1   **Adaptive evolution of hybrid bacteria by horizontal gene transfer**

2

3   Jeffrey J. Power[1,♦], Fernanda Pinheiro[1,♦], Simone Pompei[1,♦], Viera Kovacova[1], Melih

4   Yüksel[1], Isabel Rathmann[1], Mona Förster[1], Michael Lässig[1]*, Berenike Maier[1]*

5   [1] University of Cologne, Institute for Biological Physics, Köln, Germany

6   ♦ Joint first authors.

7   * Corresponding authors. Email: mlaessig@uni-koeln.de, berenike.maier@uni-koeln.de.

8

9   **Abstract**

10   **Horizontal gene transfer is an important factor in bacterial evolution that can act across**

11   **species boundaries. Yet, we know little about rate and genomic targets of cross-lineage**

12   **gene transfer, and about its effects on the recipient organism's physiology and fitness.**

13   **Here, we address these questions in a parallel evolution experiment with two *Bacillus***

14   ***subtilis* lineages of 7% sequence divergence. We observe rapid evolution of hybrid**

15   **organisms: gene transfer swaps ~12% of the core genome in just 200 generations, and**

16   **60% of core genes are replaced in at least one population. By genomics, transcriptomics,**

17   **fitness assays, and statistical modeling, we show that transfer generates adaptive evolution**

18   **and functional alterations in hybrids. Specifically, our experiments reveal a strong,**

19   **repeatable fitness increase of evolved populations in the stationary growth phase. By**

20   **genomic analysis of the transfer statistics across replicate populations, we infer that**

21   **selection on HGT has a broad genetic basis: 40% of the observed transfers are adaptive.**

22   **At the level of functional gene networks, we find signatures of negative and positive**

23   **selection, consistent with hybrid incompatibilities and adaptive evolution of network**

24   **functions. Our results suggest that gene transfer navigates a complex cross-lineage fitness**

25   **landscape, bridging epistatic barriers along multiple high-fitness paths.**

26

27   **Significance statement**

28   In a parallel evolution experiment, we probe lateral gene transfer between two *Bacillus subtilis*

29   lineages close to the species boundary. We show that laboratory evolution by horizontal gene

30   transfer can rapidly generate hybrid organisms with broad genomic and functional alterations.

31   By combining genomics, transcriptomics, fitness assays and statistical modeling, we map the

32   selective effects underlying gene transfer. We show that transfer takes place under genome-

33   wide positive and negative selection, generating a net fitness increase in hybrids. The

34   evolutionary dynamics efficiently navigates this fitness landscape, finding viable paths with

35   increasing fraction of transferred genes.

36

37   **Introduction**

38   Horizontal gene transfer (HGT) plays an important role in bacterial evolution (1, 2), which

39   includes speeding up the adaptation to new ecological niches (3, 4) and mitigating the genetic

40   load of clonal reproduction (5, 6). On macro-evolutionary time scales, HGT occurs between

41 bacteria of different species and also between bacteria and eukaryotes (1, 2, 7); these dynamics
42 have very heterogeneous rates (8) and the permanent integration of transferred genes into
43 regulatory networks is slow (9). An important mechanism of HGT is transformation, the active
44 import and inheritable integration of DNA from the environment (10, 11). In this process,
45 extracellular DNA binds to the surface of a recipient cell, is transported into the cytoplasm by
46 the recipient's uptake machinery, and then integrated into its genome by homologous
47 recombination.

48 The rate of transformation depends on multiple physiological and selective factors (12, 13). The
49 efficiency of the DNA uptake machinery is a major determinant of the probability of
50 transformation (11). At the level of recombination, this probability decreases exponentially as
51 a function of the local sequence divergence, likely because nucleotide mismatches suppress
52 sequence pairing at the initiation of the recombination step (14-16). As shown in a recent study,
53 laboratory experiments can induce genome-wide transformation between close lineages (17).
54 This study also describes the inhibition of gene uptake by restriction-modification systems, but
55 fitness effects of gene transfer are not addressed. Other evolution experiments give evidence of
56 complex selective effects of HGT (18-20). On the one hand, HGT can impose fitness costs that
57 increase with genetic distance between donor and recipient (21). Observed costs include codon
58 usage mismatch, reduction in RNA and protein stability, or mismatch of regulation or
59 enzymatic activity in the recipient organism (22, 23), and can be mitigated by subsequent
60 compensatory mutations (22, 24) or gene duplications (25). On the other hand, transferred
61 genes can confer adaptive value, for example, resistance against antibiotics or shift of carbon
62 sources (26, 27). These experiments elicit HGT of a few genes with a specific function, which
63 are a confined genomic target of selection. However, we know little about positive and negative
64 fitness effects of HGT on a genome-wide scale. In particular, it remains to be shown to what
65 extent selection can foster genome-wide HGT across lineages, and where selective barriers
66 significantly constrain recombination.

67 These questions are the subject of the present paper. We map genome-wide horizontal transfer
68 between two model lineages (28) of *Bacillus subtilis,* a species with high genetic and
69 phenotypic diversity (29). The two lineages are subspecies with an average sequence
70 divergence of 6.8% in their core genomes, which is close to the species boundary: it is larger
71 than typical diversities within populations and smaller than cross-species distances. We find
72 pervasive HGT on a genome-wide scale, which covers hundreds of genes within about 200
73 generations and confers a repeatable net fitness increase to evolved hybrid organisms. Using
74 Bayesian analysis, we jointly infer the effects of sequence diversity and selection on the
75 observed transfer pattern. We find evidence for genome-wide positive and negative selection
76 with significant hot- and cold spots in functional gene networks. We discuss the emerging
77 picture of horizontal gene transfer navigating a complex cross-lineage fitness landscape.

78

79 **Results**

80 **Parallel evolution experiments.** Our experiments record HGT from the donor species
81 *Bacillus subtilis subsp. spizizenii W23* to the recipient *Bacillus subtilis subsp. subtilis 168.* The
82 two lineages (subspecies) share a 3.6 Mbp core genome with 3 746 genes. The average

83 divergence of 6.8% in the core genome is small enough so that high-rate transfer is
84 physiologically possible (30) and large enough so that transfer segments can reliably be
85 detected in sequence alignments (SI Appendix, Fig. S1). In addition, there are unique accessory
86 genomes of 0.4 Mbp in the donor and 0.6 Mbp in the recipient, which allow for non-orthologous
87 sequence changes. Our evolution experiment consists of two-day cycles with a 6-step protocol
88 shown in Fig. 1a. On day 1, recipient cells are diluted and cultured in liquid for 4.5 h, subject
89 to UV radiation, and again diluted and inoculated onto an agar plate. On day 2, a single (clonal)
90 colony is grown for 2.5 h in liquid culture, competence is induced and transformation in the
91 presence of donor DNA takes place for 2 h, and then the culture is washed and grown overnight.
92 The last step involves population dynamics with an exponential growth phase of about 4 hrs
93 and a stationary growth phase of about 14 hrs. In this step, selection acts on HGT. In each run
94 of the experiment, a recipient population evolves over 21 consecutive cycles, corresponding to
95 about 200 generations. The full experiment consists of 7 replicate runs. We also perform 3
96 control runs without donor DNA, which are mutation-accumulation experiments under the
97 same population-dynamic conditions as the primary runs. Our laboratory protocol is designed
98 to study HGT at controlled levels of donor DNA and of induced competence. This differs from
99 natural environments, where both factors are subject to environmental fluctuations, regulatory
100 tuning, and ecological interactions between lineages (11, 31).

101 **Gene transfer is the dominant mode of evolution.** The experiments reveal fast and repeatable
102 genome evolution that is driven predominantly by HGT. To infer these dynamics, we perform
103 whole-genome sequencing after cycles 9, 15, and 21, and align the evolved genomes to the
104 donor and the ancestral recipient genome (SI Appendix, Fig. S1). The genome-wide pattern of
105 sequence evolution in hybrids is displayed in Fig. 1b and Fig. S2. Transferred genome segments
106 (green bars) are seen to be distributed over the entire core genome of the recipient. We find
107 HGT by orthologous recombination of genome segments at an approximately constant rate of
108 about 11 genes/hr (counted over the transfer period of 2 hrs/cycle) in all replicate runs. These
109 changes accumulate to an average of about 100 transferred segments per replicate, which cover
110 12% ($\pm$ 3%) of the core genome and affect 500 ($\pm$ 133) core genes (Fig. 1c). Thus, HGT by
111 orthologous recombination causes sequence evolution at an average rate of 100 bp/generation
112 (counted over all cycles). This drastically exceeds the mutation rate of 2×10^{-3} bp/generation
113 determined by a fluctuation assay in the absence of radiation (SI Appendix), as well as the
114 corresponding rate recorded under UV treatment, 3×10^{-1} bp/generation (corresponding to an
115 average of 55 de novo mutations per replicate, Table S1). The HGT rate is also higher than
116 found previously in *Helicobacteri pylori* (17), despite the substantially larger average sequence
117 divergence between donor and recipient.

118 Besides orthologous recombination, we find insertions from the donor accessory genome
119 (average 5 genes per replicate), deletions (average 31 genes per replicate) including the deletion
120 of the mobile element ICEBs1 (in all replicates but not in the control runs, indicating clearance
121 by HGT), and multi-gene duplications (192 genes in one replicate); this data is reported in Table
122 S2. Thus, the by far dominant mode of genome evolution is HGT by orthologous recombination
123 (simply called HGT in the following).

124 **Transfer depends on segment length and sequence similarity.** Transferred segments have
125 an average length of 4200 bp, contain an average of 5.1 genes, and have an approximately

126   exponential length distribution (Fig. 1d), in agreement with previous results in *Streptococcus*
127   *pneumoniae* (32) and *Haemophilus influenza* (33). These statistics show ubiquitous multi-gene
128   transfers with no sharp cutoff on segment length.

129   Fig. 1e displays the distribution of donor-recipient sequence divergence, d, in transferred
130   segments (recorded in 100 bp windows around the inferred recombination start site). By
131   comparing this distribution to the divergence distribution recorded in randomly positioned 100
132   bp windows, we infer a transfer rate $u(d)$ that decreases exponentially with increasing d; see
133   SI Appendix and Fig. S3 for details. This result is in agreement with previous work (14, 15). In
134   our system, the dependence on sequence similarity already generates a significant variation of
135   transfer rates independently of selection on specific genes. About 16% of potential
136   recombination start sites in the core genome have a local divergence $< 5\%$, leading to a transfer
137   rate higher than 3-fold above the genome average; other 3% have a divergence $> 15\%$ and a
138   transfer rate lower than 9-fold below the average (Fig. S3). We conclude that the transfer rate
139   is significantly modulated by the local efficiency of the recombination machinery, but transfer
140   is physiologically possible at the vast majority of genomic sites.

141   **Gene transfer is broadly distributed.** Next, we ask how HGT is distributed over replicate
142   runs and genome loci. For a given gene, the transfer frequency, θ, is defined as the fraction of
143   runs in which that gene is hit by HGT. Fig. 1f shows the histogram of transfer frequencies
144   evaluated over the entire core genome. We find 60% of core genes are transferred in at least
145   one of 7 runs, and 1.7% of core genes are replaced in 4 or more runs. This is consistent with
146   the broad genomic distribution of HGT shown in Fig. 1b. We find only few exceptions from
147   this pattern. First, two hot spots of about 8 kbp and 16 kbp show strong enhancement of HGT,
148   with more than 50% of their genes repeatably swapped in all replicates (Fig. 1f, SI Appendix).
149   Both hot spots encode functional differences between donor and recipient, which are discussed
150   below. To assess their statistical significance, we compare the observed multiple-hit statistics
151   with a null model that is obtained by simulations of a positionally scrambled HGT dynamics
152   with local rates $u(d)$ (SI Appendix, Extended Data Fig. S3). Both hot spots turn out to be highly
153   significant ($P < 2 \times 10^{-3}$); that is, their multiple-hit statistics cannot be explained by local
154   sequence divergence alone. Second, there are two genomic cold spots, which are extended
155   segments where HGT is repeatably suppressed in all replicates. Both cold spots are about 50
156   kb long and are significant deviations from the null model; that is, the absence of transfer is
157   unlikely to be caused by local sequence divergence alone ($P < 0.04$, Fig. S4, SI Appendix).

158   Analyzing HGT by gene ontology (GO) leads to a similarly broad distribution (SI Appendix,
159   Table S3). In most GO categories, the observed average transfer frequencies are consistent with
160   the null model; only essential genes show significantly enhanced transfer ($> 30\%$, $P < 0.03$).
161   Taken together, most genomic loci and most functional classes are accessible to evolution by
162   cross-lineage HGT, allowing the evolution of complex hybridization patterns.

163   **Gene transfer generates a net fitness gain.** We measure the selection coefficients of evolved
164   hybrids and of control populations in competition with ancestral recipient cells. Selection is
165   evaluated separately for the exponential and stationary growth phases, which are two
166   consecutive selection windows in our population dynamics following HGT (Fig. 2a). The
167   stationary phase covers most of the population dynamics interval (14/18 hrs). We find a
168   repeatable fitness increase of about 5% in this phase, which is statistically significant ($P <$

169   0.03, Mann-Whitney U-test) and repeatable across replicate populations. This net fitness
170   increase signals adaptive evolution by HGT; that is, the uptake of donor genes does more than
171   just repair deleterious mutations caused by UV radiation. Consistent with this interpretation,
172   the control populations evolving by mutation accumulation show no comparable fitness
173   increase (Fig. 2a).

174   In the exponential growth phase, the replicate-average fitness remains constant, so that the total
175   balance in average fitness is dominated by the gain in the stationary phase. The control
176   populations decline in average fitness in both phases, as expected for mutation accumulation
177   under UV treatment. Evolved hybrids acquire a large exponential-phase fitness variation across
178   replicate runs (Fig. 2a). This pattern indicates that complex fitness landscapes with positive and
179   negative components govern the evolution of hybrid organisms; the shape of such landscapes
180   will be further explored below.

181   **Gene transfer has genome-wide selective effects.** How are the strong fitness effects of HGT
182   compatible with the ubiquity of transfer across the core genome? In particular, is adaptive
183   evolution by HGT limited to few genomic loci with large effects or can we identify a genetic
184   basis of multiple genes with potentially smaller individual effects? To establish a link between
185   genotype and fitness, we first analyze the multiple-hit statistics of transfers in comparison to
186   the null model of scrambled transfers with rates $u(d)$ (SI Appendix, Fig. S3). By simulations
187   of the null model, we obtain a transfer probability p_0 for each gene that depends only on the
188   donor-recipient mutation pattern in its vicinity. This probability accounts for the fact that local
189   sequence divergence affects homologous recombination. The null model is neutral with respect
190   to gene function; i.e., all genes in loci with the same p_0 have the same multiple-hit probability
191   distribution $P_0(\theta|p_0) = B(7, 7\theta, p_0)$, which is a binomial distribution with expectation value
192   p_0 (SI Appendix). To quantify target and strength of selection, we use a minimal model with
193   selectively enhanced (i.e., adaptive) transfer probabilities $p_+ = \varphi_+ \, p_0$ in a fraction c of the core
194   genes and reduced probabilities $p_- = \varphi_- \, p_0$ in the remainder of the core genome. This mixture
195   model generates a multiple-hit distribution

$$Q(\theta|\, p_0) = c \, B\big(7, 7\theta, \varphi_+ p_0\big) + (1 - c) \, B\big(7, 7\theta, \varphi_- p_0\big). \qquad [1]$$

197   Importantly, the model jointly captures neutral and selective variation of HGT rates, properly
198   discounting physiological effects from the inference of selection. Comparing the observed
199   distribution of multiple hits, $\hat{Q}(\theta|p_0)$, and the null distribution shows an excess count of no-hit
200   ($\theta = 0$) genes and of multiple hits ($\theta \geq 3/7$). Importantly, this pattern is observed in all p_0
201   classes; i.e., independently of the local sequence similarity. The deviations from the null model
202   are strongly significant ($P < 10^{-20}$, SI Appendix), and we ascribe them to selection by HGT.
203   The selection mixture model, with parameters $c = 0.2 \, [0.1, 0.4]$, $\varphi_+ = 1.9 \, [1.6, 2.4]$, and
204   $\varphi_- = 0.75 \, [0.6, 0.84]$ (maximum-likelihood value, confidence interval in brackets) estimated
205   from a Bayesian posterior distribution, explains the observed multiple hits (Fig. 2c) and
206   infers 40 $[20, 60]$ % of the observed transfers to be adaptive (SI Appendix, Fig. S6). These
207   results suggest that HGT is a broad target of positive and negative selection in hybrid
208   organisms, in tune with the experimental results (Fig. 2a).

Frequently transferred genes tend to be upregulated and positively selected. Next, we characterize the effect of HGT on gene expression. Using whole-genome transcriptomics data, we compare gene expression in evolved strains and their ancestor (Fig. 3, Fig. S4). The overall distribution of $\log_2$ fold changes of RNA levels ΔR in the entire genome is balanced and similar to that observed in the control runs, as expected for viable cells (raw data are reported in Table S4). To map correlations between HGT and expression, we partition genes into a class with low transfer frequency, $0 \leq \theta \leq 3/7$, and a class with high transfer frequency, $4/7 \leq \theta \leq 1$. Within each class, we further partition genes by their ancestral lineage (recipient: R, donor: D). We observe that genes with low transfer frequency are not substantially affected in their average expression level. In contrast, genes with high transfer frequency show upregulation in hybrids, which is strongest for genes hit by HGT. A priori, upregulation can compensate for reduced translational or functional efficiency, or it can signal an enhanced functional role of the affected genes. In the first case, we would expect a signal for transferred (D) genes independently of their θ class. The fact that we find upregulation specifically for genes with high transfer frequency but not for others may point to an enhanced functional role. This is consistent with our inference of selection: in the high-transfer class, 88% of the genes are inferred to be under positive selection (Methods).

Selection on gene transfer in functional networks. To link selection to cellular functions, we map the transfer patterns onto gene networks of the recipient organism. Specifically, we use complex, multi-gene operons as units of genomic and functional organization above the level of individual genes. To explain the observed transfer statistics, we use a minimal cross-lineage fitness landscape of the form

$$F(q) = aq + bq(1 - q), \qquad [2]$$

which quantifies the selective effect of an operon as a function of the transfer fraction q, relative to the ancestral ($q = 0$) state (Fig. 4a). The linear term describes directional selection on HGT, the quadratic term intra-network epistasis. Specifically for $b < 0$, this term captures hybrid incompatibilities within operons, leading to a fitness trough at intermediate q and a rebound at larger values of q.

Fig 4b shows the distribution of the average transfer frequency $\bar{\theta}$ for operons containing at least 7 genes (bars denote averages over all genes in a given operon); this threshold captures the 56 most complex operons. For a given operon, $\bar{\theta}$ equals the fraction of its HGT-affected genes, q, averaged over all replicate runs. The corresponding distribution of $\bar{\theta}$ under neutral evolution, which is again obtained by simulations of positionally scrambled transfers (SI Appendix), discounts all HGT correlations between genes due to their spatial clustering in a common operon. To compare data and neutral model, we omit the two hotspots of HGT identified above, which contain operons discussed below. In the remaining data, we find a statistically significant excess of operons with small observed transfer frequency (46% of counts in the data with $\bar{\theta} < 0.1$ vs. 31% in the null model), signaling selection against HGT at the network level ($P = 0.01$, SI Appendix). As shown by simulations of the HGT dynamics in the fitness landscapes of Fig. 4a (SI Appendix), the observed $\bar{\theta}$ distribution in complex operons is consistent with hybrid incompatibilities ($b < 0$) or with negative directional selection ($a < 0$) (Fig. 4c). In Fig. 4d, we show two examples of operons with suppressed transfer. One operon is involved in

251 pectin utilization, while the other one is involved in inositol utilization (Table S5). Both operons
252 have none of their genes replaced in any of the 7 replicates, which is consistent with negative
253 directional selection ($a < 0$). Below each plot, we also display the Protein-Protein Interaction
254 (PPI) links within the operon, which may be indicative of functional links between the operon
255 genes (34).

256 The two genomic hotspots of HGT harbor operons with high average transfer frequency (Fig.
257 4b,e). The *leu* operon has $\bar{\theta} = 0.84$ and 100% of its genes are coherently transferred in 3
258 replicate runs. The *leu* operon confers the ability to grow without external leucine supply.
259 Evolved hybrids gain this function, which is present in the donor but absent in the ancestral
260 recipient (SI Appendix). Uptake of *leu* has been observed previously in transformation essays
261 under leucine starvation (36); our stationary-phase protocol may generate selection for leucine
262 synthesis. The *eps* operon, which contains 13 genes, has $\bar{\theta} = 0.64$ and up to 100% transferred
263 genes. This operon is important for biofilm formation, a function that is strongly impaired in
264 the recipient (35) but potent in the donor. Both operons show an enhancement of $\bar{\theta}$ that deviates
265 significantly from the neutral null model and is consistent with positive directional selection
266 ($a > 0$) (Fig. 4e, SI Appendix). This signals adaptive evolution by HGT at the level of gene
267 networks, generating coherent transfer of its constituent genes. Additional examples of operons
268 with strongly enhanced $\bar{\theta}$ are shown in Fig. S7.

269

270 **Discussion**

271 In this work, we integrate experimental evolution and evolutionary modeling to map effects of
272 horizontal gene transfer on genome dynamics, gene expression and fitness. We show that
273 genome-wide cross-lineage HGT is a fast evolutionary mode generating hybrid bacterial
274 organisms. This mode is repeatably observed in all replicate runs of our experiment. After about
275 200 generations, *Bacillus* hybrids have acquired about 12% donor genes across the entire core
276 genome, but coherently transferred functional gene networks reach up to 100% transferred
277 genes in individual runs.

278 Despite its broad genomic pattern, the HGT dynamics is far from evolutionary neutrality.
279 Evolved hybrids show a substantial fitness increase compared to the ancestral strain in
280 stationary growth, which occurs repeatably across all replicates. Hence, HGT does more than
281 just repairing deleterious mutations, which are caused by UV radiation in our protocol: it carries
282 a net adaptive benefit. The adaptive dynamics has a broad genomic target; we infer some 40%
283 of the observed transfers to be under positive selection. In addition, we find positive and
284 negative selection on HGT in a range of functional gene networks. We conclude that in our
285 system, unlike for uptake of resistance genes under antibiotic stress (26, 27), HGT does not
286 have just a single dominant target of selection. Rather, evolution by orthologous recombination
287 appears to tinker with multiple new combinations of donor and recipient genes, using the
288 combinatorial complexity of hundreds of transferred genes in each replicate run. This picture
289 is consistent with the substantial fitness variation across replicates, which is most pronounced
290 in the exponential growth phase.

291 To capture HGT under genome-wide selection, we introduce the concept of a cross-lineage
292 fitness landscape. Such landscapes describe the fitness effects of orthologous recombination
293 between two (sub-)species, starting from the unmixed genomes as focal points of *a priori* equal
294 rank. This feature distinguishes cross-lineage landscapes from empirically known landscapes
295 for mutation accumulation within lineages, e.g. for antibiotic resistance evolution, most of
296 which have a single global fitness peak. Here we begin to map building blocks of cross-lineage
297 landscapes (Fig. 4). These include positive selection for uptake of gene networks, well as
298 selection against transfer in genes and operons, suggesting that hybrid incompatibilities
299 between donor and recipient and epistasis may play a role in the observed HGT dynamics.
300 Future massively parallel experiments will allow a more systematic mapping of cross-species
301 fitness landscapes, which combines these building blocks into a systems picture. Such
302 experiments will also show how different compositions of donor DNA and different selection
303 pressures applied in the experimental cycle modulate the HGT dynamics on cross-lineage
304 landscapes.

305 Our observation of fast HGT under broad selection suggests that evolution navigates the cross-
306 lineage landscape in an efficient way. That is, deep cross-lineage fitness valleys must be sparse
307 enough for evolution by HGT to find viable paths of hybrid evolution with increasing transfer
308 fraction. Stronger and more ubiquitous fitness barriers at larger donor-recipient distances or
309 larger transfer fractions may eventually halt gene uptake, but we have yet to see where the limits
310 of genome-wide HGT are.

311 Two key features of our experiments enable these dynamics. First, choosing a donor-recipient
312 pair with a sequence divergence close to the species boundary generates more, often subtle
313 functional differences between orthologous genes – that is, more potential targets of selection
314 on HGT – than closer pairs or single populations. Second, permissive population dynamics with
315 recurrent bottlenecks allow the (transient) fixation of deleterious recombinant genomes, which
316 can bridge fitness valleys of hybrid incompatibilities and act as stepping stones for subsequent
317 adaptation. Together, our results suggest that laboratory evolution by HGT can become a
318 factory for evolutionary innovation. An exciting perspective is to use cross-lineage HGT
319 together with artificial evolution in order to engineer functional novelties.

320

321

References

1. J. P. Gogarten, J. P. Townsend, Horizontal gene transfer, genome innovation and evolution. *Nature Reviews Microbiology* **3**, 679-687 (2005).

2. G. Schönknecht *et al.*, Gene transfer from bacteria and archaea facilitated evolution of an extremophilic eukaryote. *Science* **339**, 1207-1210 (2013).

3. R. Niehus, S. Mitri, A. G. Fletcher, K. R. Foster, Migration and horizontal gene transfer divide microbial genomes into multiple niches. *Nature communications* **6**, 8924 (2015).

4. S. M. Soucy, J. Huang, J. P. Gogarten, Horizontal gene transfer: building the web of life. *Nature Reviews Genetics* **16**, 472-482 (2015).

5. A. S. Kondrashov, Classification of Hypotheses on the Advantage of Amphimixis. *J Hered* **84**, 372-387 (1993).

6. T. Held, D. Klemmer, M. Lässig, Survival of the simplest in microbial evolution. *Nature communications* **10**, 1-11 (2019).

7. M. Lin, E. Kussell, Inferring bacterial recombination rates from large-scale sequencing datasets. *Nature methods* **16**, 199-204 (2019).

8. T. Sakoparnig, C. Field, E. van Nimwegen, Whole genome phylogenies reflect long-tailed distributions of recombination rates in many bacterial species. *BioRxiv* 10.1101/601914 (2019).

9. M. J. Lercher, C. Pal, Integration of horizontally transferred genes into regulatory interaction networks takes many million years. *Mol Biol Evol* **25**, 559-567 (2008).

10. I. Chen, D. Dubnau, DNA uptake during bacterial transformation. *Nature Reviews Microbiology* **2**, 241-249 (2004).

11. B. Maier, "Competence and Transformation" in Bacillus: Molecular and Cellular Biology, P. Graumann, Ed. (Caister Academic Press, 2017), pp. 395-414.

12. O. Popa, T. Dagan, Trends and barriers to lateral gene transfer in prokaryotes. *Current Opinion in Microbiology* **14**, 615-623 (2011).

13. C. M. Thomas, K. M. Nielsen, Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nature reviews microbiology* **3**, 711 (2005).

14. B. Carrasco, E. Serrano, H. Sanchez, C. Wyman, J. C. Alonso, Chromosomal transformation in Bacillus subtilis is a non-polar recombination reaction. *Nucleic Acids Research* **44**, 2754-2768 (2016).

15. M. Vulić, F. Dionisio, F. Taddei, M. Radman, Molecular keys to speciation: DNA polymorphism and the control of genetic exchange in enterobacteria. *Proceedings of the National Academy of Sciences* **94**, 9763-9767 (1997).

16. D. Kidane, S. Ayora, J. B. Sweasy, P. L. Graumann, J. C. Alonso, The cell pole: the site of cross talk between the DNA uptake and genetic recombination machinery. *Critical reviews in biochemistry and molecular biology* **47**, 531-555 (2012).

17. S. Bubendorfer *et al.*, Genome-wide analysis of chromosomal import patterns after natural transformation of Helicobacter pylori. *Nature communications* **7**, 1-12 (2016).

18. D. A. Baltrus, K. Guillemin, P. C. Phillips, Natural transformation increases the rate of adaptation in the human pathogen Helicobacter pylori. *Evolution: International Journal of Organic Evolution* **62**, 39-49 (2008).

19. A. L. Utnes *et al.*, Growth phase-specific evolutionary benefits of natural transformation in Acinetobacter baylyi. *The ISME journal* **9**, 2221-2231 (2015).

20. D. J. P. Engelmoer, I. Donaldson, D. E. Rozen, Conservative Sex and the Benefits of Transformation in Streptococcus pneumoniae. *Plos Pathog* **9** (2013).

21. O. H. Ambur, J. Engelstadter, P. J. Johnsen, E. L. Miller, D. E. Rozen, Steady at the wheel: conservative sex and the benefits of bacterial transformation. *Philos T R Soc B* **371** (2016).

372 22. S. Bershtein *et al.*, Protein Homeostasis Imposes a Barrier on Functional Integration of
373   Horizontally Transferred Genes in Bacteria. *PLoS Genet* **11**, e1005612 (2015).
374 23. J. D. Bloom, S. T. Labthavikul, C. R. Otey, F. H. Arnold, Protein stability promotes
375   evolvability. *Proc Natl Acad Sci U S A* **103**, 5869-5874 (2006).
376 24. S. Bedhomme *et al.*, Evolutionary Changes after Translational Challenges imposed by
377   Horizontal Gene Transfer. *Genome Biol Evol* 10.1093/gbe/evz031 (2019).
378 25. P. A. Lind, C. Tobin, O. G. Berg, C. G. Kurland, D. I. Andersson, Compensatory gene
379   amplification restores fitness after inter-species gene replacements. *Mol Microbiol* **75**,
380   1078-1089 (2010).
381 26. C. B. Wadsworth, B. J. Arnold, M. R. A. Sater, Y. H. Grad, Azithromycin resistance
382   through interspecific acquisition of an epistasis-dependent efflux pump component and
383   transcriptional regulator in Neisseria gonorrhoeae. *MBio* **9**, e01419-01418 (2018).
384 27. H. Y. Chu, K. Sprouffske, A. Wagner, Assessing the benefits of horizontal gene transfer
385   by laboratory evolution and genome sequencing. *BMC evolutionary biology* **18**, 54
386   (2018).
387 28. D. R. Zeigler, The genome sequence of Bacillus subtilis subsp. spizizenii W23: insights
388   into speciation within the B. subtilis complex and into the history of B. subtilis genetics.
389   *Microbiol-Sgm* **157**, 2033-2041 (2011).
390 29. P. H. Brito *et al.*, Genetic Competence Drives Genome Diversity in Bacillus subtilis.
391   *Genome Biology and Evolution* **10**, 108-124 (2018).
392 30. B. Carrasco, E. Serrano, A. Martin-Gonzalez, F. Moreno-Herrero, J. C. Alonso, Bacillus
393   subtilis MutS Modulates RecA-Mediated DNA Strand Exchange Between Divergent
394   DNA Sequences. *Front Microbiol* **10**, 237 (2019).
395 31. P. Stefanic, B. Kraigher, N. A. Lyons, R. Kolter, I. Mandic-Mulec, Kin discrimination
396   between sympatric Bacillus subtilis isolates. *Proc Natl Acad Sci U S A* **112**, 14042-
397   14047 (2015).
398 32. N. J. Croucher, S. R. Harris, L. Barquist, J. Parkhill, S. D. Bentley, A high-resolution
399   view of genome-wide pneumococcal transformation. *Plos Pathog* **8**, e1002745 (2012).
400 33. J. C. Mell, J. Y. Lee, M. Firme, S. Sinha, R. J. Redfield, Extensive cotransformation of
401   natural variation into chromosomes of naturally competent Haemophilus influenzae. *G3*
402   *(Bethesda)* **4**, 717-731 (2014).
403 34. D. Szklarczyk *et al.*, STRING v11: protein-protein association networks with increased
404   coverage, supporting functional discovery in genome-wide experimental datasets.
405   *Nucleic Acids Res* **47**, D607-D613 (2019).
406 35. R. Gallegos-Monterrosa, E. Mhatre, A. T. Kovacs, Specific Bacillus subtilis 168
407   variants form biofilms on nutrient-rich medium. *Microbiology* **162**, 1922-1932 (2016).
408 36. M. Albano, J. Hahn, D. Dubnau, Expression of competence genes in Bacillus subtilis. *J*
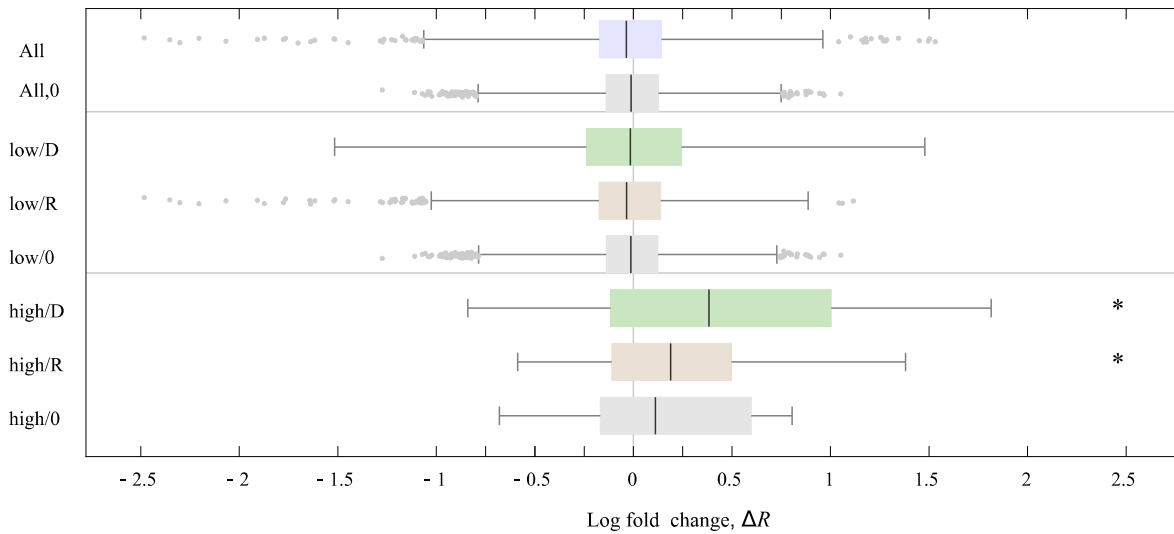409   *Bacteriol* **169**, 3110-3117 (1987).

410
411

Figures
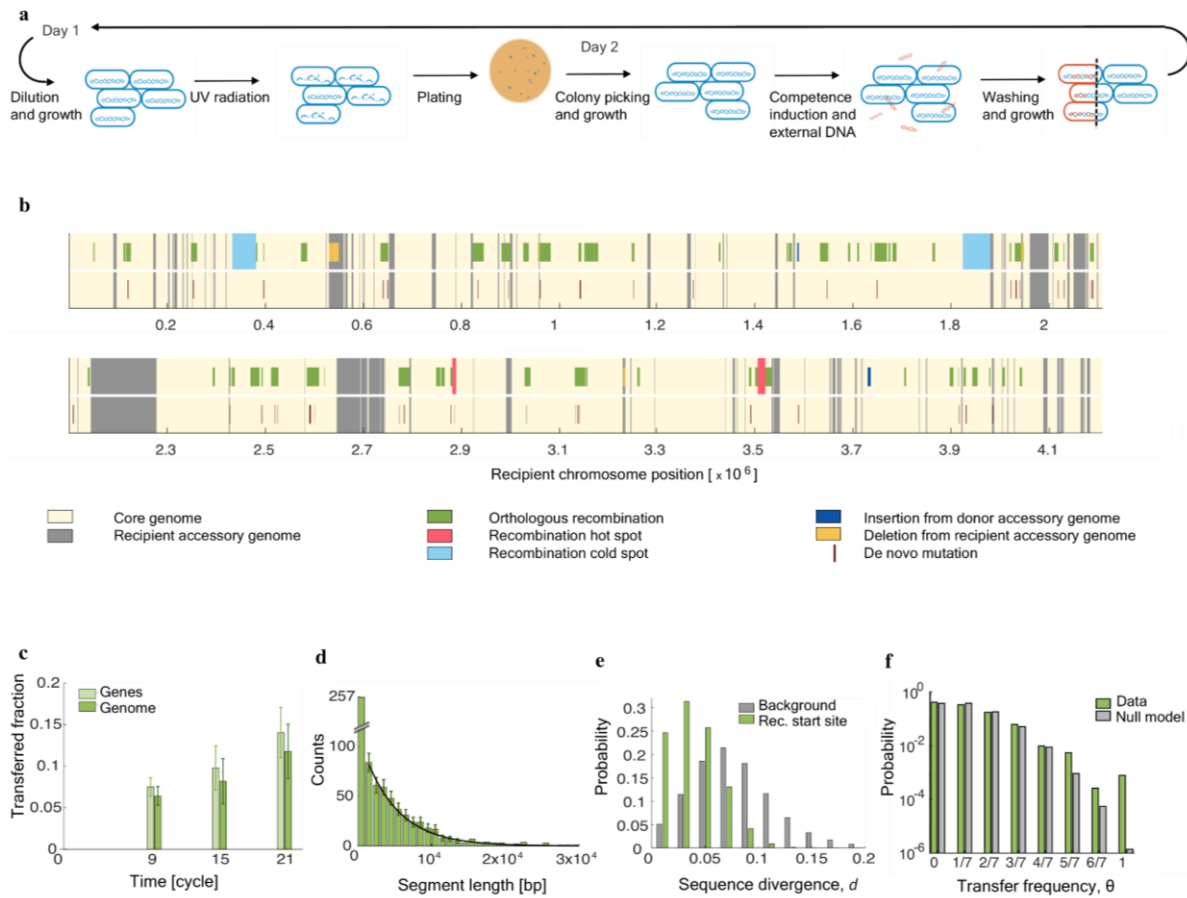
Fig. 1. Evolution of *B. subtilis* hybrids. **(a)** Experimental design. A single two-day (10 generation) cycle contains 6 steps, including two growth phases with irradiation and horizontal gene transfer (HGT), respectively. **(b)** Transferred segments (green) are broadly distributed over the recipient core genome (light shading) with two hot spots (red) and two cold spots (blue). De novo mutations occur throughout the recipient genome, and there are a few deletions from the recipient genome (orange). Data are shown for run 4; see Fig. S3 and Tables S1 and S2 for data from all runs. **(c)** Time-dependence of HGT. Fraction of core genes (light green) and of core genome (dark green) affected by HGT after 9, 15, and 21 cycles (mean and standard deviation across the 7 parallel runs). **(d)** Length distribution of transferred genome segments (bars: count histogram, black line: exponential fit for segments > 1000 bp). **(e)** Inferred distribution of donor-recipient sequence divergence, d, in 100 bp windows around the recombination start site of transferred segments (green) and in scrambled 100 bp windows (gray). **(f)** Distribution of the transfer frequency per gene, $\hat{Q}(\theta)$, in the 7 parallel runs, counted across the recipient core genome (green); corresponding distribution $P_0(\theta)$ from simulations of the null model (gray).
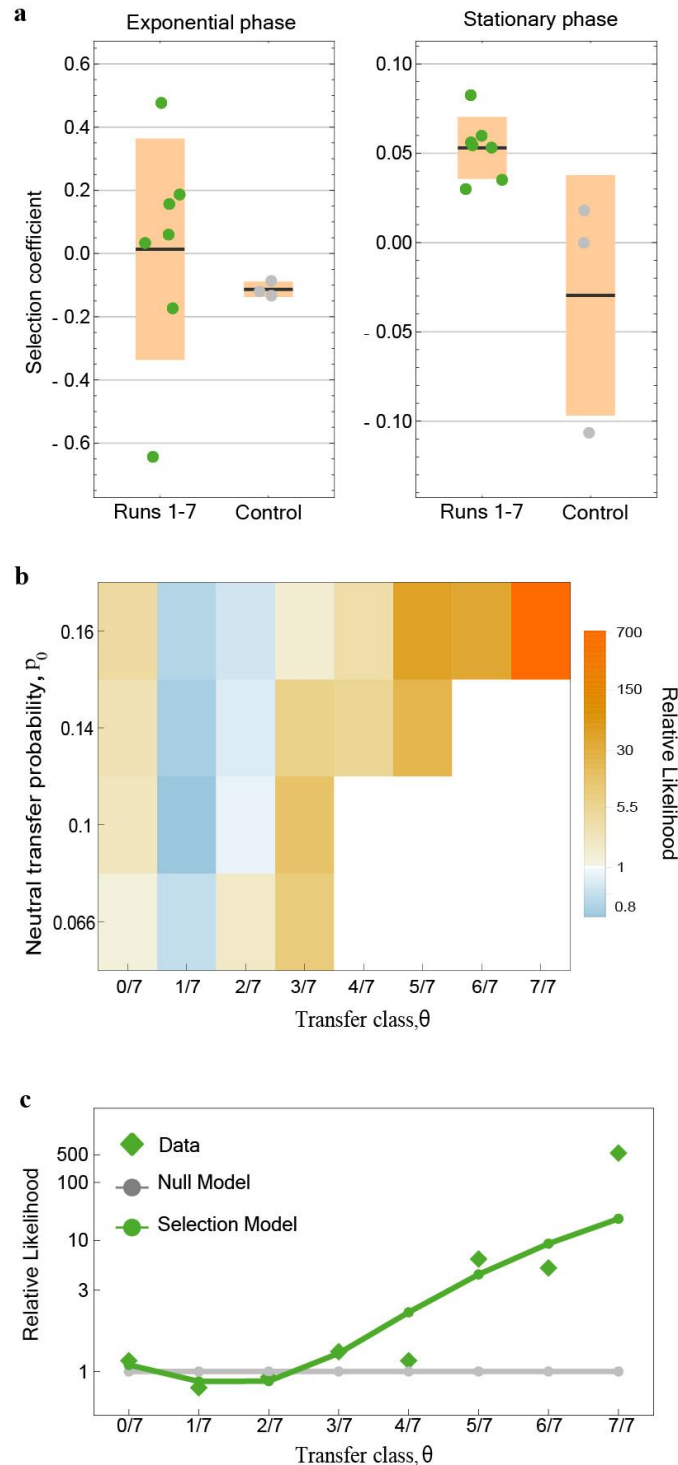
Fig. 2. Selective effects of horizontal gene transfer. **(a)** Transfer affects hybrid fitness. Mean selection coefficient of hybrids compared to the ancestral recipient population in the exponential phase and in the stationary phase (colored dots: data from individual runs, bars and boxes: mean and standard deviation over all 7 runs); control data of runs without donor DNA (gray). **(b)** Genome-wide selection on transfer. The relative likelihood of transfer, $\hat{Q}(\theta|p_0)/P_0(\theta|p_0)$, is shown for genes binned by transfer class ($\theta = 0/7, 1/7, ..., 7/7$) and by the neutral transfer probability p_0, which depends on sequence similarity. **(c)** The relative likelihood of transfer aggregated over p_0 bins, $\hat{Q}(\theta)/P_0(\theta)$, is shown together with the corresponding likelihood ratio $Q^*(\theta)/P_0(\theta)$ of the maximum-likelihood selection model (green circles); see SI Appendix. Relative to the neutral null model (gray baseline), the selection model has an enhanced transfer probability ($p_+/p_0 = 1.9$) in a fraction $q = 0.2$ of the recipient core genes and a reduced probability ($p_-/p_0 = 0.75$) in the remainder of core genes.

12

447
448
449

Fig. 3. Gene expression in hybrids. Log$_2$ RNA fold changes, ΔR, with respect to the ancestral recipient (whisker plots, blue line: mean, box: 1st and 3rd quartiles, bars: 99% percentiles, dots: outliers) are shown for different gene classes: all genome, genes with low transfer frequency ($0 \leq \theta \leq 3/7$), and genes with high transfer frequency ($4/7 \leq \theta \leq 1$); cf. Fig. 1f. In each gene class, we show ΔR whisker plots separately for non-transferred recipient genes (R) (red) and for transferred orthologous donor genes (D) (green), together with the corresponding changes in the control experiments without donor DNA (0) (grey). For example, a gene subject to HGT in replicates 1, 2, 5 is in the low-transfer class; its ΔR values from replicates 1, 2, 5 (3, 4, 6, 7) contribute to the low/D (low/R) statistics. Asterisks mark statistically significant changes of the average ΔR for upregulation of high/D and high/R genes ($P < 10^{-3}$, T-test) compared to the ancestor.

450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468

469
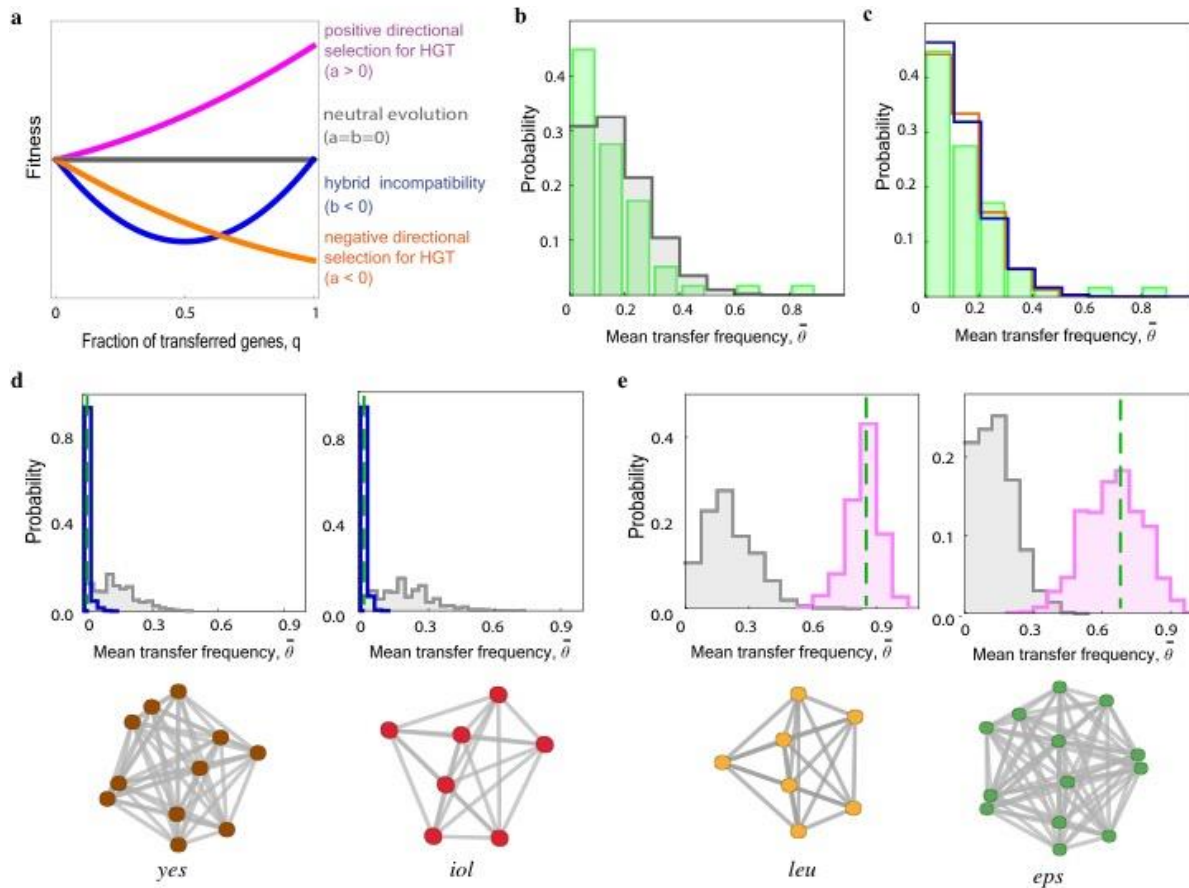
470
471
Fig. 4. Cross-lineage fitness effects in gene networks. **(a)** Cross-lineage fitness landscapes given by eq. [1] with hybrid incompatibilities ($b < 0$, blue), predominantly negative directional selection ($a < 0$, orange), and predominantly positive directional selection ($a > 0$, magenta). **(b-c)** Distribution of the observed average transfer frequency, $\bar{\theta}$, for recipient operons containing at least 7 genes (green). Corresponding distributions (b) for the neutral null model (gray) and (c) for fitness models given by eq. [1] with hybrid incompatibilities ($b < 0$, blue) and with negative directional selection ($a < 0$, orange). **(d)** Examples of low-transfer operons, *yes* and *iol*. Observed values of $\bar{\theta}$ (dashed lines) are reduced in comparison to the neutral null model (gray) and consistent with negative directional selection ($a < 0$, blue); see SI Appendix. **(e)** Examples of high-transfer operons, *leu* and *eps*. Observed values of $\bar{\theta}$ (dashed lines) are enhanced in comparison to the neutral null model (gray) and consistent with positive directional selection ($a > 0$, magenta). The deviation from the neutral model is statistically significant (*leu*: $P < 10^{-3}$, *eps*: $P < 10^{-3}$; SI Appendix). PPI links for these operons are shown below (34). Simulations of transfer under selection use the fitness landscape of eq. [2] with selection coefficients a, b listed in SI Appendix.

486
487

488

14

489 **SI Appendix**

490 The SI Appendix describes experimental procedures, sequence analysis, and evolutionary
491 analysis of HGT.

492

496

497 **Author Contributions.** First authors: JJP developed, performed, and analyzed the core
498 experiments; FP and SP developed and performed evolutionary analysis. JJP, FP, SP, MY, ML,
499 and BM designed research; JJP, MY, IR, and MF performed the experiments; VK and JJP
500 processed sequence data; VK and MY processed transcriptomics data; ML and BM wrote the
501 article; all authors discussed and approved the article.

502
503
504

505   **SI Appendix**

506

507   **1.   Experimental procedures**

508   **Strains, Media, and Growth.** The donor strain used was 2A9 (*B. subtilis subsp. spizizenii*
509   *W23*). All other strains used in this study were derived from strain BD630 (*B. subtilis subsp.*
510   *subtilis 168 his leu met*). *B. subtilis* strains were grown either in liquid LB medium or in
511   competence medium (1), supplemented with 0.5 % glucose, 50 µg/ml L-histidine, L-leucine,
512   L-methionine, and kanamycin (5µg/ml), or spectinomycin (100 µg/ml) at 37°C. Growth was
513   monitored by measuring the OD_{600} on an Infinite M200 plate reader (Tecan, Männedorf,
514   Switzerland). Competent cells were prepared as described previously and transformed by
515   following the standard protocol (1).

516   In *B. subtilis*, natural competence for transformation develops at high cell densities.
517   Competence development is regulated by an extended network of regulatory proteins (2). This
518   regulatory network is a large mutational target and mutations are likely to affect the probability
519   of competence development and thus HGT. Since competence development is not subject of
520   this study, we circumvented this problem by setting the expression of the master regulator for
521   competence, *comK*, under the control of an IPTG-inducible promoter. The recipient strain
522   Bs166 (*his leu met amyE::P$_{hs}$comK(spc) comK::kan*) was generated by transforming strain
523   BD3836 (3) with genomic DNA from strain Bs075 (4). For competition experiments, we
524   generated the reporter strain Bs175 (*his leu met amyE::P$_{hs}$comK(spc) comK::kan lacA::PrrnE-*
525   *gfp (erm)*) as follows. *PrrnE-gfpmut2* was amplified from genomic DNA of Bs139 (4) using
526   primers MeY87 (5'- AAGGAATTCCGGATCCGAAGCAGGTTAT-3') and MeY88 (5'-
527   GGACTAGTTTATTTGTATAGTTCATCCATGC -3'). Both the PCR-fragment and the donor
528   plasmid pBS2E (obtained from the *Bacillus subtilis* Stock Center) were double digested with
529   EcoRI and SpeI. The PCR-fragment was ligated into the linearized plasmid, after
530   dephosphorylation of the 5'-sticky ends of pBS2E. The ligation product pBS2E_*PrrnE-*
531   *gfpmut2* was then transformed into Bs166 to generate Bs175.

532   **Whole Genome Sequencing.** Clonal genomes of ancestral and evolved populations were
533   obtained using next generation sequencing (NGS) methods, in particular Illumina HiSeq.
534   Samples were prepared by growing a frozen culture overnight on an LB Agar plate at 37° C,
535   5% CO_2. The subsequent day, an individual colony was selected and grown overnight in
536   competence medium. A 2 mL aliquot of that culture was pelleted at 16.7 x g for 3 min, decanted,
537   and then frozen at -20° C. Additionally, a 1 mL aliquot of the overnight CM culture was mixed
538   with DMSO (10% v/v) and stored for reference at -80° C. Genomic DNA was isolated from the
539   frozen pellet using the Qiagen Dneasey Blood & Tissue Kit (Hilden, Germany) according to
540   the manufacturer's instructions. A small aliquot of the Isolated DNA was run on a 1% agarose
541   gel with a 1 kb plus DNA Ladder (Thermo Scientific) to check for degradation. Non-degraded
542   samples were sent to GATC Biotech (Konstanz, Germany) for NGS. Sequencing was
543   performed on an Illumina HiSeq 3000/4000 system with 150 bp paired reads and an average
544   depth of >500.

545   **RNA sequencing.** Clonal transcripts of ancestral and evolved populations were obtained using
546   next generation sequencing (NGS) methods, in particular Illumina HiSeq. Samples were
547   prepared by growing a frozen culture overnight on an LB Agar plate at 37° C, 5% CO_2. The

subsequent day, individual colonies were selected and grown in 1 ml competence medium. After 2.5 h growth, the cultures were diluted to an 0.1 OD in competence medium and were incubated for 3 h. Total RNA was isolated by using the Qiagen RNeasy Protect Bacteria Mini Kit (Hilden, Germany) according to the manufacturer's instructions. Samples were sent to Cologne Genomic Center (Cologne, Germany) for NGS. Prior to sequencing ribosomal RNA was depleted from the samples by using the Illumina Ribo-Zero rRNA Removal Kit (San Diego, USA). Sequencing was performed on an Illumina HiSeq system with 75 bp paired reads and on average 5 million reads per sample.

Determination of mutation rate. Spontaneous mutation rates to nalidixic acid resistance were determined using a fluctuation assay (5). Twenty parallel cultures grown overnight in LB medium were inoculated in competence medium with a starting OD_{600} 0.01 and grown for 6 h. Subsequently, each culture was plated on LB plates containing 15 µg/ml nalidixic acid and incubated at 37°C. After 24 h, the number of colonies per plate was determined. The total number of cells was determined by plating dilutions of three parallel cultures on LB agar plates. The mutation rate was calculated with the MSS-Maximum-Likelihood Method (6). These experiments were repeated at least three times. The mutation rate was determined to be $(5.04 \pm 3.82) \times 10^{-10}$ per nucleotide per generation.

Evolution Experiment. The evolution experiment was composed of continuous repetitions of a two-day cycle, consisting of six steps: dilution, radiation, plating, colony selection and regrowth, competence induction and addition of extracellular DNA, washing and overnight growth (Fig. 1a). Seven replicate ancestral clones of the recipient strain $\Delta comK\ comK_{ind}$ (Bs166) were used for transformation with genomic DNA from *B. subtilis* W23 and three clones were used for the control experiments. Initially, all strains were grown overnight in 1 mL LB medium at 37°C, 250 rpm.

Overnight cultures were diluted $1:10^3$ in fresh competence medium and grown for 4.5 h (37°C, 250 rpm) in a 24-well microtiter plate (1 mL, final OD ≈ 0.25; Greiner Bio-one). Microtiter plates were covered with rayon film adhesive covers (VWR). Cultures were radiated in the microtiter plate (without cover) in a 600 Jcm^{-2} UV-C light chamber (Bio-Link BLX-E crosslinker). We included the irradiation step to damage the chromosomal DNA hypothesizing that donor DNA was used as a template for DNA repair as proposed earlier (7). A 100 µL aliquot of each radiated culture was diluted $1:10^4$, plated onto a 10 cm LB-agar plate, and incubated overnight at 37°C, 5% CO_2. A random colony was picked from each plate using a 200µL pipette tip, mixed into 1 mL fresh competence medium, and grown for 2.5 h (OD ≈ 0.25). IPTG [600µM] was added to each culture along with genomic DNA equivalent to two genomes per cell. Genome equivalents were calculated assuming 1 bp = 650 Da and a culture density of 10^8 CFU/mL. After growing induced cultures for an additional 2 h (37°C), each culture was washed twice (16,800xg, 1 mL competence medium) and grown overnight (37°C, 250 rpm). This completed one cycle and was repeated, starting again with dilution. Samples were frozen for later analysis every second cycle, starting with cycle 3. For each culture, a 500µL mixture with DMSO (10% v/v) was stored at -80°C.

Gain-of-function experiments. Our statistical analysis identifies two HGT hot spots (section 3, Fig. S3). These hot spots contain operons encoding cellular functions that are present in the donor but absent in the recipient strain: the *leu* operon enables leucine synthesis, and the *eps*

17

591 operon is important for biofilm formation. (i) We tested the gain of leucine synthesis in evolved
592 hybrid strains (cycle 21) that have acquired the *leu* operon by growing the hybrids overnight in
593 defined medium without leucine. All hybrid strains but not the recipient grew indicating that
594 the hybrid strains are *leu* prototroph. (ii) The experiments testing biofilm formation in evolved
595 hybrids have remained inconclusive.

596 **Fitness measurements.** The selection coefficients reported in Fig. 2a are determined by
597 competition experiments between evolved strains and the ancestral recipient, which are carried
598 out separately in the exponential and stationary growth phase. The selection coefficient between
599 an evolved (hybrid or control) population (e) and the ancestor (a) in a competition period $t - t_0$
600 t_0 is defined as

$$s_{e,a} = \frac{1}{t - t_0} \log\left(\frac{x_e(t)}{x_a(t)} \Big/ \frac{x_e(t_0)}{x_a(t_0)}\right), \tag{S1}$$

601

602 where $x_e(t_0) = 1 - x_a(t_0)$ and $x_e(t) = 1 - x_a(t)$ are the population fractions of the
603 competing lineages at the beginning and at the end of the competition period, respectively. As
604 described in (4), we measure selection coefficients $s_{e,g}$ against a *gfp*-expressing reporter strain
605 and we correct for the fitness effect $s_{g,a}$ of the reporter; i.e., $s_{e,a} = s_{e,g} - s_{g,a}$. Each selection
606 coefficient is obtained as the average over at least 3 independent competition runs. We note that
607 selection coefficients in the exponential phase are essentially differences between cell
608 duplication (birth) rates, whereas stationary-phase selection coefficients are determined by
609 time-dependent birth and death rates. This provides a rationale for the different fitness effects
610 of HGT in the exponential and stationary phase.

611

612 **2. Sequence analysis**

613 **Sequencing pipelines.** DNA-seq reads from each library were properly paired and trimmed in
614 Trimmomatic (version 0.36) (8). We did quality checks of raw and filtered reads using FastQC
615 (v0.11.5)(9).

616 We used BWA package (v0.7.12-r1039) (10) to align filtered paired reads against the reference
617 genomes (default settings). First, we corrected reference genomes of *B. subtilis* 168 (NCBI,
618 NC_000964.3) and *Bacillus subtilis* subsp. *spizizenii* str. W23 (NCBI, NC_014479.1) for point
619 mutations and short indels present in our ancestral/donor strain with
620 *FastaAlternateReferenceMaker* (GATK, v3.5-0-g36282e4) (11). Adequately, we
621 accommodated all the gene positions in the reference annotation files (gff3, gtf, bed) using
622 custom-made scripts. We mapped filtered paired reads from evolved strains against the adjusted
623 reference genomes using BWA tool (option mem) under two criteria: 1.) loose mapping - a
624 *mismatch penalty* (-B) and *gap open penalty* (-O) of 1 - to determine mutations between the
625 evolved strain and the reference *B. subtilis* 168 genome (NCBI, NC_000964.3), 2.) hard
626 mapping - a *mismatch penalty* (-B) and *gap open penalty* (-O) of 100 - to determine de novo
627 insertions from *Bacillus subtilis* subsp. *spizizenii* str. W23 (NCBI, NC_014479.1) in the
628 evolved species.

629 We used SAMtools (v1.3.1) (12) to call and pileup SNPs from BWA output files, for each
630 evolutionary run in cycles 9, 15 and 21. We kept only homozygous variants. Based on the

631 quality score distributions for SAM files (Picard v2.6.0) and per base coverage distributions
632 (bedtools, v2.26.0) (13), we filtered out insufficiently covered SNPs and SNPs with poor
633 quality.

634 **Detection of de novo mutations.** Using the hard mapping pipeline outlined in Sequencing
635 Pipelines, we aligned the reference genomes of the donor and the recipient. We created a master
636 list of positions, and bases, representing point differences between the donor and the recipient
637 genome. For each DNA-seq library in cycle 9, 15 and 21, we called all the mutations that differ
638 in position information, and base composition, compared to the master list.

639 **Inference of HGT segments.** The ancestral recipient and the donor strain share 3648489 bp of
640 core genome, which is partitioned in 142 segments alternating with segments of the recipient
641 accessory genome (Fig. 1b). The core genome sequence contains 248511 single-nucleotide sites
642 with donor-recipient divergence. At each of these sites, the evolved sequence of a given run has
643 a recipient (R) or donor (D) consensus. We infer transfer segments in the evolved sequence
644 sequentially in the direction of increasing genome coordinate, using the following algorithm
645 (Fig. S1a): (1) The start of each transfer segment is marked by a sequence of two consecutive
646 D alleles (5' marker). The starting coordinate of the segment is then assigned to the midpoint
647 between the 5' marker and its left flanking site, which is either the previous R site or the start
648 of the core genome segment. (2) The end of each segment is marked by a sequence of k
649 consecutive R alleles (3' marker, the optimal value of the parameter k will be inferred below).
650 The end coordinate of the segment is then assigned to the midpoint between the 3' marker and
651 its left flanking site, which is the previous D site.

652 These rules reflect the fact that the sequence read of an evolved run is determined not only by
653 horizontal transfer, but also by point mutations, mismatch repair, and sequencing errors. These
654 processes introduce local noise to the R/D allele assignment, predominantly singlet allele
655 changes. In previous work (14), transfer segments have been inferred as clusters of D alleles in
656 evolved sequences, whereas our method uses the full local information of R and D alleles.

657 The results of the transfer inference are robust with respect to variation of the inference
658 algorithm; for example, the distribution of transfer segment lengths and of inter-segment gap
659 lengths depends only weakly on k (Fig. S1b-c; cf. Fig.1d and Fig. S4). This is consistent with
660 most gaps between consecutive transfer segments being much longer than the typical sequence
661 length of 5' and 3' markers.

662 To further evaluate the performance of the transfer inference algorithm, we record the fraction
663 of R alleles that fall within inferred transfer segments, $x_R = n_R/(n_R + m_R)$, and the fraction
664 of D alleles that fall outside inferred transfer segments, $y_D = m_D/(n_D + m_D)$, where n_R, n_D
665 and m_R, m_D are the allele counts within and outside segments, respectively. In Fig. S1d, the
666 replicate averages of x_R and y_D are plotted for different values of the parameter k. Both of these
667 error fractions are small for all k, which signals that gene transfer is the dominant process
668 shaping the evolved genomes. Furthermore, the majority (70%) of the remaining R alleles
669 within segments are singlets (i.e., flanked by two D sites), which is consistent with low-rate
670 local point processes independent of gene transfer. We choose $k = 5$ for the analysis of the
671 main text; this parameter value minimizes $x_R + y_D$.

672 To test the positional accuracy of our inference, we apply the optimized ($k = 5$) algorithm to
673 simulated sequence data representing evolved genomes. These data are generated in two steps:
674 (1) We randomly mark a set S of non-overlapping sequence segments in the recipient core
675 genome, using number and length distribution of the transfer segments recorded in the actual
676 evolved genomes. (2) For all loci with donor-recipient sequence divergence, we randomly
677 assign R/D alleles with probabilities $\tilde{x}_R = n_R/(n_R + n_D)$ and $1 - \tilde{x}_R$ within S, and with
678 probabilities $\tilde{y}_R = m_R/(m_R + m_D)$ and $1 - \tilde{y}_R$ outside S, respectively. This step preserves
679 the allele frequency distributions recorded in the actual genomes. We then apply the inference
680 algorithm to this artificial genome in order to reconstruct the set S. A scatter plot of the inferred
681 length against the input length of S segments shows the high fidelity of the algorithm on
682 individual segments (Fig. S1e). In most cases, an input segment of length l_{in} is uniquely
683 mapped to an inferred segment of length $l \approx l_{in}$. In about 0.5% of the cases, two close input
684 segments of length $l_{in,1}$ and $l_{in,2}$ are reported as a paired segment of length $l \approx l_{in,1} + l_{in,2}$; this
685 affects the total coverage only weakly. Defining the inference error $\Delta l = |l - l_{in}|$ for uniquely
686 mapped segments and $\Delta l = |l - (l_{in,1} + l_{in,2})|$ for paired segments, we find an average value
687 $\Delta l = 18$ bp in the simulated genome data. This uncertainty is of order $d^{-1} = 15$ bp, where d
688 denotes the core genome sequence divergence of the donor and recipient lineages. That is, the
689 positional accuracy of the transfer inference algorithm is set by the density of divergent loci
690 between donor and recipient.

691 **Inference of de novo insertions.** A separate algorithm was designed to detect the integration
692 of *B. subtilis* W23 specific genes in the evolved replicates. Start and end positions of the
693 accessory *B. subtilis* W23 genes were taken from Table S1 of (15). The genome coverage, per
694 base, was calculated from the stringently mapped reads using bedtools (v2.26.0) (13) and the
695 per base pair depth setting (-d). Segments had to be at least 200 bp long and have an average
696 coverage greater than that of the genome-wide average coverage (ignoring positions with zero
697 mapped reads).

698 **Inference of deletions.** Using bedtools (version 2.17.0, function coverage) (13), we compute
699 read coverage across the whole genome (window size 100 bp, step by 100 bp) for each DNA-
700 seq library. Genes overlapping with depressed regions (the mean of coverage - 2x standard
701 deviation) are visualized in R packaged TEQC (16) to inspect coverage patterns. Altogether,
702 we find 74 deleted genes. Each of the 74 genes also has zero read counts in a relevant RNA-
703 seq library.

704 **RNA-seq analysis.** Reads from each library are properly paired and trimmed in Trimmomatic
705 (version 0.38) (8). We map the filtered reads against the reference genome (NCBI,
706 NC_000964.3) adjusted for each evolutionary run (a custom-made script based on the detected
707 and filtered SNPs, avaible at https://github.com/vierocka) using STAR (version 2.5.3a) (17).
708 We call read counts for each known annotated gene using subread packaged (version 1.6.4-
709 Linux-x86_64, tool featureCounts) (18). We normalize the count data using regularized log
710 transformation implemented in R package DESeq2 (19). To evaluate batch effects in sequenced
711 libraries, we use principal component analysis (PCA, R function prcomp). Input for the PCA is
712 a matrix with 30 columns representing sequenced libraries and 4 413 rows representing all
713 detected transcripts. Each cell of the matrix contains a log2-scaled normalized read count. Batch
714 effects can be recognized in the primary libraries as clusters of the PCA components (Fig. S6a).

20

715   To remove these effects, we use the sva R package (function combat with parametric
716   adjustments, 2 iterations) (20). The corrected data shows no statistically significant differences
717   between batches (Fig. S6b); this data is reported in Fig. 3 and Table S4.
718

719   **3. Evolutionary analysis of HGT**

720   **Transfer rates of segments.** In the following, we describe our inference procedure for the
721   segment transfer rate, $u(d)$. Here d is the local donor-recipient sequence divergence in a 100
722   bp window around the recombination start site, which can be at the 5' or 3' terminal end.
723   Because the actual start site for a given transfer segment is unknown, we use a statistical
724   procedure to infer the function $u(d)$. (i) In each of the 595 transfer segments with length > 200
725   bp found across 7 replicate runs, we record the donor-sequence divergence in 100 bp windows
726   at the 5' and 3' terminal ends. This defines an ordered pair $(d_{\min} = \min(d_{5'}, d_{3'}), d_{\max} =$
727    $\max(d_{5'}, d_{3'}))$ for each segment; normalized histograms then yield empirical density
728   distributions $\hat{\rho}_{\min}(d)$ and $\hat{\rho}_{\max}(d)$ for the set of segments (Fig. S3a). We also record a null
729   density $\hat{\rho}_0(d)$ from scrambled 100 bp windows in the core genome. (ii) Guided by previous
730   work(21, 22), we compare this data with a transfer rate model of exponential form,

$$u(d_{rs}|\lambda) = u_0 \, \exp(-\lambda d_{rs}), \tag{S2}$$

735
731   where d_{rs} is the local divergence at the recombination start site, λ is the decay exponent, and
732   u_0 is a global rate. The parameters λ and u_0 are to be inferred from the data. In each segment,
733   recombination can start either at the 5' or at the 3' end; this information is a priori unknown.
734   The rate model (S1) defines a divergence probability density

$$\rho_{rs}(d|\lambda) = C \, u(d|\lambda) \, \rho_0(d) \tag{S3}$$

737   for recombination start sites, where $\rho_0(d)$ is the divergence density in randomly positioned
738   core genome segments and C is a normalization constant. Assuming that the divergence at the
739   recombination end site follows the distribution $\rho_0(d)$, the model predicts density distributions
740   of the minimum and maximum terminal-end distance,

$$\rho_{\min}(d|\lambda) = \rho_{rs}(d|\lambda) \, \Phi_0(d) \tag{S4}$$

$$\rho_{\max}(d|\lambda) = \rho_0(d) \left(1 - \Phi_{rs}(d|\lambda)\right), \tag{S5}$$

743   where $\Phi_0(d)$ and $\Phi_{rs}(d|\lambda)$ are the corresponding cumulative distributions. (iii) We calibrate
744   the decay parameter λ by a combined least-square fit of the model distributions $\rho_{\min}(d|\lambda)$ and
745    $\rho_{\max}(d|\lambda)$ to the empirical distributions $\hat{\rho}_{\min}(d)$ and $\hat{\rho}_{\max}(d)$ (Fig. S3). The null distribution
746   is estimated from its empirical counterpart, $\rho_0(d) = \hat{\rho}_0(d)$. The good fit of both distributions
747   provides an a-posteriori validation of the exponential rate model given by equation (S2). The
748   divergence-dependent rate $u(d) = u(d|\lambda^*)$ with the optimal decay parameter $\lambda^* = 28$ bp is
749   plotted in Fig. S3b. The global rate $u_0 = 4 \times 10^{-6}/(\text{bp x hr})$ is obtained by calibrating
750   equation (S2) with the total number of observed transfers in the competence periods of the
751   experiment.

752   **Null model for HGT of genes and operons.** For each gene g, we obtain a null probability of
753   transfer, $p_0(g)$, by simulation of a positionally scrambled transfer dynamics with rates $u(d)$

754   given by equation (S2) with decay parameter $\lambda^* = 28$ bp. First, we build a record of the relative
755   transfer frequencies by simulations of 10^6 independent single-segment transfers. Each
756   candidate transfer consists of a random starting position in the core genome, a random direction
757   of recombination (i.e., from 5' to 3' end or vice-versa), and a length sampled from the observed
758   length distribution (Fig. 1d). We compute the divergence d_{rs} in the first 100bp from the
759   recombination start site. We then accept the transfer with probability

$$\pi_0(d_{rs}) = \tau_0\, u(d_{rs}), \tag{S6}$$

761   where τ_0 is a normalization factor, provided the segment can be contiguously integrated into
762   the recipient core genome (i.e., it does not overlap with an accessory region). For each accepted
763   transfer, we record which genes are affected. This determines probabilities $p_0(g)$ up to the
764   overall normalization τ_0, which is obtained by matching $\sum p_0(g)$ with the average number of
765   genes hit by transfer observed in the data (Fig. 1f). This null model also describes the multiple-
766   hit statistics of transfer in independently evolving replicate populations. Specifically, the
767   distribution of the transfer frequency θ in r replicate runs is

$$P_0(\theta|p_0) = B(r, \theta r, p_0) \qquad \text{(genes)}, \tag{S7}$$

769   where $B(r, k, p_0)$ is the binomial probability for k hits in a sample of size r. This distribution
770   has mean $\mathrm{E}(\theta) = p_0$ and variance $\mathrm{Var}(\theta) = r^{-1} p_0 (1 - p_0)$.

771   The multiple-hit statistics of transfer for operons goes beyond the single-gene model, because
772   it involves spatial correlations between non-overlapping transfer segments in the same replicate
773   run. Here we use a variant of the simulations, where non-overlapping transfer segments are
774   accumulated until the global genome coverage matches the data. For a given operon (23) or PPI
775   community (24), we simulate transfers in the corresponding genome region plus 10000 bp of
776   flanking sequence. We record the mean transfer frequency of the operon, $\bar{\theta} = n_{op}^{-1} \sum_{g \in op} \theta(g)$,
777   where n_{op} denotes the number of genes in the operon and $\theta(g)$ the transfer frequency of each
778   gene. For a given operon (op), we obtain the distribution of the mean transfer frequency

$$P_0(\bar{\theta}|op) \qquad \text{(operons)} \tag{S8}$$

780   from simulations in batches of $r = 7$ independent replicates. This distribution contains all
781   correlations due to the spatial proximity of genes in the operons. These correlations typically
782   increase the variance of $\bar{\theta}$ compared to the expectation from the single-gene model.

783   This null model serves for the inference of HGT hot spots and cold spots (Fig. S2 and S4), the
784   GO analysis of HGT (Table S3), and the inference of selection on HGT (Fig. 2c, Fig. 4).

785   **Inference of hot spots.** We identify HGT hotspots as genome loci containing an operon with
786   significantly enhanced mean transfer frequency, $\bar{\theta} = n_{op}^{-1} \sum_{g \in op} \theta(g)$, compared to the null
787   model described above. We find two significant hot spots: (i) A locus of 7 genes (chromosome
788   positions 2888940 – 2896972) contains the *leu* operon and has $\bar{\theta} = 0.84$. All of these genes
789   are coherently replaced in at least 4 replicates, three of them in all replicates. (ii) A locus of 15
790   genes (chromosome positions 3514115 – 3529855) contains the two *eps* operons (23) and has
791     $\bar{\theta} = 0.66$. All of these genes are coherently replaced in at least 4 replicates. The statistical
792   significance of these hot spots is evaluated by comparing the observed value of $\bar{\theta}$ with the null
793   distribution (S8) obtained by simulations (Fig. 4e), including a Bonferroni correction for

794   multiple testing in the set of 141 operons with ≥ 5 genes. We obtain $P < 3 \times 10^{-4}$ for the *leu*
795   locus and $P < 2 \times 10^{-3}$ for the *eps* locus.

796   **Inference of cold spots.** Cold spots of HGT are genome loci with no observed transfers in any
797   of the replicate runs. We identify statistically significant cold spots from the statistics of gaps
798   between consecutive transfers under the null model of independent transfers. First, in each
799   replicate, the null distribution of gaps is exponential,

806
$$P_0^{(1)}(\delta) = C\, e^{-\nu_1 \delta} \qquad \text{(gaps)}, \qquad\qquad\qquad \text{(S9)}$$

800   where C is a normalization constant. The empirical gap distribution recorded in the seven
801   replicate runs is consistent with this form with $\nu_1 = 5.22 \times 10^{-5}$/bp for $\delta > 5000$ (Fig. S4a);
802   deviations for short gaps can be attributed to short-range correlations between non-overlapping
803   transfers, constraints imposed by the accessory regions of the genome, and statistical noise in
804   the segment mapping. Next, the distribution of common gaps common in r independent
805   replicates is also exponential,

807
$$P_0^{(r)}(\delta) = rC\, e^{-r\nu_1 \delta} \qquad \text{(gaps)}. \qquad\qquad\qquad \text{(S10)}$$

808   The observed distribution of common gaps in 7 replicate runs is asymptotically exponential,
809    $\hat{Q}^{(7)}(\delta) = C_7\, e^{-\nu_7 \delta}$ with an exponent $\nu_7 = 1.88 \times 10^{-4} < 7\nu_1$ for $\delta > 5000$ bp (Fig. S4b).
810   This inequality reveals weak correlations between replicates, which are in tune with the
811   genome-wide selection model discussed below. We find two significant cold spots as outliers
812   with respect to $\hat{Q}^{(7)}(\delta)$. The first cold spot has a length of 48815bp (chromosome positions
813   332356 – 381171) and contains 43 genes; the second has a length of 55895bp (chromosome
814   positions 1823905 – 1879800) and contains 28 genes, including the ultra-long genes *pksN*
815   (16467bp) and *pksR* (7632bp). We evaluate the significance of the cold spots from the Gumbel
816   distribution (parameters 33839, 5409) obtained from $\hat{Q}^{(7)}(\delta)$ for 477 draws, corresponding to
817   the number of gaps common to all replicates. We obtain $P < 0.041$ for the first and $P < 0.01$
818   for the second cold spot.

819   **GO analysis.** We study the association of HGT with cellular functions, using GO categories
820   obtained from(25). For each category containing at least 10 genes, we evaluate the mean
821   transfer frequency $\bar{\theta} = n_{\text{cat}}^{-1} \sum_{g \in \text{cat}} \theta(g)$. We compare this quantity with its expectation value
822   under the null model, $E(\bar{\theta}) = \bar{p}_0 = n_{\text{cat}}^{-1} \sum_{g \in \text{cat}} p_0(g)$, by evaluating the ratio $\varphi = \bar{\theta}/\bar{p}_0$ (Table
823   S3). Because our null model incorporates variations in local sequence similarity, we discount
824   correlations between GO categories and HGT generated by their mutual dependence on
825   sequence similarity. To evaluate the statistical significance of over- or underrepresentation of
826   HGT in a given category, we compare the observed number of transfer hits in 7 replicates,
827    $t_{\text{cat}} = 7\, n_{\text{cat}}\, \bar{\theta}$, with the expected number, $E(t_{\text{cat}}) = 7\, n_{\text{cat}}\, \bar{p}_0$, using a standard Binomial test
828   with a Bonferroni multiple-testing correction in the set of GO categories (Table S3).

829   **Selection model for HGT of genes.** Selection affects the probability to observe the transfer of
830   an individual gene (g),

831
$$p(g) = p_0(g)\, \varphi\big(2N_e s(g)\big). \qquad\qquad\qquad \text{(S11)}$$

832   Here $p_0(g)$ is the transfer probability under the null model; φ is the ratio of fixation
833   probabilities under the full model and under the null model for genetic variants with

<div align="center">23</div>

834 orthologous replacement of the gene. The function $\varphi\big(2N_e s(g)\big)$ depends on a gene-specific

835 selection coefficient of transfer, $s(g)$, and the effective population size N_e underlying the

836 population dynamics in the evolution experiments (26). For the inference of genome-wide

837 selection acting on HGT, we determine $p_0(g)$ by simulations, as described above. Next, we

838 partition the genome into two classes with uniform selection in each class, i.e., $\varphi(g) = \varphi_+$ in

839 the (+) class and $\varphi(g) = \varphi_-$ in the (-) class. This assumption is expected to dilute the gene-

840 specific signal of selection, but is necessary to avoid overfitting of the selection model

841 parameters from the data. The two-class model has a transfer probability

$$p\big(c, \varphi_+, \varphi_-, p_0\big) = c\,\varphi_+\,p_0 + (1-c)\,\varphi_-\,p_0\,, \tag{S12}$$

843 where the mixture parameter c measures the fraction of core genes in the (+) class. The resulting

844 distribution of the transfer frequency θ in r replicate runs,

$$Q\big(\theta|c, \varphi_+, \varphi_-, p_0\big) = c\,B\big(r, \theta r, \varphi_+ p_0\big) + (1-c)\,B\big(r, \theta r, \varphi_- p_0\big), \tag{S13}$$

846 is to be compared with the corresponding probability under the null model, which is given by

847 equation (S7).

848 **Bayesian inference of selection on HGT.** We infer the mixture model parameters c, φ_+, φ_-

849 using the log likelihood score

$$S(c, \varphi_+, \varphi_-) = \sum_g \log \frac{Q\big(\theta(g)|c, \varphi_+, \varphi_-, p_0(g)\big)}{P_0(\theta(g)|p_0(g))} \tag{S14}$$

851 with probability distributions Q and P_0 given by equations (S7) and (S13). Evaluating this score

852 function in our data set with flat prior probabilities of $(c, \varphi_+, \varphi_-)$ leads to the posterior score

853 distribution shown in Fig. S6a. Varying the parameter c, we evaluate the conditional score

854 maximum $S^*(c)$ (Fig. S5b) and the corresponding parameters $\varphi_+^*(c), \varphi_-^*(c)$ (Fig. S5c). This

855 determines the absolute maximum-likelihood score $S^* = \max_q S^*(c) = 47$ and the maximum-

856 likelihood   parameters $(c^*, \varphi_+^*, \varphi_-^*) = \arg\max_{c,\varphi_+,\varphi_-} S\big(c, \varphi_+, \varphi_-\big) = (0.2, 1.9, 0.75)$.   We

857 estimate the parameter regime of high posterior probability by integration of the function $S^*(c)$.

858 Using conservative confidence thresholds (90% for $c < c^*$, 50% for $c > c^*$), we obtain

$$c \in [0.1, 0.4],\ \varphi_+ \in [1.6, 2.4],\ \varphi_- \in [0.6, 0.84]. \tag{S15}$$

860 The expected fraction of (+) genes can also be evaluated for each transfer class. Specifically, in

861 the high-transfer (low-transfer) class defined in the main text, 88% (20%) of the genes are

862 inferred to be under positive selection. This calculus also yields an estimate of the fraction of

863 observed transfers that are under positive selection,

$$\eta = \frac{\varphi_+ c}{\varphi_+ c + \varphi_-(1-c)}. \tag{S16}$$

864

865 We obtain $\eta = 0.40\ [0.20, 0.60]$; the c-dependent conditional fraction $\eta^*(c)$ is shown in Fig.

866 S5d. To estimate the statistical significance of this inference, we rewrite the score (S14) in the

867 form

$$S = n\,H\big(Q_2\,|\,P_{0,2}\big), \tag{S17}$$

24

870 where n is the number of core genes and $H(Q_2 \mid P_{0,2})$ is the relative entropy (Kullback-Leibler
871 distance) between the joint distributions $Q_2(\theta, p_0) = Q(\theta|p_0) P(p_0)$ and $P_{0,2}(\theta, p_0) =$
872 $P_0(\theta|p_0) P(p_0)$ (here we omit the parameters c, φ_+, φ_- from the notation). At the maximum-
873 likelihood point, Sanov's theorem then gives the probability of the observed data under the null
874 model, $P < \exp(-S^*) \sim 10^{-20}$.

875 To display the deviations of the transfer statistics from the neutral null model, Fig. 3b shows
876 the empirical likelihood ratio $\hat{Q}(\theta|p_0)/P_0(\theta|p_0)$ in four bins of p_0. In Fig. 2c, we plot the
877 likelihood ratio $Q^*(\theta)/P_0(\theta)$, where $Q^*(\theta) = \int Q_2^*(\theta, p_0) \, dp_0$ is the marginal distribution of
878 the maximum-likelihood selection model and $P_0(\theta) = \int P_{0,2}(\theta, p_0) \, dp_0$ is the marginal
879 distribution of the null model. Comparison with the corresponding empirical ratio
880 $\hat{Q}(\theta)/P_0(\theta)$ shows that the maximum-likelihood selection model provides a good fit to the
881 data, whereas the null model does not. We emphasize that our inference of selection and
882 significance estimate are based on the full joint data (θ, p_0), in order to properly discount local
883 variations of p_0.

884 Importantly, the null model given by equation (S6) takes into account local variations of the
885 sequence divergence between donor and recipient, avoiding a spurious signal of selection from
886 this variation. Our inference of positive and negative selection is based on the assumption that
887 this null model is a faithful representation of the neutral transfer dynamics. However, even if
888 the null model contains some uniform selection on HGT, our inference still provides evidence
889 of two gene classes with selective differences ($\varphi_+ > \varphi_-$).

890 **HGT in operons.** Our analysis of HGT in gene networks uses a list of operons in *B. subtilis*
891 retrieved from (23). To infer selection on HGT for a given operon (*op*), we compare the
892 observed mean transfer frequency of the network genes, $\bar{\theta} = n_{op}^{-1} \sum_{g \in op} \theta(g)$, with the neutral
893 null distribution $P_0(\bar{\theta}|op)$ given by equation (S8).

894 For the aggregate inference of negative selection on operons (Fig. 4a), we use the set of all *B.*
895 *subtilis* operons in the core genome region that contain at least 7 genes; the threshold is chosen
896 to ensure any signal of selection can be considered a network property (i.e., is not dominated
897 by single genes). This set contains 56 operons with an average of 10 genes per operon.
898 Compared to the neutral null model, the data shows overrepresentation of transfers at low
899 frequency, 46% of counts of $\bar{\theta} < 0.1$ in the data vs. 31% in the null model. These deviations
900 are statistically significant, signaling selection against HGT at the network level ($<$
901 10^{-3}, computed again by Sanov's theorem from the histogram of data and the null
902 distribution). We have excluded the hotspots for recombination (*leu* and *eps* operons), which
903 are analyzed separately for positive selection.

904 The individual operons *leu* and *eps* differ significantly from the null model ($P < 3 \times 10^{-4}$ for
905 *leu*, $P < 2 \times 10^{-3}$ for *eps*; see above). The additional examples of high- and low-transfer
906 operons (Fig. 4d, Fig. S7) are marginally significant at the individual level ($P < 0.05$ without
907 stringent Bonferroni correction). The list of genes in these operons is given in Table S5.

908 **HGT dynamics in a cross-lineage fitness landscape.** Intra-network selection on HGT can be
909 described by a minimal cross-lineage fitness landscape $F(q|a, b)$ that depends on the fraction

910 of transferred network genes, q, as given by equation [2]. Simulations of the HGT dynamics in
911 the landscape $F(q|a, b)$ use a transfer probability for segments,

$$\pi(d_{rs}, q, q'|a, b) = \pi_0(d_{rs}) \, \varphi\big(2N_e \, s(q, q'|a, b)\big), \qquad (S18)$$

913 where q, q' are the network transfer fractions before and after the event in question, and
914 $\pi_0(d_{rs})$ is the neutral transfer probability given by equation (S6). The fixation probability
915 φ depends on the effective population size N_e and on the selection coefficient $s(q, q'|a, b) =$
916 $F(q'|a, b) - F(q|a, b)$; cf. equation (S11). These dynamics determine distributions of the
917 mean transfer frequency $Q(\bar{\theta}|a, b)$ in the fitness landscape $F(q|a, b)$, which is to be compared
918 with the corresponding null distribution $P_0(\bar{\theta})$ given by equation (S8). The model distributions
919 $Q(\bar{\theta}|a, b)$ reported in Fig. 4a are obtained from the fitness model of equation [2] with scaled
920 parameters $2N_e a = -0.4$, $2N_e b = 0$ (negative directional selection, orange) and $2N_e a = 0$,
921 $2N_e b = -2.2$ (hybrid incompatibilities, blue).

922 The distributions of Fig. 4d (*yes* and *iol*) have parameters $2N_e b = 0$ and $2N_e a =$
923 -10 (negative directional selection). The distributions of Fig. 4e and Fig. S7 have parameters
924 $2N_e a = 5.4$ (*leu*), $2N_e a = 3.5$ (*eps*), $2N_e a = 1.5$ (*bkd*), $2N_e a = 1.4$ (*xse*), and $2N_e b = 0$
925 (positive directional selection, magenta). These selection coefficients are calibrated such that
926 the model distributions $Q(\bar{\theta}|a, b)$ have the same average of $\bar{\theta}$ as the histogram of operon data.
927 The analysis shows that the observed HGT in operons is broadly consistent with selection given
928 by a fitness landscape of the form of equation [2]. However, more data is needed to calibrate
929 mixture models with maximum-likelihood parameters delineating directional and epistatic
930 selection.

931

932 **Data availability.** Data is available in Table S1 (de novo mutations), Table S2 (genomic
933 positions of orthologous recombinations, insertions from accessory genome of donor, deletions,
934 duplications), Table S3 (GO categories), Table S4 (log$_2$ fold changes ΔR of mRNA levels),
935 Table S5 (operons with constituent genes). All custom-made scripts will be available at
936 https://github.com/vierocka.

937

938
939
940 1.   M. Albano, J. Hahn, D. Dubnau, Expression of competence genes in Bacillus subtilis. *J*
941   *Bacteriol* **169**, 3110-3117 (1987).
942 2.   B. Maier, "Competence and Transformation" in Bacillus: Molecular and Cellular
943   Biology, P. Graumann, Ed. (Caister Academic Press, 2017), pp. 395-414.
944 3.   H. Maamar, D. Dubnau, Bistability in the Bacillus subtilis K-state (competence) system
945   requires a positive feedback loop. *Mol Microbiol* **56**, 615-624 (2005).
946 4.   M. Yuksel, J. J. Power, J. Ribbe, T. Volkmann, B. Maier, Fitness Trade-Offs in
947   Competence Differentiation of Bacillus subtilis. *Front Microbiol* **7** (2016).
948 5.   W. A. Rosche, P. L. Foster, Determining mutation rates in bacterial populations.
949   *Methods* **20**, 4-17 (2000).
950 6.   S. Sarkar, W. T. Ma, G. H. Sandri, On fluctuation analysis: a new, simple and efficient
951   method for computing the expected number of mutants. *Genetica* **85**, 173-179 (1992).

952   7.   M. F. Wojciechowski, M. A. Hoelzer, R. E. Michod, DNA repair and the evolution of
953   transformation in Bacillus subtilis. II. Role of inducible repair. *Genetics* **121**, 411-422
954   (1989).
955   8.   A. M. Bolger, M. Lohse, B. Usadel, Trimmomatic: a flexible trimmer for Illumina
956   sequence data. *Bioinformatics* **30**, 2114-2120 (2014).
957   9.   S. Andrew (2010) Fastqc: A quality control tool for high throughput sequence data.
958   (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/).
959   10.   H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler
960   transform. *Bioinformatics* **25**, 1754-1760 (2009).
961   11.   A. McKenna *et al.*, The Genome Analysis Toolkit: A MapReduce framework for
962   analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297-1303 (2010).
963   12.   H. Li *et al.*, The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**,
964   2078-2079 (2009).
965   13.   A. R. Quinlan, I. M. Hall, BEDTools: a flexible suite of utilities for comparing genomic
966   features. *Bioinformatics* **26**, 841-842 (2010).
967   14.   S. Bubendorfer *et al.*, Genome-wide analysis of chromosomal import patterns after
968   natural transformation of Helicobacter pylori. *Nature communications* **7**, 1-12 (2016).
969   15.   D. R. Zeigler, The genome sequence of Bacillus subtilis subsp. spizizenii W23: insights
970   into speciation within the B. subtilis complex and into the history of B. subtilis genetics.
971   *Microbiol-Sgm* **157**, 2033-2041 (2011).
972   16.   M. Hummel, S. Bonnin, E. Lowy, G. Roma, TEQC: an R package for quality control in
973   target capture experiments. *Bioinformatics* **27**, 1316-1317 (2011).
974   17.   A. Dobin *et al.*, STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21
975   (2013).
976   18.   Y. Liao, G. K. Smyth, W. Shi, featureCounts: an efficient general purpose program for
977   assigning sequence reads to genomic features. *Bioinformatics* **30**, 923-930 (2014).
978   19.   M. I. Love, W. Huber, S. Anders, Moderated estimation of fold change and dispersion
979   for RNA-seq data with DESeq2. *Genome Biol* **15** (2014).
980   20.   J. T. Leek *et al.*, sva: Surrogate Variable Analysis. R package version 3.32.1.
981   *Bioconductor* (2019).
982   21.   M. Vulic, F. Dionisio, F. Taddei, M. Radman, Molecular keys to speciation: DNA
983   polymorphism and the control of genetic exchange in enterobacteria. *Proc Natl Acad
984   Sci U S A* **94**, 9763-9767 (1997).
985   22.   P. H. Oliveira, M. Touchon, E. P. C. Rocha, Regulation of genetic flux between bacteria
986   by restriction-modification systems. *Proc Natl Acad Sci U S A* **113**, 5658-5663 (2016).
987   23.   B. Taboada, R. Ciria, C. E. Martinez-Guerrero, E. Merino, ProOpDB: Prokaryotic
988   Operon DataBase. *Nucleic Acids Res* **40**, D627-D631 (2012).
989   24.   D. Szklarczyk *et al.*, STRING v11: protein-protein association networks with increased
990   coverage, supporting functional discovery in genome-wide experimental datasets.
991   *Nucleic Acids Res* **47**, D607-D613 (2019).
992   25.   B. Y. Zhu, J. Stulke, SubtiWiki in 2018: from genes and proteins to functional network
993   annotation of the model organism Bacillus subtilis. *Nucleic Acids Res* **46**, D743-D748
994   (2018).
995   26.   J. F. Crow, M. Kimura, An introduction to population genetics theory. *An introduction
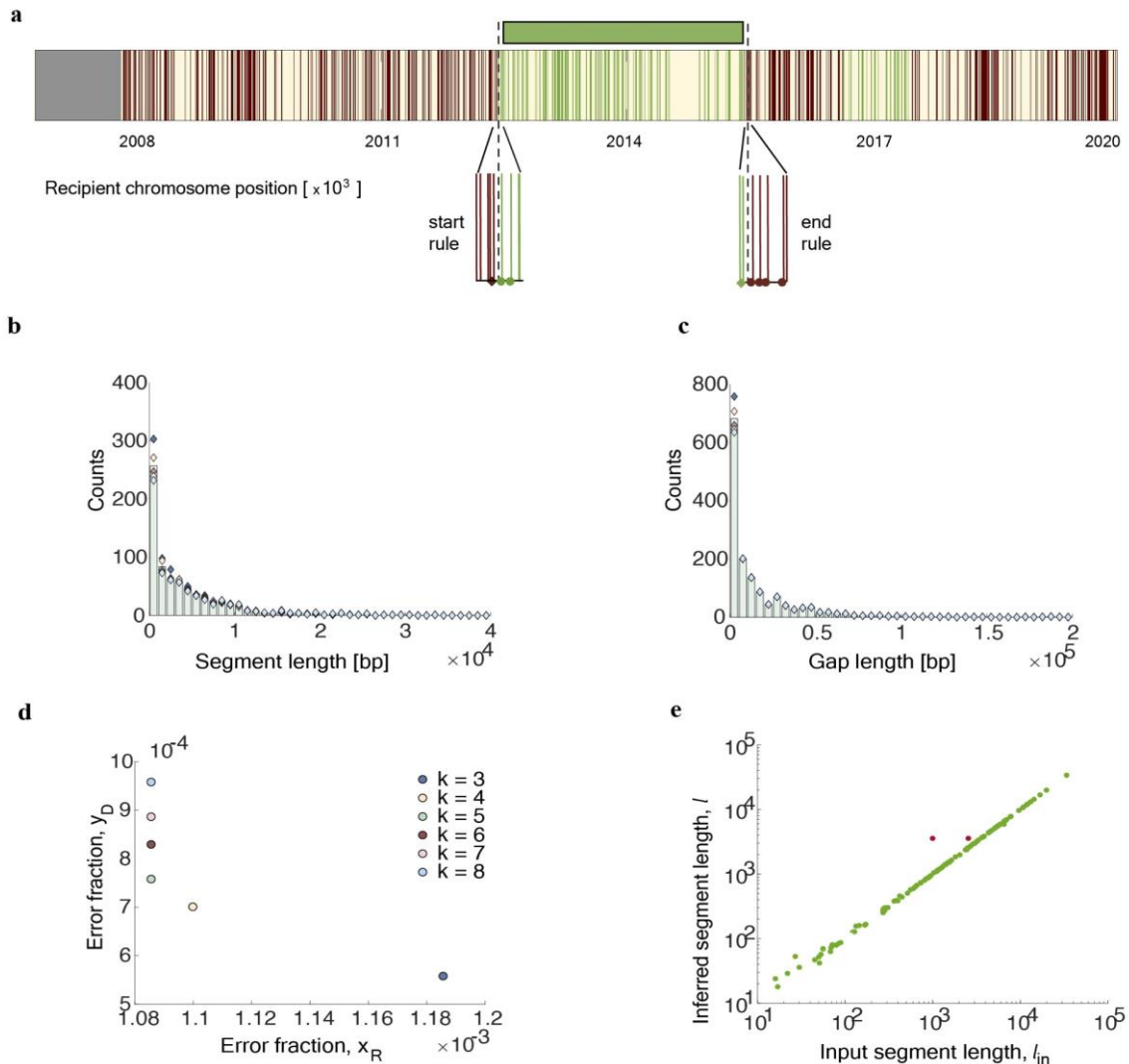996   to population genetics theory.* (1970).

997
998
999
1000
1001

1002
1003
1004
1005
1006
1007
1008
1009
1010

Supporting Figures and Tables

Fig. S1. Inference of orthologous recombination. (a) Algorithm. The recipient genome is partitioned into core (yellow) and accessory (grey) segments. Core sequence sites with donor-recipient divergence are marked by vertical lines. The evolved sequence has recipient (R, red) or donor (D, green) alleles at each of these sites (data shown from run R1, genomic coordinates 2006770 – 2020000). The start of each segment is determined by a 5' marker (two consecutive D alleles, green dots); the end is determined by a 3' marker ($k = 5$ consecutive R alleles, red dots). These markers and their left flanking sites (diamonds) set the start/end coordinates of inferred transfer segments (green bars) by a midpoint rule. See SI Appendix for details. **(b,c)** Length distribution of inferred transfer segments and of inter-segment gaps for different values of the parameter k. **(d)** Fraction of R alleles that fall within transfer segments, x_R, and fraction of D alleles that fall outside inferred transfer segments, y_D, for different values of the parameter k. **(e)** Test of the algorithm on simulated data. Scatter plot of the inferred length vs. the input length for a set of 117 scrambled transfer segments. This set produces 115 inferred segments uniquely mapped to an input segment (green dots) and one inferred segment containing two close input segments (red dots). The average inference error is $\Delta l = 20$ bp. See section 2 of SI Appendix for details.
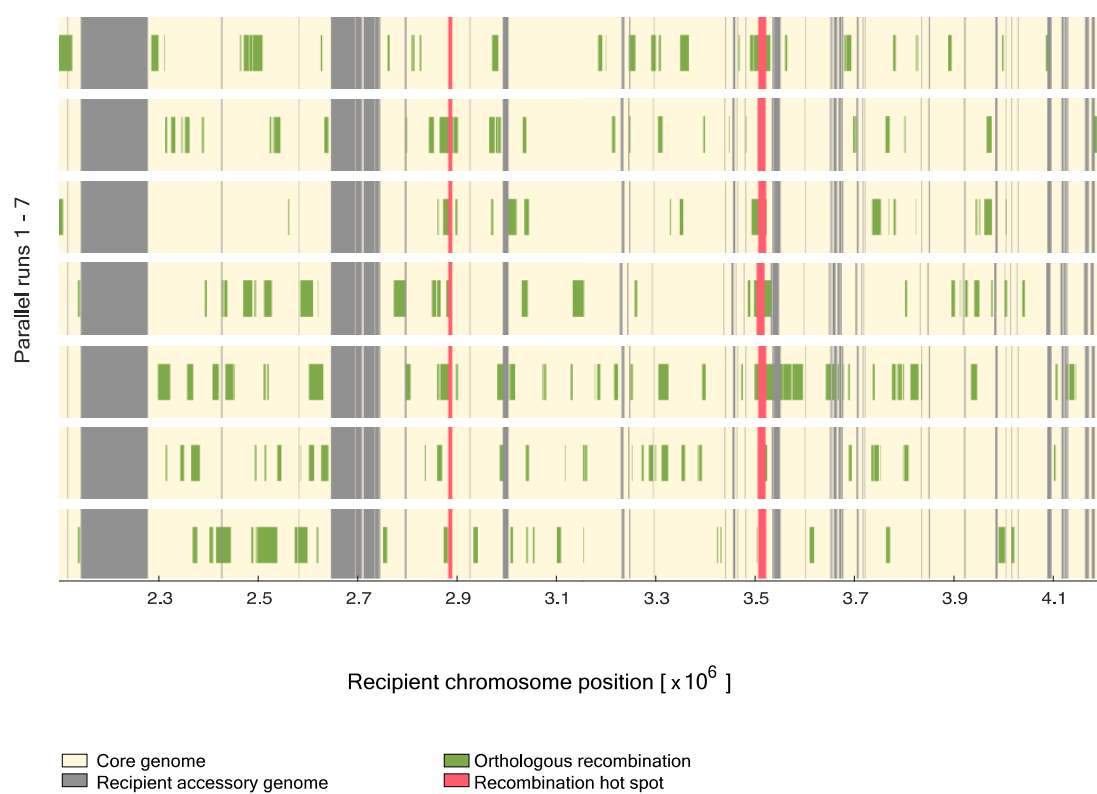
1032
1033
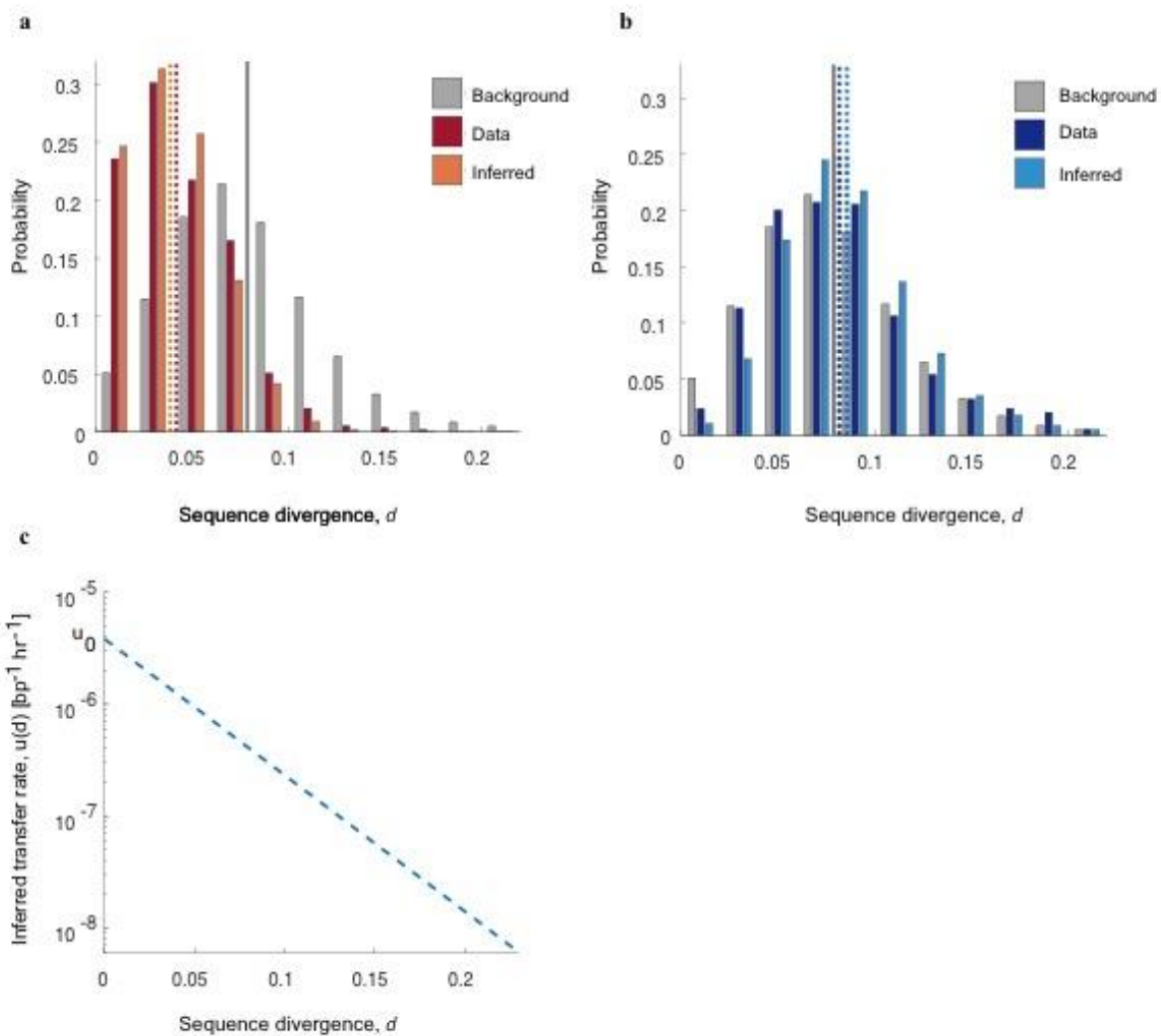1034
1035 **Fig. S2. Genomics of HGT across replicate runs.** The aligned HGT pattern in all 7 parallel runs is
1036 shown for a part of the genome (annotation as in Fig. 1b).
1037
1038

1039
1040
1041

1042
1043
1044 **Fig. S3. Transfer rate depends on local sequence divergence. (a)** Observed sequence divergence d in
1045 100bp terminal regions of transferred genome segments: distribution $\hat{\rho}_{min}(d)$ recorded from the low-
1046 divergence end within each segment (left panel, red) and **(b)** distribution $\hat{\rho}_{max}(d)$ recorded from high-
1047 divergence ends (right panel, blue). These data are shown together with the background distribution
1048 $\hat{\rho}_0(d)$ obtained from scrambled 100 bp windows (gray) and with the distributions $\rho_{min}(d)$ and
1049 $\rho_{max}(d)$ computed from the optimal local-divergence model. Dashed lines mark the mean values of the
1050 corresponding distributions. **(c)** Inferred transfer rate depending on the local sequence divergence at the
1051 recombination start site, $u(d) = u_0 \exp(-\lambda d)$ with $\lambda = 28$ bp and $u_0 = 4 \times 10^{-6}/(\text{bp x hr})$. See
1052 section 3 of SI Appendix for definitions of these distributions and for details of the inference procedure.
1053
1054
1055
1056

a

b
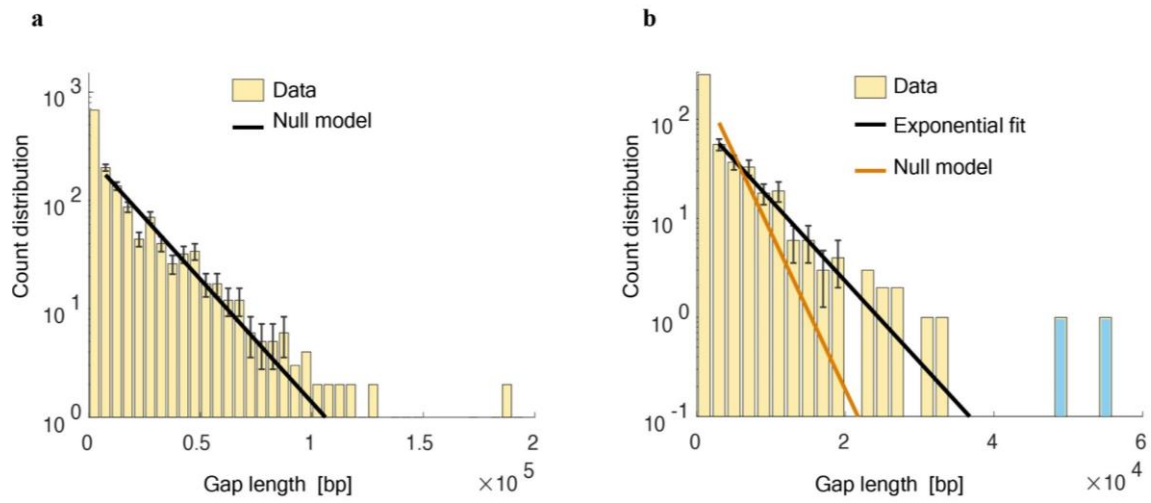
Fig. S4. Inference of HGT cold spots. (a) Distribution of the gap length δ between subsequent transfers in single runs with exponential fit, $P_0^{(1)}(\delta) \sim \exp(-\nu_1 \delta)$ for $\delta > 3000$ bp (black line). **(b)** Length distribution of common gaps in all 7 runs together with exponential fit, $Q^{(7)}(\delta) \sim \exp(-\nu_7 \delta)$ (black line), shown together with null expectation for independent replica, $P_0^{(7)}(\delta) \sim \exp(-7\nu_1 \delta)$ (orange line). Two cold spots of length $l = 48815$ bp and $l = 55895$ bp (blue) are identified as statistically significant outliers to the observed exponential pattern ($P < 0.04$). See section 3 of SI Appendix for the statistical analysis.

1069
1070
1071
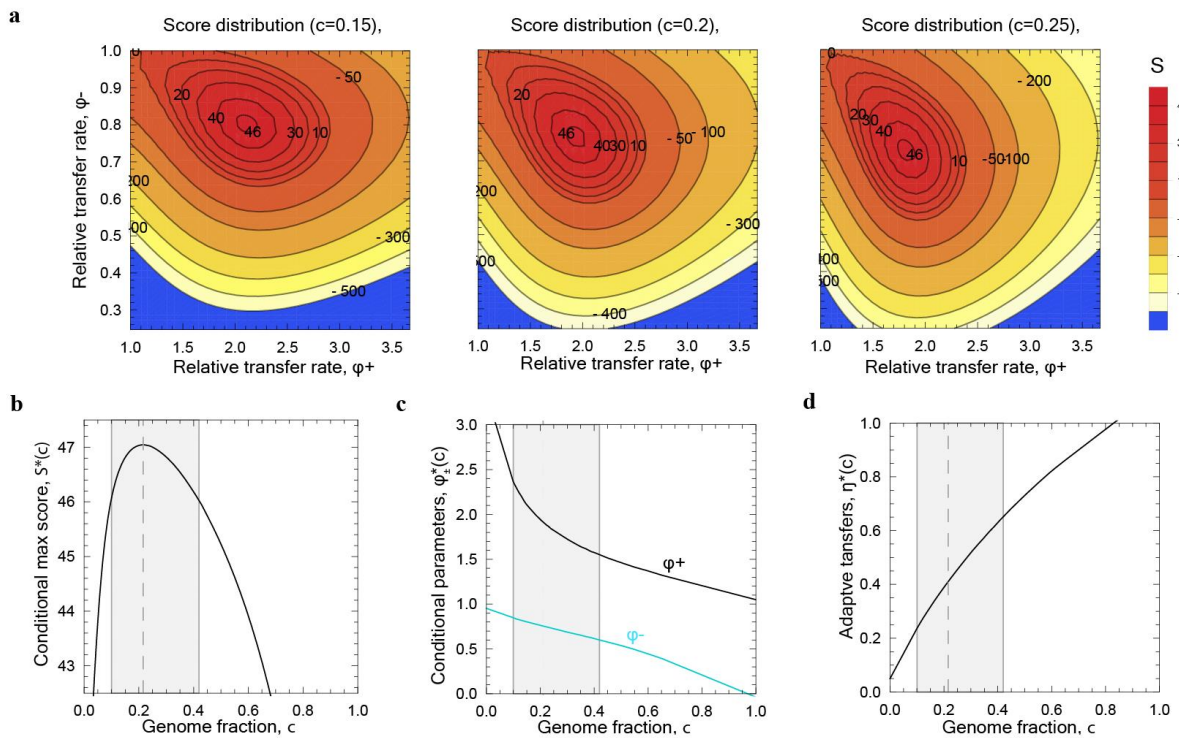
Fig S5. Inference of the selection model for transfer. (**a**) Log likelihood score of the selection model compared to the neutral null model as a function of the relative transfer rates $\varphi_+ = p_+/p_0$ and $\varphi_- = p_-/p_0$ for given genome fractions $c = 0.15, 0.25, 0.35$. (**b**) Conditional maximum log likelihood score $S^*(c)$. The c interval of high posterior probability is indicated by shading. (**c**) Conditional maximum-likelihood parameters $\varphi_+^*(c)$ and $\varphi_-^*(c)$. (**d**) Partitioning of HGT by selection class. Conditional maximum likelihood function $\eta^*(c)$, where $\eta = \varphi_+ c/(\varphi_+ c + \varphi_-(1 - c))$ is the fraction of the observed transfers predicted to be under positive selection. See section 3 of SI Appendix for details of the inference procedure.
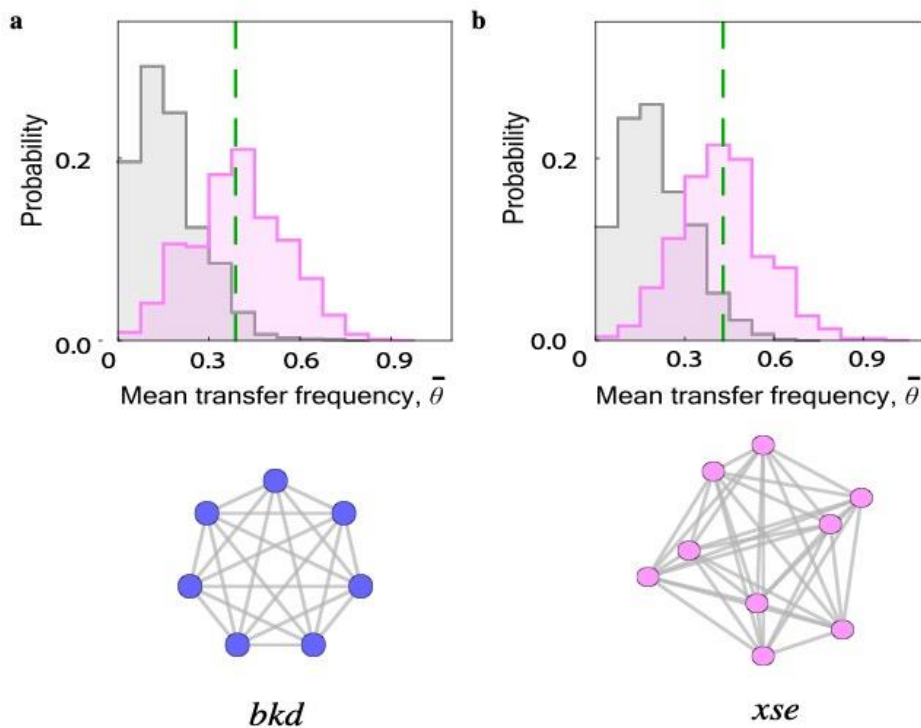
1083
1084
1085
1086
1087

1088
1089
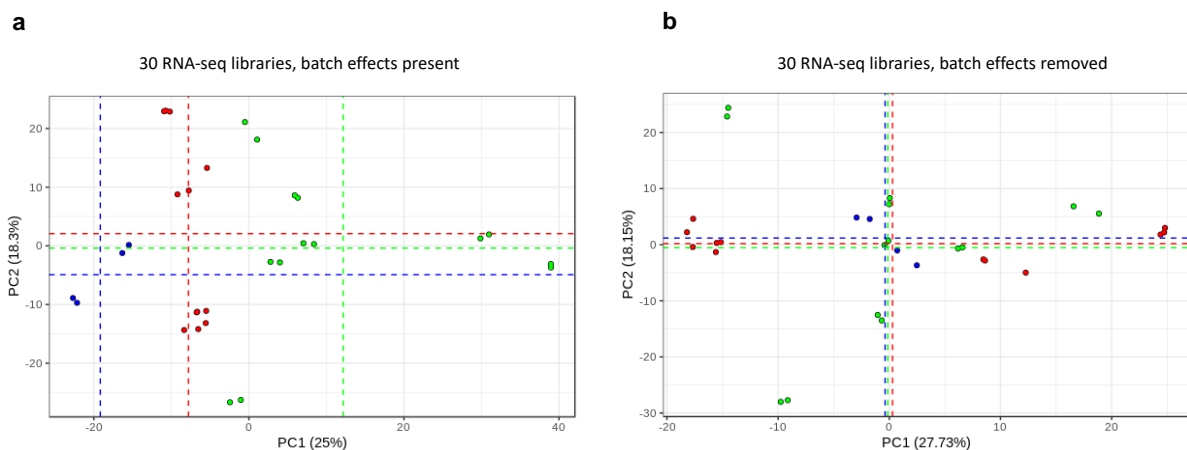
1090
1091 **Fig. S6. Batch correction of transcriptomics data. (a)** 30 RNAseq libraries from 7 evolved
1092 populations, 3 control runs, and the ancestral recipient were obtained in three batches (red, blue, green).
1093 We perform a principal component analysis (PCA) of the $\log_2$ -scaled normalized read counts of these
1094 libraries; batch averages of the first two PC components are shown as dashed lines. Batch effects can be
1095 inferred primarily in the component PC1. **(b)** After batch correction (SI Appendix), the data show no
1096 longer statistically significant differences between batches.
1097
1098
1099

1100
1101
1102
1103
1104

1105
1106
1107
1108 **Fig S7. Additional operons with enhanced HGT**. **(a)** *bkd* operon, with 7 genes and $\bar{\bar{\theta}} = 0.39$. **(b)** *xse*
1109 operon with 9 genes and $\bar{\bar{\theta}} = 0.43$. The observed transfer frequency $\bar{\bar{\theta}}$ (green, dashed line) is
1110 substantially enhanced in comparison to the neutral null model (gray) and consistent with positive
1111 directional selection ($a > 0$, magenta). Both operons are associated to metabolism of amino acids and
1112 small compounds (24). The corresponding PPI networks are shown below.

1113
1114
1115
1116
1117
1118
1119
1120
1121
1122
1123
1124
1125
1126
1127
1128
1129
1130
1131