PNAS
www.pnas.org

Supplementary Information for

Horizontal gene transfer overrides mutation in *Escherichia coli* colonizing the mammalian gut

Nelson Frazão, Ana Sousa, Michael Lässig, Isabel Gordo

Isabel Gordo, Michael Lässig
E-mail: igordo@igc.gulbenkian.pt, mlaessig@uni-koeln.de.

This PDF file includes:

Supplementary Text
Supplementary Figures S1 to S11
References for SI

Other supplementary materials for this manuscript include the following:

Supplementary Tables (S1 to S27) were submitted in a single separated Excel file (.xlsx).

www.pnas.org/cgi/doi/10.1073/pnas.1906958116

Supplementary Text

Invader *Escherichia coli* lineage. Clones expressing fluorescence were derived from *E. coli* K12 strain MG1655 (1, 2). Serial plating of 1X PBS dilutions of feces in LB agar plates supplemented with the appropriate antibiotics were incubated overnight and YFP- or CFP-labeled bacterial numbers were assessed by counting the fluorescent colonies using a fluorescent stereoscope (SteREO Lumar, Carl Zeiss).

Isolation and characterization of the resident *E. coli* lineage. For species identification, we used McConkey + 0.4% lactose medium, *E. coli* phylogenetic group multiplex PCR (3) and Multi-Locus Sequence Typing (MLST) (4). Genetic diversity was analyzed using an ERIC-based typing technique (5). A resident *E. coli* lineage was isolated from the mouse microbiota (*SI Appendix,* Fig. S4*A*) along the evolution experiment. Twelve isolates per mouse per week were confirmed to belong to the *E. coli* species by using a multiplex PCR (3) that amplifies specific genes, from the four *E. coli* phylogenetic groups. All the isolates (*n* = 192) belonged to the phylogenetic group B1 (*SI Appendix,* Fig. S4*B*). Fingerprinting 48 isolates for repetitive sequences (5) showed that a single *E. coli* lineage naturally colonizes the intestinal microbiota of these laboratory mice (*SI Appendix,* Fig. S4*C* and Table S4).

Furthermore, upon whole-genome sequencing and MLST of a resident *E. coli* clone isolated at day -2 from the mouse G2 intestine (*SI Appendix,* Table S1) we further confirmed that the resident lineage belongs to the *E. coli* species. The sequences of seven housekeeping genes (*adk*, *fumC*, *gyrB*, *icd*, *mdh*, *purA*and *recA*) were extracted from the resident's genome and compared with *E. coli* sequences deposited in the *E. coli* MLST database (http://mlst.warwick.ac.uk/mlst/dbs/Ecoli), thus retrieving the Sequence Type of the resident *E. coli* clone, identified as ST602 (*SI Appendix,* Table S2).

Antibiotics supplemented culture media. Streptomycin (100 µg/mL), tetracycline (30 µg/ mL), ampicillin (100 µg/mL), chloramphenicol (30 µg/mL), nalidixic acid (40 µg/mL), nitrofurantoin (640 µg/mL), rifampicin (100 µg/mL), fosfomycin (32 µg/mL), ciprofloxacin (1 µg/mL) or gentamicin (10 µg/mL) were used to supplement the media when specified.

Emergence of streptomycin resistance in the resident *E. coli* lineage. In the animals where coexistence of resident and invader lineages occurred (mice G2 and H2), we tested whether *de novo* evolution towards streptomycin resistance occurred in the resident lineage (the invader YFP clone already carried a resistance mutation). All resident isolates investigated for phylogenetic group classification (n = 192) and tested for growth in streptomycin-supplemented plates (100 µg/mL) were susceptible two days prior to the 24h streptomycin perturbation (*SI Appendix,* Fig. S5). All the resident *E. coli* clones from mouse H2 acquired resistance after antibiotic treatment (*SI Appendix,* Fig. S5), being consistently found at a high load (>10^6 CFU/g of feces) (Fig. 1*B*). On the contrary in mouse G2, the clones of the resident lineage maintained a susceptibility profile (*SI Appendix,* Fig. S5). This lineage suffered a strong bottleneck, leading to undetectable levels at day 2, but then recovered to high loads at day 8 (3×10^8 CFU/g of feces) (Fig. 2*A*). Thus in 2/5 mice the resident could be stably maintained and in 1/2 where it survived and evolved resistance after a single day of streptomycin.

***Escherichia coli* niche size is independent of microbiota**. The mice in the evolution experiment have different microbiota states as assayed by 16s rRNA sequencing (*SI Appendix,* Fig. S3). A linear mixed effects model analysis of the temporal loads along this experiment indicates that, as a species, *E. coli* is maintained at an average load of 10^8 (±1.9) CFU/g feces (*SI Appendix,* Fig. S6 and *SI Appendix,* Table S3), largely independent of microbiota compositional state (*SI Appendix,* Fig. S3).

Muller plots of phage-mediated horizontal gene transfer (HGT) and adaptive mutation dynamics. The Muller plot in Fig. 4*A* represents phage-mediated HGT and adaptive mutation dynamics during the evolution experiment in mouse G2 (*SI Appendix,* Table S20). Phage-mediated HGT events were assessed at days 2, 13 and 20 by PCR (*SI Appendix,* Table S21), and at days 8 and 27 via whole-genome population sequencing (*SI Appendix,* Tables S22 and S23). Adaptive mutation dynamics was investigated in the *frlR* gene, which at day 27 presented one of the most common and parallel (adaptive) mutations (*SI Appendix,* Table S5), via amplicon sequencing for days 2, 9, 13 and 20 (*SI Appendix,* Table S16) and whole-genome population sequencing for days 8 and 27 (*SI Appendix,* Tables S24 and S5). HGT events and *frlR* mutations were assumed to occur mostly within the most common genetic background.

To assess phage (Nef and/or KingRac) mediated HGT dynamics using whole-genome population sequencing, we used the Breseq pipeline to compare each population at days 8 or 27 with the evolved invader YFP clone (*SI Appendix,* Table S22 and S23). This reference genome carries the Nef phage integrated between positions 1,127,918 and 1,174,087 and the KingRac prophage, replacing the Rac defective prophage, inserted between positions 2,497,344 and 2,544,177. We searched for unassigned new junction evidence between the above-mentioned positions to assess the percentage of the population lacking each of the prophages. To distinguish between KingRac or Rac defective integration, we used the average Single Nucleotide Polymorphism (SNP) difference along the respective insertion region.

At day 8 (*SI Appendix,* Table S22), evidence of a new junction was identified for the Nef prophage region at 5.5% in frequency, indicating Nef integration in the majority — 94.5% (100% - 5.5%) — of the population. By contrast, no new junction evidence was found in the KingRac or Rac defective region (2,495,832-2,544,044), indicating that the vast majority of the population carries either of the prophages (*SI Appendix,* Table S22). Analysis of the genetic changes within this region, which were present at 100% in the population (*SI Appendix,* Table S22), and blasting of the corresponding reads containing SNPs against the ancestral genome (NC_000913.2) led to the conclusion that the entire population carries the Rac defective, not the KingRac, prophage. In short, concerning the presence of Nef and/or KingRac at day 8, 5.5% of the *E. coli* population carries no prophage and 94.5% carries the Nef prophage only. At day 27 (*SI Appendix,* Table S23), evidence of a new junction was identified for the Nef prophage region at 3.3% frequency, indicating Nef integration in 96.7% of the population. Regarding the KingRac or Rac defective region, new junction evidence was observed at 31% frequency, indicating that this sub-population carries neither of these prophages (*SI Appendix,* Table S23). To distinguish between KingRac or Rac defective integration in the remaining population, we assessed the presence of genetic changes within this region, which were found in 11.6% (± 0.5% 2s.e.) of the population, with SNP analysis indicating integration of the Rac defective prophage. As the Nef prophage is present in the majority of the population (96.7%), we estimated that 30% (31% of 96.7%) harbors Nef but no KingRac or Rac defective prophages, while 11% (11.6% of 96.7%) carries the Nef + Rac defective prophages and 55.7% (96.7% - 30% - 11%) the Nef + KingRac prophages. In summary, concerning the presence of Nef and/or KingRac, at day 27, 3.3% of the *E. coli* population carries no prophage, 41% (30% + 11%) carries the Nef prophage only, while 55.7% carries the Nef + KingRac prophages.

As for *frlR* mutations (*SI Appendix,* Table S5), we assumed that these occurred in the most common background (55.7%), which carried both the Nef and KingRac prophages (*SI Appendix,* Table S20). As an example, we estimated that 21.3% (38.3% of 55.7%) of the population harboring Nef + KingRac should also have *frlR* mutation 1. In fact, we have randomly isolated a clone from this population and it turn out to carry both phages and only this particular mutation (*SI Appendix,* Table S6).

The Muller plot in Fig. 4*B* represents the combined information of the phage-mediated HGT and *psuK*/*fruA* adaptive mutation dynamics in mouse H2 during the evolution experiment (*SI Appendix,* Table S25). The same rational described above (for mouse G2) was used to estimate HGT dynamics (*SI Appendix,* Tables S21, S26 and S27) in mouse H2. Concerning the adaptive mutation dynamics we used whole-genome population sequencing data of the *psuK*/*fruA* intergenic region mutation (*SI Appendix,* Tables S24 and S3), which was observed to be both common and occurring in parallel at day 27 (*SI*

Appendix, Table S5). At day 8 (*SI Appendix,* Table S26) 5.6% of the *E. coli* population carries no prophage, 60.7% (34% + 26.7%) carries the Nef prophage only, while 33.7% carries the Nef + KingRac prophages. As for *psuK*/*fruA* mutation (*SI Appendix,* Table S24), we assumed that these occurred in the most common background (60.7%), which carried Nef prophage only. As an example, we estimated that 16.5% (27.2% of 60.7%) of the population should harbor Nef + *psuK*/*fruA* mutation.

At day 27 (*SI Appendix,* Table S27) 0% of the *E. coli* population carries no prophage, 44% (24.6% + 19.4%) carries the Nef prophage only, while 56% carries the Nef + KingRac prophages. As for *psuK*/*fruA* mutation, we observe that 100% of the population carries this mutation (*SI Appendix,* Table S5).

Microbiota analysis. Raw reads were processed using QIIME version 1.9.1 (6). Quality filtering included a minimum limit of Q20 and a maximum of three low quality bases before read truncation. Ambiguously called bases were not allowed and reads were discarded if trimmed over 75% of original length. Chimera removal was conducted with the QIIME-usearch61 method, which performs both *de novo* and reference-based detection. *Operational taxonomic unit* (OTU) clustering was performed using Uclust with an open-reference approach (7). OTU tables were subsampled without replacement in order to even sample sizes for diversity analysis. The size of the smallest sample was chosen for subsampling, in this case 11432 reads. Unifrac distance was used as beta diversity metric to compare community structure. Taxonomic assignment of OTUs was based on GreenGenes taxonomy (8). Unifrac distance matrices and OTU tables were used to calculate principal coordinates and construct ordination plots using R software package version 3.4.3 (http://www.R-project.org). Richness, as the observed number of OTUs/sample, and Shannon index (9) were calculated for alpha diversity analysis.

Whole-genome sequencing of *Escherichia coli* populations and clones. Illumina technology: Each sample was pair-end sequenced in an Illumina MiSeq Benchtop Sequencer and standard procedures produced data sets of paired-end 250 bp read pairs. The mean coverage per sample was 282x, 350x, 194x, 201x and 238x for populations A2, B2, I2, G2 and H2, respectively. For the evolved invader-YFP and resident clones, the mean coverage was 73x and 67x, respectively. Mutations were identified using the BRESEQ pipeline (v0.26) (10), with the polymorphism option on (for populations) or off (for clones), using as reference the *E. coli* genome MG1655 (NC_000913.2). For populations, the default settings were used except for: a) requirement of a minimum coverage of three reads on each strand per polymorphism; b) eliminating polymorphism predictions occurring in homopolymers of length greater than 3, except when frequency ≥ 5%; c) discarding polymorphism predictions with significant ($P < 0.05$) strand or base quality score bias. Parallel mutations were defined as mutational events that occurred in a minimum of two animals. For additional verification of mutations predicted by BRESEQ, we used the IGV software (version 2.3.93) (11). *Nanopore technology:* Third-generation sequencing (Oxford Nanopore) technology was used to obtain a fully-closed bacterial genome sequence. Libraries were prepared without shearing to maximize sequencing read length using the ligation sequencing kit (SQK-LSK108). Sequencing was performed in MinION Mk1b (Oxford Nanopore Technologies, Oxford, UK) using SpotON flow cell (R9.4) in a 48-h sequencing protocol in MinKNOW (v1.1.8 or 2.0.1). Reads were base-called with Albacore (v1.1.2 or v2.0.2) to output fastq. All reads that passed Albacore quality control thresholds were used subsequently. Adapter sequences were then trimmed (in the case of the evolved invader clone) from the reads using Porechop (v0.2.2). The hybrid read set (both Illumina and Nanopore reads) were assembled using Unicycler (v0.4.0 or 0.4.3). Briefly, Unicycler produces an Illumina reads based assembly graph with SPAdes (v3.10.1 or 3.11) that are then assembled with long reads. Unicycler polishes its final assembly with Illumina reads using Pilon (v1.22) to reduce the rate of small base-level errors and producing complete circular and closed assemblies (*i.e.* circular contigs).

Supplementary Figures

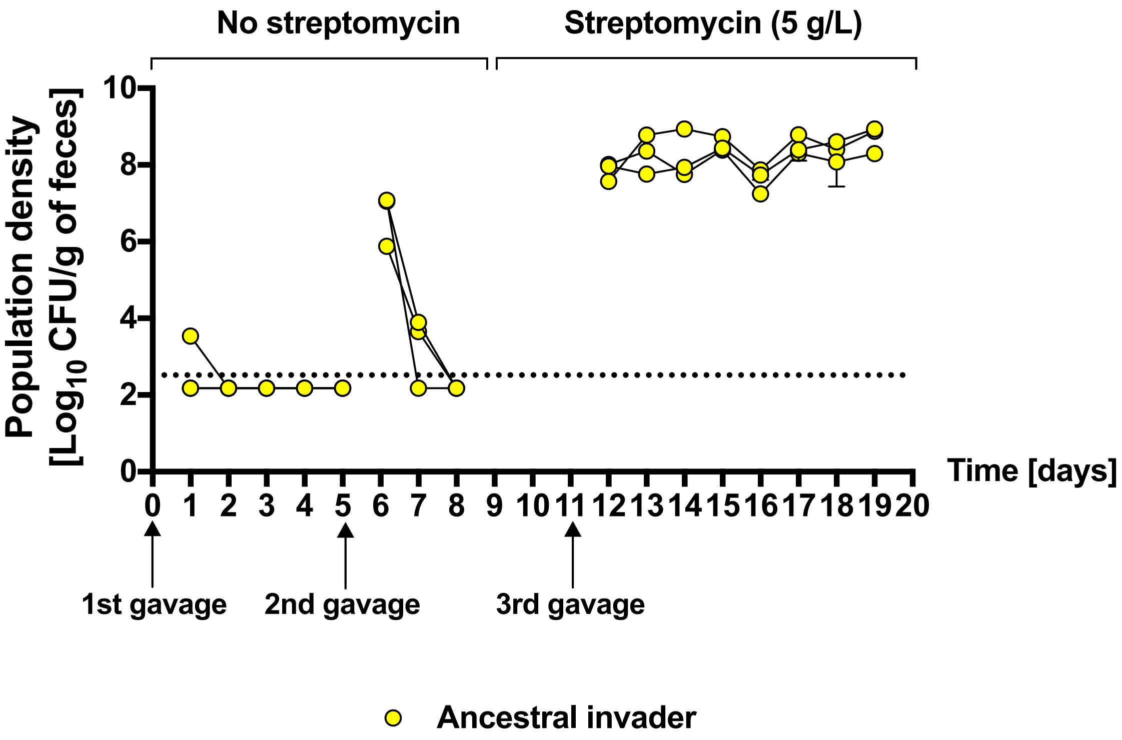

Fig. S1. The invader *E. coli* is unable to colonize the gut of mice not treated with streptomycin. Gut colonization attempts with ancestral invader *E. coli* clone without prior streptomycin treatment (1st and 2nd gavage) or continuous streptomycin (5 g/L) treatment of 3 mice. The dotted line indicates the detection limit (330 CFU/g of feces).

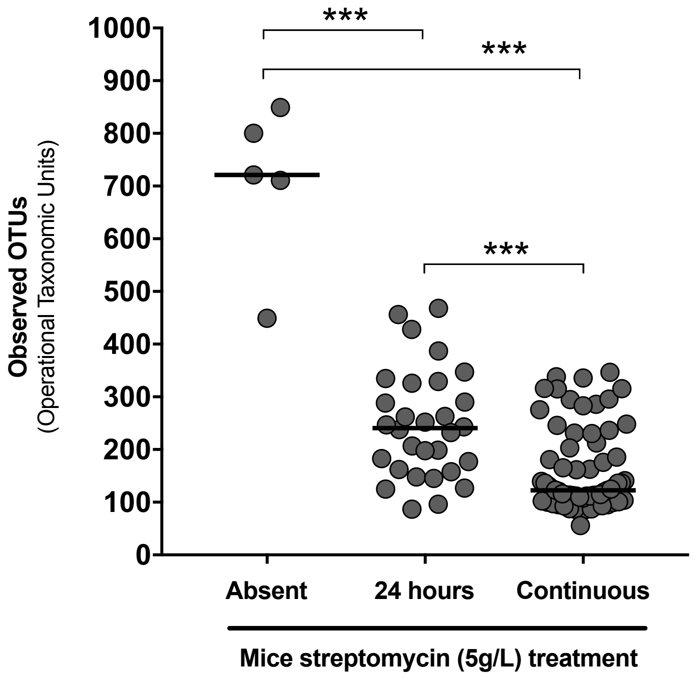

Fig. S2. A short streptomycin treatment (24h) leads to higher microbiota diversity than a continuous treatment.

Microbiota alpha diversity in the absence, after a short perturbation (24h) or during continuous streptomycin treatment of different mice (gray circles). The untreated and 24-h treated mice correspond to the 5 animals analyzed in the present study at day -2 and days 2, 8, 9, 13, 20 and 27, respectively. The continuous treatment refers to 21 mice (days 1, 2, 3, 4, 7, 12, 17 and 18) analyzed in a previous publication from our laboratory (12) with the same mouse strain (SPF C57BL/6J animals). Alpha diversity comparison based on the number of observed OTUs (Mann Whitney test; *** P < 0.0001). The horizontal bar indicates the median number of observed OTUs.

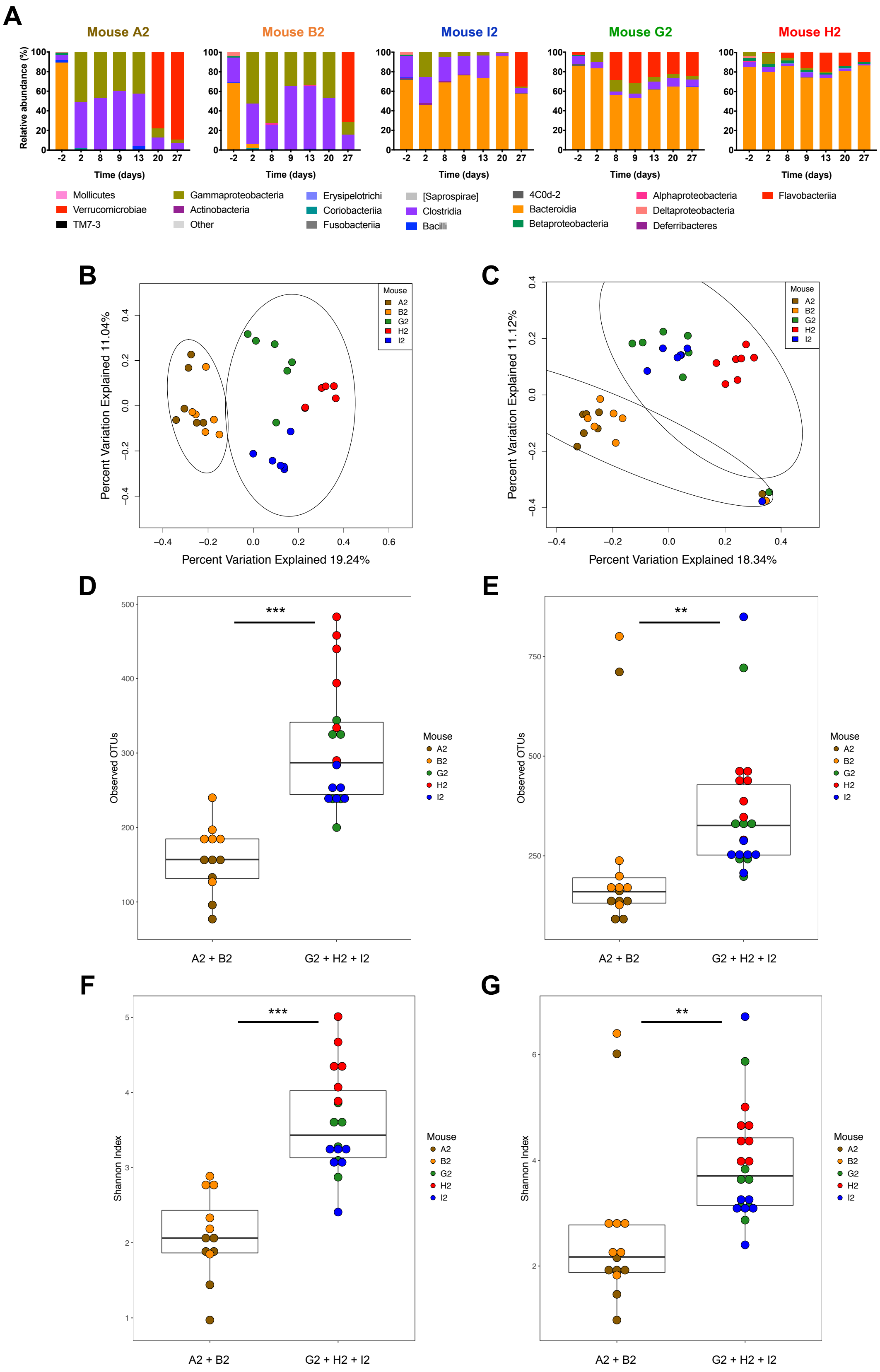
A
Mouse A2
Mouse B2
Mouse I2
Mouse G2
Mouse H2
Relative abundance (%)
100
80
60
40
20
0
-2
2
8
9
13
20
27
Time (days)
Mollicutes
Verrucomicrobiae
TM7-3
Gammaproteobacteria
Actinobacteria
Other
Erysipelotrichi
Coriobacteriia
Fusobacteriia
[Saprospirae]
Clostridia
Bacilli
4C0d-2
Bacteroidia
Betaproteobacteria
Alphaproteobacteria
Deltaproteobacteria
Deferribacteres
Flavobacteriia
B
Percent Variation Explained 11.04%
Percent Variation Explained 19.24%
Mouse
A2
B2
G2
H2
I2
−0.4
−0.2
0.0
0.2
0.4
0.6
C
Percent Variation Explained 11.12%
Percent Variation Explained 18.34%
D
Observed OTUs
***
A2 + B2
G2 + H2 + I2
100
200
300
400
500
E
**
250
500
750
F
Shannon Index
***
1
2
3
4
5
G
**
2
4
6

Fig. S3. Gut microbiota analyses during evolution experiment reveal heterogeneity between mice. (A) Microbiota composition as relative OTU abundance assayed by 16S rRNA amplicon sequencing and clustered at the class level (colored segments) of each colonized mice during the evolution experiment (see Fig. 1*B*). The first bar (day -2) represents the microbiota composition before colonization. **(B** and **C)** Microbiota beta diversity visualization by principal coordinate analysis (PCoA) based on Unweighted UniFrac distances excluding (B) or including (C) samples before antibiotic treatment and colonization with the invader *E. coli* (day -2). Ellipses represent the standard deviation of point scores with a 95% confidence limit for each group (ANOSIM test, $P<0.05$). Microbiota alpha diversity comparisons based on the number of observed OTUs **(D** and **E)** or on Shannon index **(F** and **G)** excluding **(D** and **F)** or including **(E** and **G)** day -2 (Mann Whitney test; **$P<0.001$, *** $P<0.0001$).

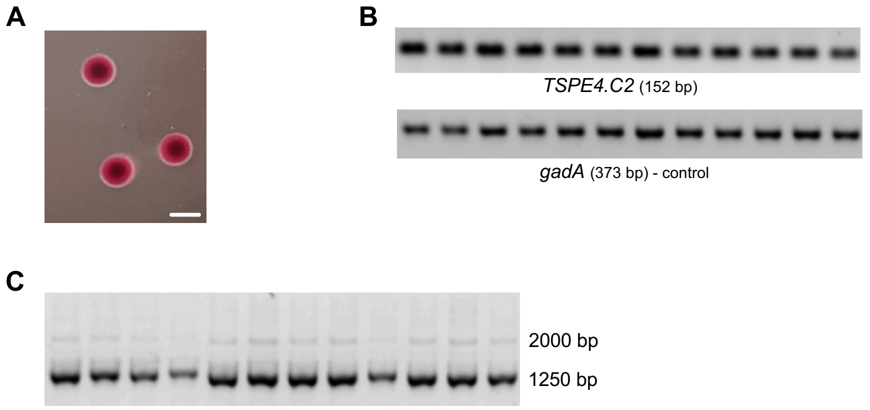

Fig. S4. Isolation and characterization of resident *E. coli* clones. (A) Resident *E. coli* **l**actose-positive colonies grown in Enterobacteriacea selective medium (McConkey + 0,4% lactose). Lactose-fermentation leads to the formation of pink colonies due to the production of acid, which changes the neutral red pH indicator from colorless to red. Bar = 3 mm. **(B)** Clones isolated from mouse fecal material amplified the the internal control gene *gadA* and the *TSPE4.C2* gene indicating they belong to the same phylogenetic group B1 (3). **(C)** Clones isolated from mouse fecal material exhibited the same ERIC-based pattern (5), suggesting no genetic variability.

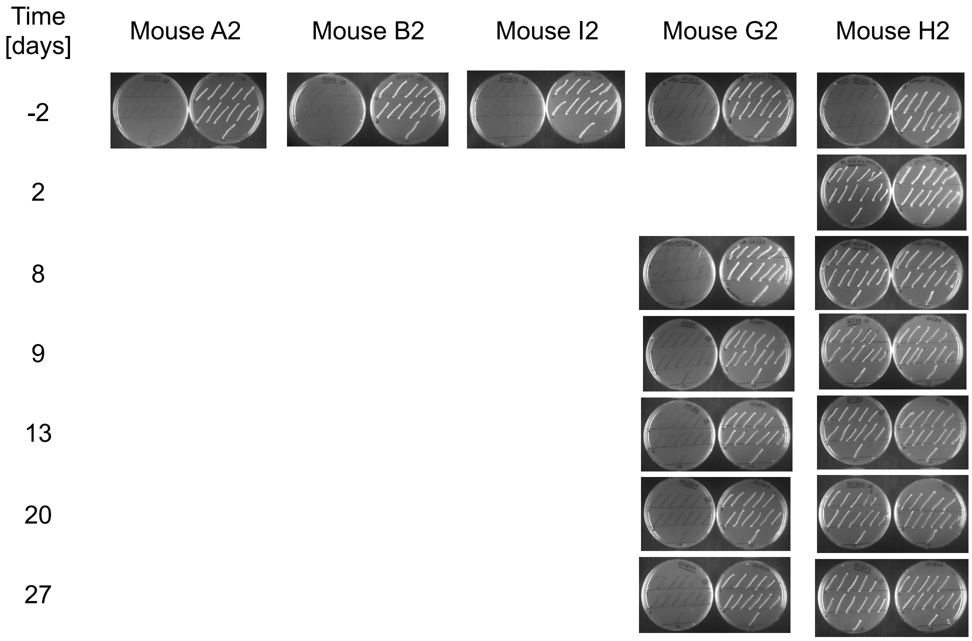

Fig. S5. Streptomycin susceptibility of resident *E. coli* isolates. Isolates were streaked on LB agar plates supplemented or not with streptomycin (100 µg/mL), to determine antibiotic susceptibility of the 192 resident *E. coli* clones belonging to phylogenetic group B1. In each image, the streptomycin-supplemented plate is on the left and the non-supplemented on the right. Blank spaces correspond to days when isolation of resident *E. coli* clones was not possible (absent or below the detection limit of 330 CFU/g of feces).

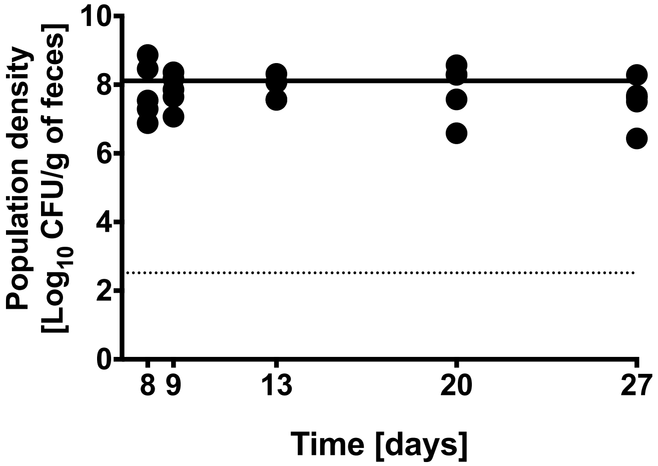

Fig. S6. Size of the *E. coli* ecological niche. Quantification of the total *E. coli* (invader + resident) load in the gut of mice A2, B2, I2, G2 and H2. The dotted line indicates the detection limit (330 CFU/g of feces).

A - Ancestral invader *E. coli* genome (MG1655 - NC_000913.2)

E - Evolved invader *E. coli* genome from clone isolated from mouse G2 at day 27 (Table S1)

R - Resident *E. coli* genome from clone isolated from mouse G2 at day -2 (Table S1)

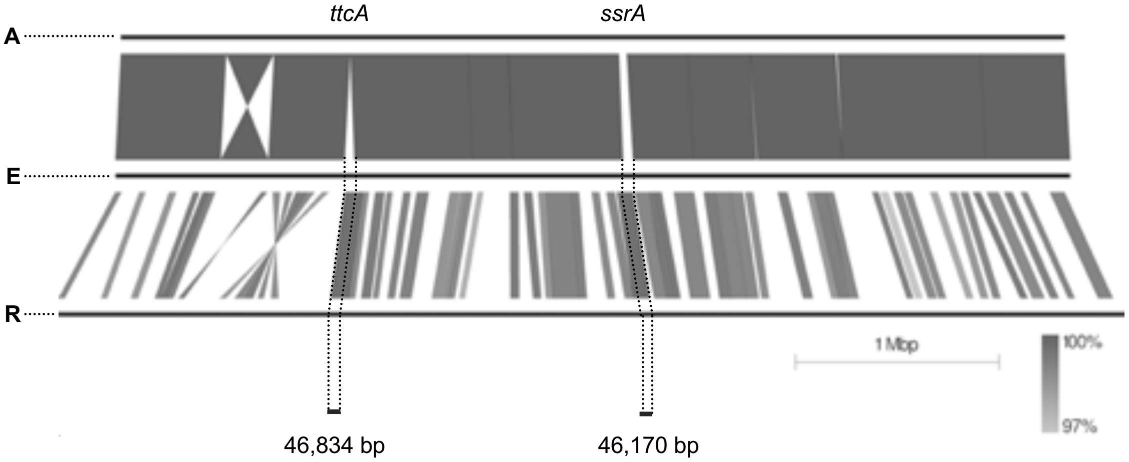

Fig. S7. Comparison of ancestral, evolved and resident genomes. Two genomic regions were found to be absent from the ancestral genome and common to the evolved and resident genomes. These regions, 46,834 and 46,170 bp in length, are inserted in the evolved invader genome at bacterial genes *ttcA* and *ssrA*.

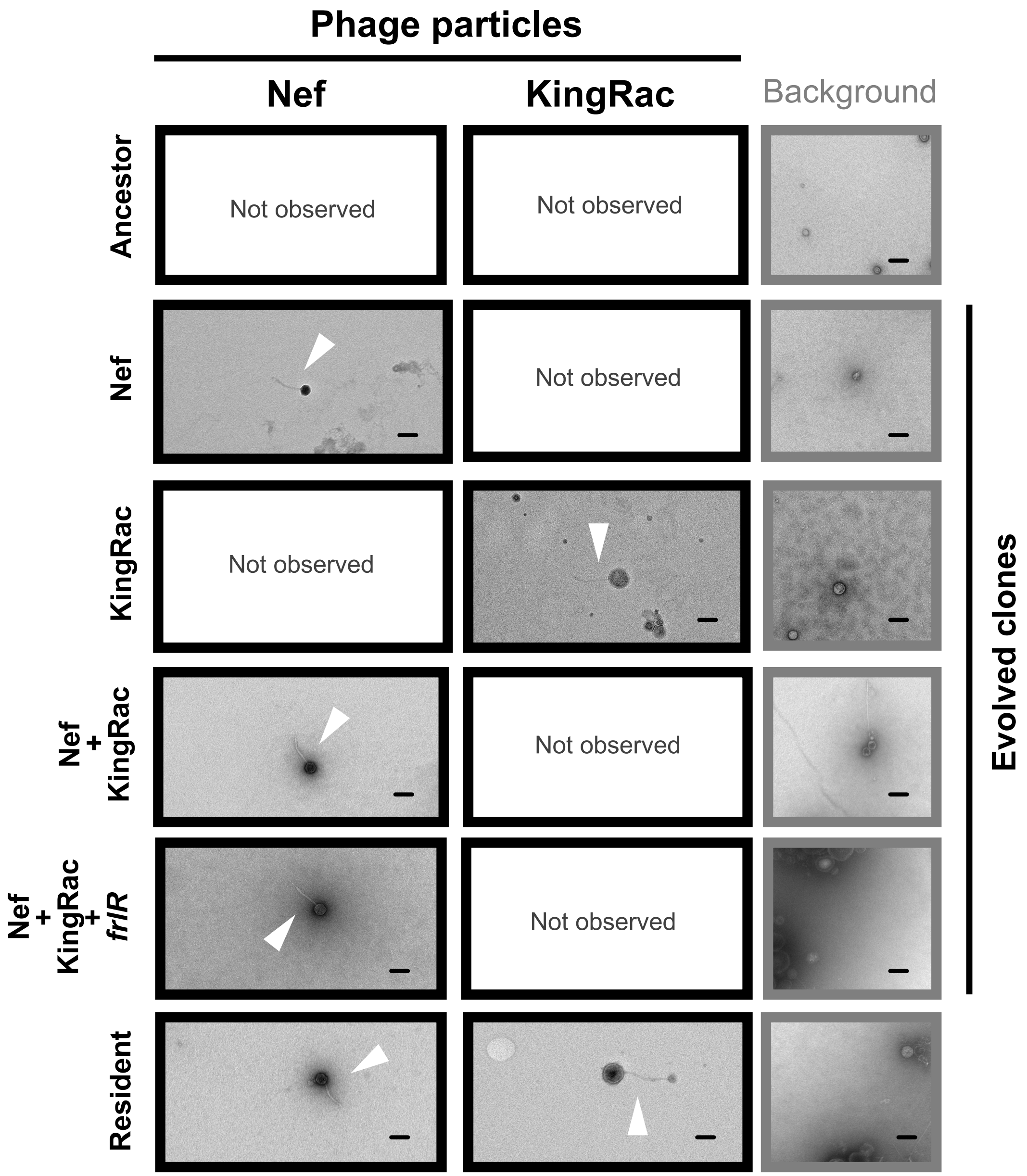

Fig. S8. Transmission electron microscopy analysis of the phage particles produced by the *E. coli* clones. Micrographs of the phage lysate suspensions obtained after mitomycin C induction of bacterial cultures of the ancestral, evolved (Nef, KingRac, Nef+KingRac and Nef+KingRac+*frlR*) and resident *E. coli* clones. The white arrows indicate the observed phage particles. Bars = 100 nm. Direct magnification: 20 000 x.

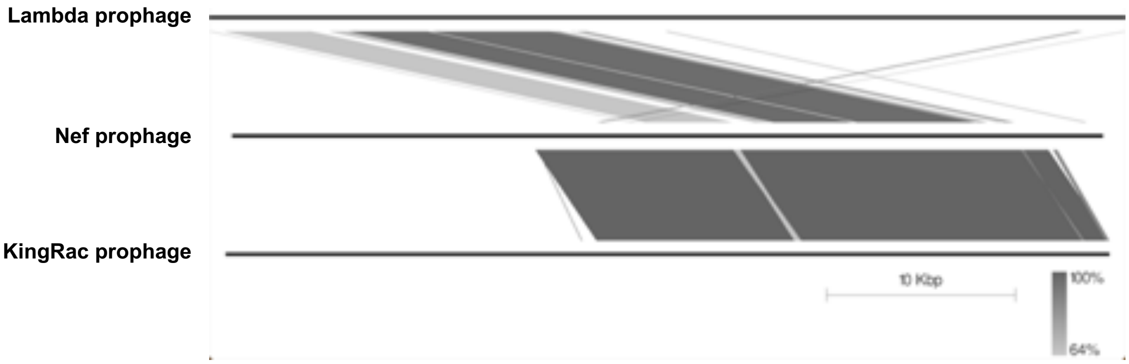

Fig. S9. Comparison between the Lambda, Nef and KingRac prophage sequences. Extensive sequence similarity between Lambda (J02459.1), Nef and KingRac prophage sequences.

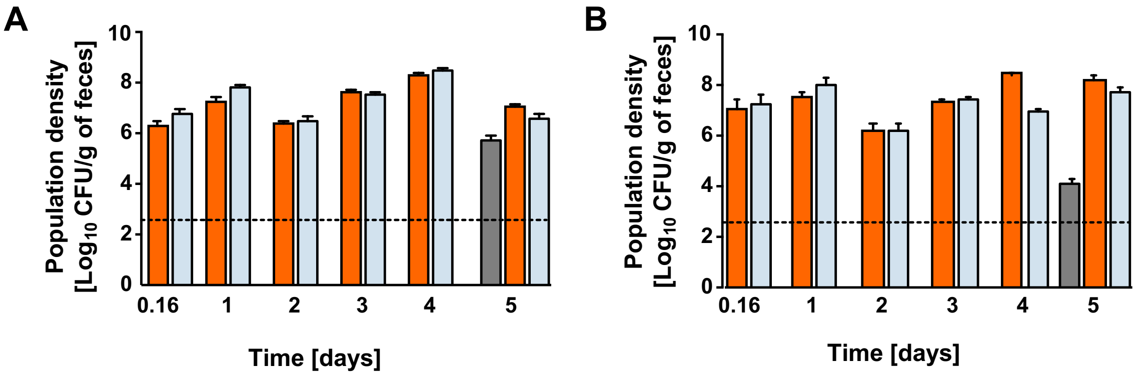

Fig. S10. Coexistence of the evolved, ancestral and resident *E. coli* lineages in the mouse gut. Loads of the evolved invader (phage donor - orange bars), ancestral invader (phage recipient - blue bars) and resident (phage donor - gray bars) *E. coli* populations colonizing two mice (**A** and **B**, also represented in Fig. 4*C* and *D*, respectively) during the co-colonization experiment. Error bars represent 2X standard error (SE), and the dotted line indicates the detection limit (330 CFU/g of feces).

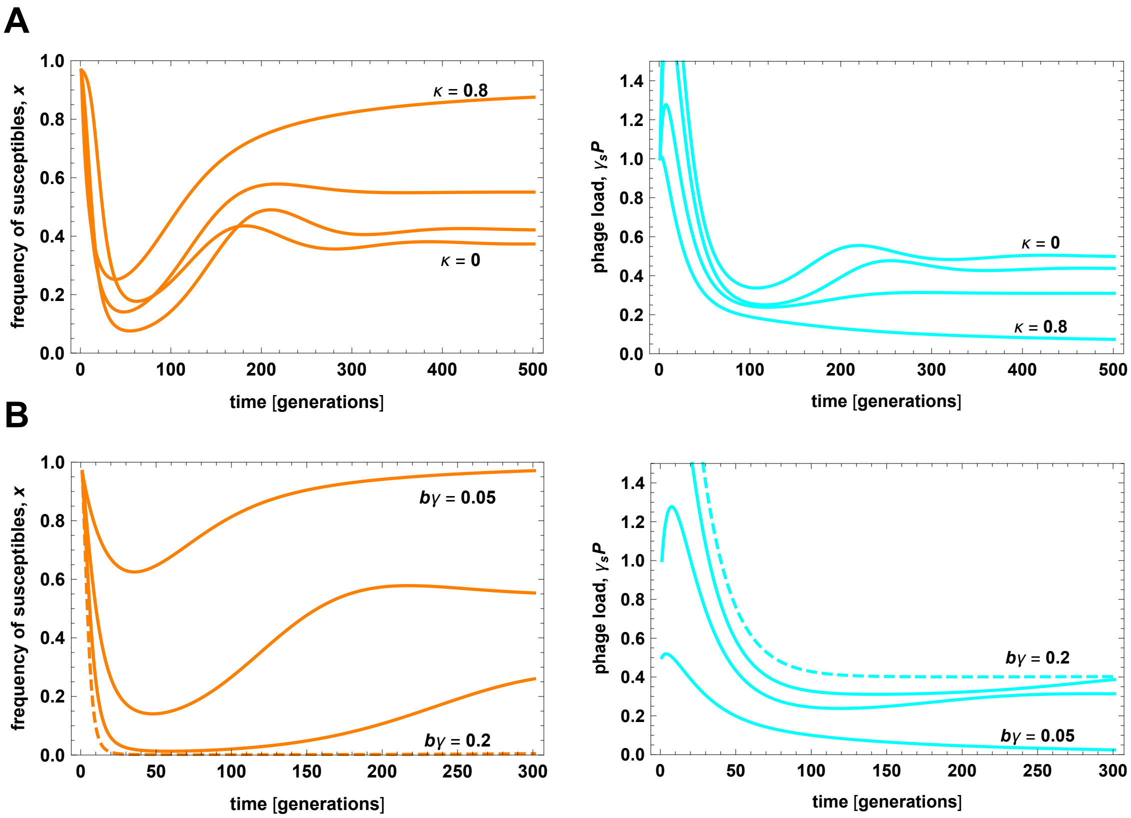

Fig. S11. Parameter dependence of phage epidemics. ***(A)*** The population frequency of susceptible bacteria in the invader population (orange) and the phage load (cyan) are plotted against time for different values of the lysogenization fraction of susceptible bacteria, κ = 0, 0.2, 0.5, 0.8. Pattern and characteristic time scale of the epidemic depend only weakly on lysogenization. ***(B)*** The same data are plotted for different values of the scaled infection cost $\boldsymbol{\tilde{\gamma} = b\gamma}$ = 0.05, 0.1, 0.15, 0.2; see Materials and Methods for the underlying scaling properties of the model. The epidemic pattern shows similar dynamical regimes as in Fig. 5*A*: initial decline of susceptibles at high phage levels and subsequent rebound at lower phage levels (low and intermediate $\boldsymbol{\tilde{\gamma}}$, solid), pandemic with rapid loss of susceptibles (high $\boldsymbol{\tilde{\gamma}}$, dashed). Other model parameters are as in Fig. 5*A*: scaled infection cost $\boldsymbol{\tilde{\gamma}} =$ 0.1 (in *A*), induction cost $\boldsymbol{\delta_R}$ = 0.01, lysogenization fraction κ = 0.5 (in *B*),

background fitness $r_S = 0.15$, $r_I = 0.11$, carrying capacity $\boldsymbol{c^{-1} = 0.1}$, niche overlap $\boldsymbol{q = 1}$, phage clearance rate $\boldsymbol{\lambda} = 0.05$.

References

1. Barroso-Batista J, et al. (2014) The First Steps of Adaptation of Escherichia coli to the Gut Are Dominated by Soft Sweeps. *PLoS Genet* 10(3):e1004182.

2. Mason TG, Richardson G (1982) Observations on the in vivo and in vitro competition between strains of Escherichia coli isolated from the human gut. *J Appl Bacteriol* 53(1):19–27.

3. Doumith M, Day MJ, Hope R, Wain J, Woodford N (2012) Improved multiplex PCR strategy for rapid assignment of the four major Escherichia coli phylogenetic groups. *J Clin Microbiol* 50(9):3108–3110.

4. Maiden MC, et al. (1998) Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc Natl Acad Sci U S A* 95(6):3140–3145.

5. Versalovic J, Koeuth T, Lupski JR (1991) Distribution of repetitive DNA sequences in eubacteria and application to fingerprinting of bacterial genomes. *Nucleic Acids Res* 19(24):6823–6831.

6. Caporaso JG, et al. (2010) QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* 7(5):335–336.

7. Rideout JR, et al. (2014) Subsampled open-reference clustering creates consistent, comprehensive OTU definitions and scales to billions of sequences. *PeerJ* 2:e545.

8. McDonald D, et al. (2012) An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME J* 6(3):610–618.

9. Shannon CE (1948) A Mathematical Theory of Communication. *Bell Syst Tech J* 27(3):379–423.

10. Barrick JE, et al. (2009) Genome evolution and adaptation in a long-term experiment with Escherichia coli. *Nature* 461(7268):1243–1247.

11. Robinson JT, et al. (2011) Integrative genomics viewer. *Nat Biotechnol* 29(1):24–26.

12. Sousa A, et al. (2017) Recurrent Reverse Evolution Maintains Polymorphism after Strong Bottlenecks in Commensal Gut Bacteria. *Mol Biol Evol* 34(11):2879–2892.