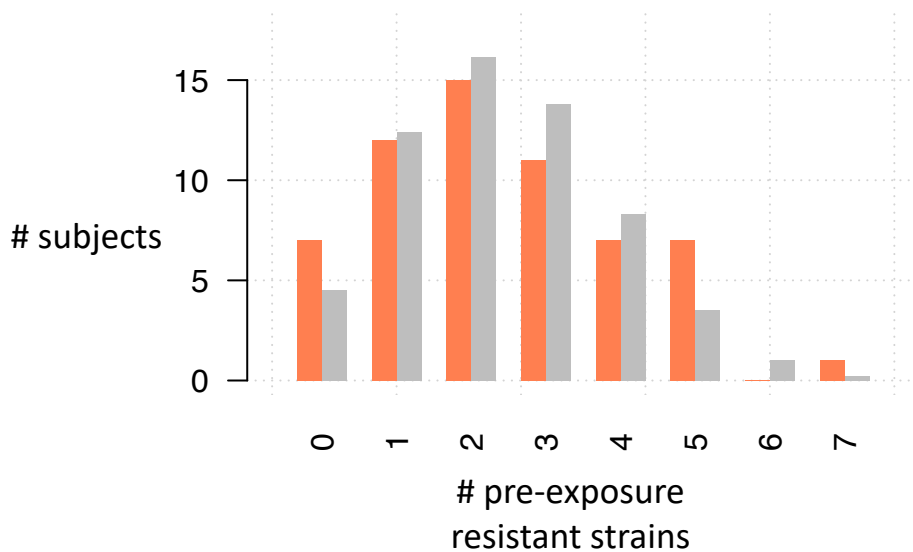




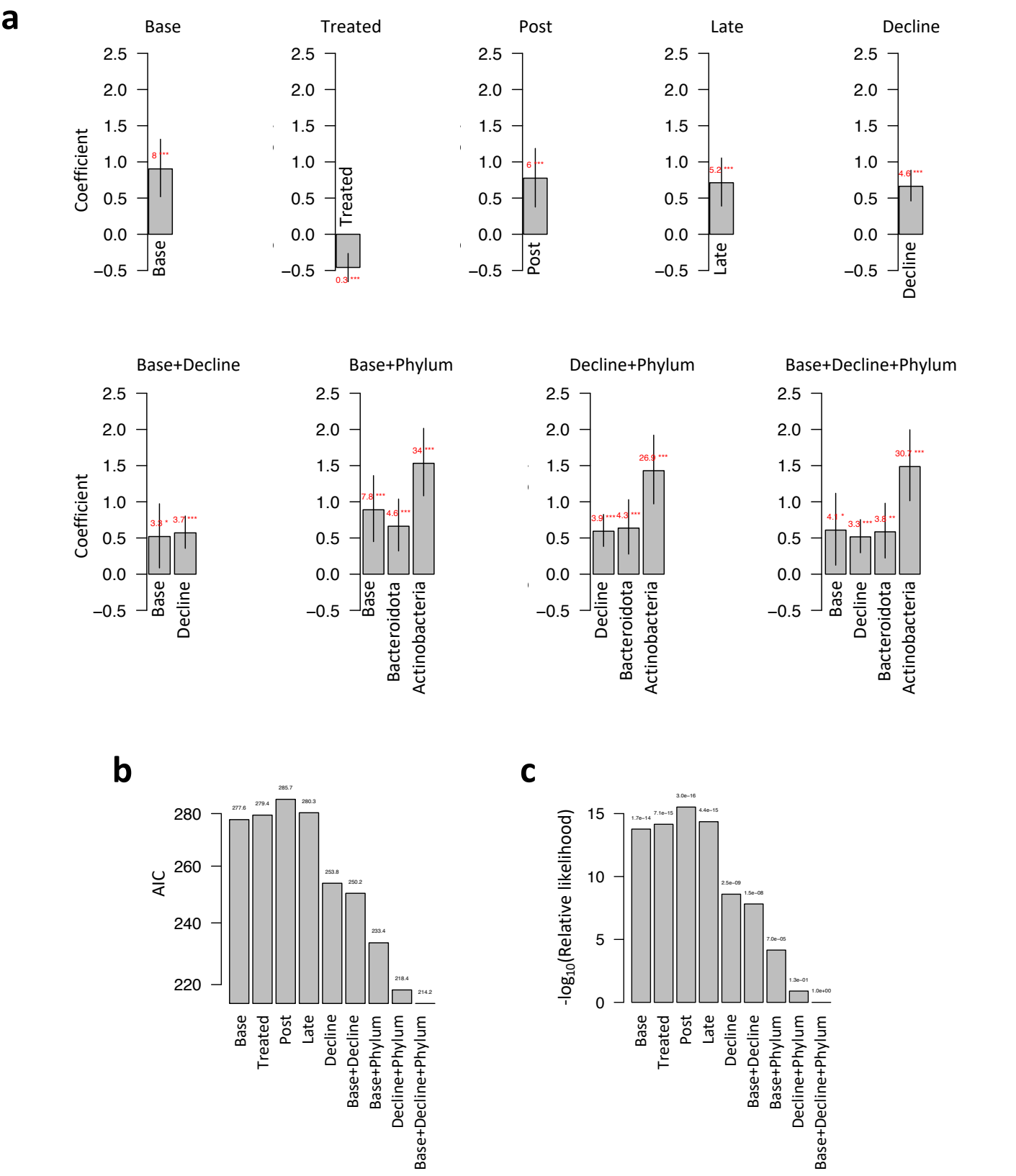
Supplementary Figure 9. Examples of evolving populations not associated with mutations in *gyrA*. There were 58 populations that (1) had serine in position 83 in *gyrA*, (2) no dynamic variants associated with *gyrA*, and (3) had sweeping variant associated with non-synonymous substitutions in genes other than *gyrA*. Shown are 8 representative populations, with subjects and genome identifiers specified in parathesis. See Figure 2h for a description of all figure components. Sweeping variants are colored in green, using different shades if variants are not in same linkage group. Gene annotations are written on the bottom right and were assessed through KO, Uniref100 and PFAMs.



Supplementary Figure 10. Converging evolution controls. **a)** Comparison of the converging analysis performed using “recent” sweeping variants (associated with LGs that were up to 100 in size) vs. all sweeping variants (in genomes with up to 1000 variants), and a weighted scheme (weighting variants by LG size) and a uniform weight scheme. The main analysis shown in the paper was performed using recent and weighted scheme. Significant KOs shown in red, others in gray, and K02469 (*gyrA*) highlighted in blue. **b)** Comparison of log-enrichments values for pairs of runs. KOs found significant in both runs are colored black, KOs significant in the run shown on the y-axis (resp. x-axis) are colored orange (resp. dark green), and the remaining KOs are colored gray. Log-enrichment for KOs with no data in a run are set to 0.

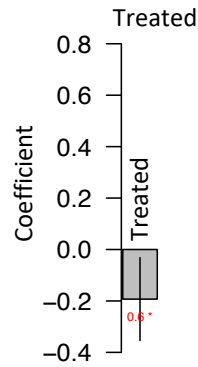


Supplementary Figure 11. Pre-exposure resistant strains across subjects. The number of resistant populations, defined as populations with a species-specific *gyrA*:83 resistant allele prior to the ciprofloxacin exposure, were counted per subject. Shown in orange is the number of subjects (y-axis), as a function of the number of resistant populations (x-axis). Shown in gray is a background model, generating through n=10,000 permutations of the data that randomly shuffled *gyrA*:83 alleles between subjects while respecting species. The Real and permuted distributions of the number of resistant populations per subject were not significantly different, based on an asymptotic two-sample Kolmogorov-Smirnov test.

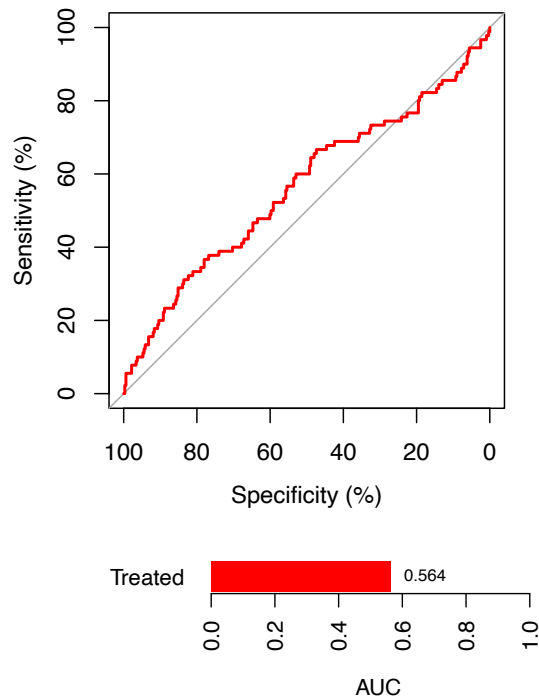


Supplementary Figure 12. *GyrA* evolvability models. Comparison of 9 logistic regression models with *gyrA* evolvability as the response variable. **a)** Coefficient values (y-axis, log₁₀ scale) for all coefficients (x-axis). Exponentiated coefficient values depicted in red text. *p*-value significance represented with asterisks: P<0.001 (***), P<0.01 (**), P<0.05 (*). Confidence intervals (5% to 95%) depicted with vertical lines. **b)** Akaike information criterion (AIC, y-axis), for all models (x-axis). Values depicted in text above bars. **c)** The relative likelihood (y-axis, log-transformed) for all models (x-axis). Values depicted in text above bars.

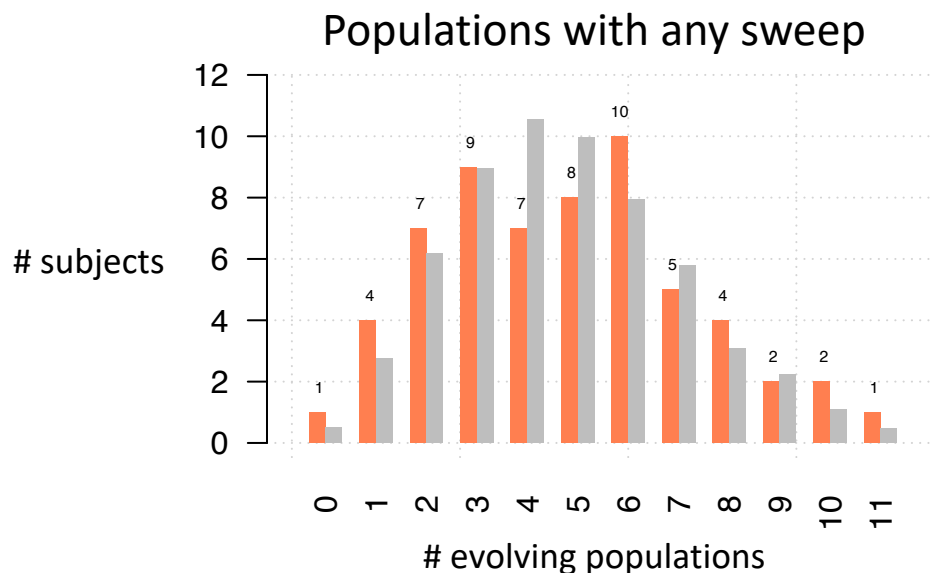
a



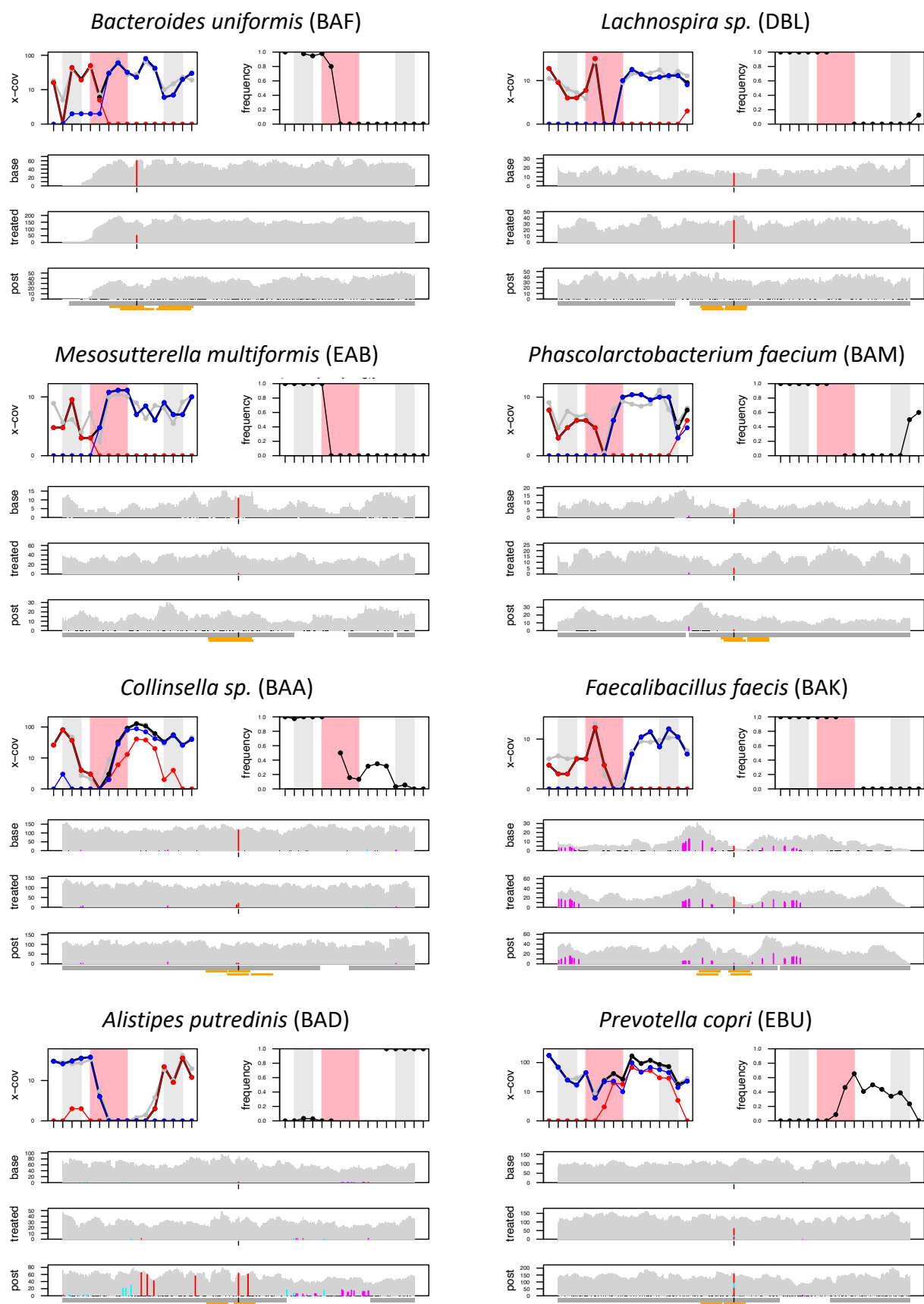
b



Supplementary Figure 13. Non-GyrA evolvability. The 9 models described in **Supplementary Figure 12** were fitted with non-*gyrA* evolvability as the response variable. Shown is the single model (based on the predictor 'Treated') for which the coefficients were significant (p -value below 0.05). **a)** Model coefficients. **b)** ROC curve and AUC value for the model. See **Supplementary Figure 12** for the legends of panel **a** and Figure 4d for the legend of panel **b**.



Supplementary Figure 14. Evolving populations across subjects. The number of evolving populations (involving *gyrA* sweeps on top, any sweep on bottom) were counted per subject. Shown in orange is the number of subjects (y-axis, numbers above bars), as a function of the number of evolving populations (x-axis). Shown in gray is a background model, generating through n=1000 permutations of the data that randomly shuffled populations between subjects while respecting genera. The Real and permuted distributions of the number of evolving populations per subject were not significantly different, based on an asymptotic two-sample Kolmogorov-Smirnov test.



Supplementary Figure 15. Amplification of region within *gyrA* to study sweeps. We focused on 8 cases of sweeping *gyrA* that selected from different species, different subjects, and different substitutions at position 83. Shown is the shotgun coverage in 2 kb region around position 83 for 8 representative cases. Figure legend as in **Extended Data Figure 7**.

Supplementary Note 1

PolyPanner removes two types of non-polymorphic variants. In this note we describe how we leverage temporal data to enhance the detection and removal of these spurious variants and compare our approach to existing approaches.

Paralog variants are a result of two or more similar stretches of DNA, present in the same genome, that align to one genomic site in the metagenomic co-assembly. For example, minor variations in high-copy number genes can result in paralog variants. We identify and remove paralog variants by recognizing their linkage to each other, which results in their coverage vectors failing a Chi-Square test of independence (described in detail in ‘Dynamic variant classification’ in the Methods section). Of note, true polymorphic variants that display little change in frequency over time will also be classified as paralogs. Such variants are not of interest in our analysis, since they are unlikely to contribute to fitness differences within the population under the range of conditions experienced during longitudinal sampling. To the best of our knowledge, previous tools have not attempted to identify and remove paralog variants.

A second type of spurious variants are *ortholog variants*, which are a result of similar stretches of DNA that are present in two or more distinct genomes. In our assembly-based approach these similar sequences can be consolidated into a single assembled sequence that ends up being binned with one of the genomes. That genome is called the recipient genome, while the other genomes are called donors (which may not be recognized as such, and perhaps not assembled at all). Donor reads that align to the recipient genome will result in spurious variants that we call *ortholog variants*. This phenomenon has been called ‘read-stealing’ in previous work (e.g., Garud et al. 2019).

A key property of orthologs is that the recipient variant is linked to the recipient genome, while donor variants are linked to their own genomes. Since they are linked, they are expected to have similar coverage over a single or multiple time points. PolyPanner leverages the temporal data to assess linkage. If a variant coverage vector is linked to the regional coverage vector, it is deemed a recipient variant and the other variants at the site are classified as orthologs and dropped.

Previous approaches addressed the ortholog issue using coverage within a single sample. Within that sample, genes and loci were masked if they had an unusually high or low coverage, relative to the genome median. For a complete description we refer the readers to Section S1A in the Methods and Supplemental Information of Garud et al. 2019. Under that approach,

typical coverage thresholds were chosen to avoid masking loci due to the variation in coverage that arises from differing distances from the origin of replication and other biases. Actual thresholds used were 3-fold above and below the median genome coverage.

This approach, taken by Garud et al., correctly handles some types of ortholog variants. Given an ortholog variant, if the donor abundance is higher than the recipient abundance by the appropriate factor (e.g. >3-fold), the problematic genomic region will be masked and dropped. However, if the donor abundance is lower (e.g. 50% of the recipient genome), the donor and recipient loci would together be considered a low frequency polymorphic site ($f=0.33$) which would not be masked out. Garud et al. addressed this issue by combining multiple tactics, including blacklisting known shared genes and limiting their analysis to sites with extreme changes in allele frequency (≤ 0.2 to ≥ 0.8 , or vice versa). The main novelty of PolyPanner is the use of temporal sequencing data to identify spurious sites in a statistically sound manner, resulting in the precise identification of polymorphic sites that have may have extreme or subtle temporal changes in their frequency.

Supplementary Note 2

Calculation of a lower bound on the number of mutant cells in a population with an abundance of 0.01%. The number of mutant cells for any specific mutation is estimated to equal $\mu \times N \times \frac{(1-e^{-ts})}{s}$, where μ is the mutation rate (mutations per bp per generation), N is the population size, t is the number of generations since the last sweep or colonization event, and s is the selection coefficient representing the fitness cost of the mutant allele under prevailing conditions (i.e., prior to antibiotic exposure). For more details see Eq. 3.9 in ref.⁸³ that deals with the case in which $t = \infty$ and Eq. 7 in ref.⁵⁰. We assume that there are 10^{13} cells in the intestine of a subject, and therefore $N = 10^9$ for a population with an abundance of 0.01%. We also assume $s \leq 0.01$ (in line with results in this work) and $t \geq 70$ (indicating at least one week passed since the last selective sweep). Plugging into the formula, the number of mutant cells m for any given mutation (such as a single nucleotide change in *gyrA*) is at least 5 and 50 for $\mu = 10^{-10}$ and $\mu = 10^{-9}$ respectively.