File S1: Supplemental Materials and Methods

DNA donor strains

The following 11 strains were used as DNA donors in the experimental evolution experiments:

Bacillus subtilis strain RO-FF-1 (described in Stefanic *et al*. 2012 and in Cohan et al. 1991)

Bacillus subtilis strain RS-D-2 (described in Stefanic *et al*. 2012 and in Cohan et al. 1991)

Bacillus spizizenii strain RO-E-2 (Originally described in Stefanic *et al*. 2012 and in Cohan et al. 1991, and further classified as a separate species in Dunlap et al. 2020)

Bacillus mojavensis strain RO-H-1 (described in Stefanic *et al*. 2012 and in Cohan et al. 1991)

Halobacillus halophilus strain DSM2266 (ATCC # 35676)

Halomonas elongate strain DSM2581 (ATCC # 33173)

Aliivibrio fischeri strain ES114 (ATCC# 700601)

Haloarcula marismortui strain DSM3752 (ATCC# 43049)

Haloferax dentrificans strain DSM4425 (ATCC# 35960)

Haloferax mediterranei strain R-4 (ATCC# 33500)

Haloferax volcanii strain WFD11 (described in Charlebois *et al.* 1987)

Whole-genome DNA sequencing

DNA sequencing was performed as previously described (Blecher-Gonen et al. 2013) with the following modifications: ~2000ng of genomic DNA was sheared using the Vovaris E220X sonicator (Covaris). 1000ng of sheared DNA was used for library preparation as follows: End repair was performed at 20°C for 30 min and then purified using 0.75X Agencourt Ampure XP beads cleanup (Beckman Coulter cat# A63880). A-bases were added to both 3’ ends and the product was purified with 2.2X Ampure XP beads followed by

adapter ligation (25°C for 15 min). The ligation product was purified with 0.75X Ampure XP beads. Libraries were quantified by Qubit and qPCR analysis using Illumina primers. Libraries were pooled and sequenced on 1 lane of Hiseq2500 V4 (Illumina) using the paired end 125bp kit.

Bioinformatics pipeline for identification of foreign DNA fragments and mutations

Identification of HGT acquired fragments as well as point mutation was based on integrating two bioinformatics approaches: The first analysis included variant calling against the reference genome of *Bacillus subtilis* 168. The second analysis was based on mapping the reads from each evolved population to a reference sequence, combining all the relevant donors of each evolutionary treatment and the reference genome of *Bacillus subtilis* 168. In cases in which no reference genome was available, we sequenced the relevant donor and used the sequencing result for genome assembly (see details in section "Assembly of donor genomes"). The two analyses and integration process are detailed bellow.

- Variant calling

Reads were trimmed using cutadapt (Martin 2011) and then aligned to the *Bacillus subtilis* strain 168 genome (build GCA_000009045.1, downloaded from Ensembl) using bwa mem (Li and Durbin 2010) (v0.7.15, with -a and -M tags). The alignment files were sorted, duplicates were marked and secondary (/supplementary) alignments were filtered out using sambamba (v0.6.0) (Tarasov et al. 2015). Variants were called using mpileup (minimum mapping quality = 1, minimum base quality = 15) and VarScan2 (Koboldt et al. 2012) (–min_coverage=10, –min-var-freq=0.02, –somatic-p-value=0.05, –strand-filter=1, –min-tumor-freq=0.02 for SNPs and 0.04 for indels, –max-normal-freq=0.02 for SNPs and 0.005 for indels).

- Mapping reads to donors and recipient genomes

A compound genome was composed that included the *Bacillus subtilis* strain 168 reference genome (GCA_000009045.1) and the genomes of all donor species/strains relevant to each experiment (see below the list of accession numbers). For donors with no sequenced genome, we used genome assemblies generated in this work. The reads were aligned against the compound genome using bwa mem with higher penalties on mismatches and gap openings (-B 10 -O 10). Alignments with the highest alignment score were retained, and reads that aligned with the same alignment score to the *Bacillus subtilis* 168 reference genome were discarded, using a custom pysam script. Additional removed reads include: duplicate reads, reads from a secondary alignment, and reads with alignment quality lower than 10 (processed using sambamba). Reads with zero mapping quality were processed in parallel, to account for cases where an HGT segment could arise from more than one donor. Reads that mapped to the donor genomes (coverage>2) were used to define putative HGT regions. Putative regions from all samples of the same lineage were merged together in order to define consistent coordinates.

The number of reads on the putative donor regions was counted using bedtools (Quinlan and Hall 2010). The maximal read count across samples was compared to that of the ancestor, and high-quality donor regions were defined as having:

$$log_2\left(\frac{\max(counts)+2}{count_{ancestor}+2}\right) > 3\ .$$

- Integration of the two methods

The sequences of the high-quality donor regions, defined by the mapping to the compound genome (second analysis, see "mapping reads to donors and recipient genomes"), were extracted from the relevant donor genomes. These sequences were mapped against the *Bacillus subtilis* 168 genome (GCA_000009045.1) and variants were called. The resulting variants were compared with those called directly from the reads (first analysis, see "variant calling"). The variants found by both methods were defined as variants that are due to HGT.

The first and last shared variants in each region were used to define the start and end coordinates of the transferred region. Variants that were called directly from the reads and that did not appear in the alignments of the donor regions were considered *de novo* mutations/indels that did not come from a foreign DNA fragment.

In order to calculate the frequency of each genomic alteration event, we used the frequency of each variant, calculated by the variant caller. For foreign DNA fragments, the median frequency of all variants within a fragment was calculated and used as the foreign DNA fragment frequency.

In some cases, the pipeline considered overlapping or adjacent foreign fragments with similar frequencies and donor identity as two separate fragments. Following manual inspection, such fragments were merged, as long as the distance between them did not exceed 1kb and no other variants were expected to be found between them.

Other cases of overlapping foreign fragments that shared the same frequency, but were aligned to different donors, were also detected. These cases usually included one large fragment and smaller overlapping fragments of different donor identities. In order to correct for possible misidentification of donor identity, the fragments were compared by BLAST against each other and the *Bacillus subtilis* 168 genome and variants were classified according to their donor identity. In cases where the differences in donor identity between the overlapping fragments resulted from high similarity between donors in that region, fragments were merged according to the identity of the largest fragment.

The final list of genomic events in the different populations (i.e., point mutations, indels and foreign DNA fragments) that reached a frequency of at least 10% in at least one of the time points can be found in Table S2 and Table S3.

- Classification of replacement and duplication

In order to examine whether fragments replaced an existing region in the recipient chromosome or rather were integrated in an ectopic region, deep-sequencing signatures that characterize duplications, i.e., higher coverage and split reads overlapping non-adjacent genomic regions were assessed. Each foreign DNA fragment identified was manually inspected using the Integrative Genomics viewer (IGV) tool (Robinson et al. 2011), for changes in coverage or high abundance of split reads.

- Accession numbers of donor genomes and plasmids

Bacillus subtilis strain RO-FF-1	JACJGC000000000
Bacillus subtilis strain RS-D-2	JACJGD000000000
Bacillus spizizenii strain RO-E-2	JACJGE000000000
Bacillus mojavensis strain RO-H-1	JACJGF000000000
Halobacillus halophilus strain DSM2266	Chromosome: NC_017668.1
	Plasmid PL16: NC_017669.1
	Plasmid PL3: NC_017670.1
Halomonas elongate strain DSM2581	Chromosome: NC_014532.1
Aliivibrio fischeri strain ES114	Chromosome I: NC_006840
	Chromosome II: NC_006841
	Plasmid pES100: NC_006842
Haloarcula marismortui strain DSM3752	Chromosome I: NC_006396.1
	Chromosome II: NC_006397.1
	Plasmid pNG100: NC_006389.1

	Plasmid pNG200: NC_006390.1
	Plasmid pNG300: NC_006391.1
	Plasmid pNG400: NC_006392.1
	Plasmid pNG500: NC_006393.1
	Plasmid pNG600: NC_006394.1
	Plasmid pNG700: NC_006395.1
Haloferax dentrificans strain DSM4425	Assembly RefSeq accession: GCF_000337795.1
Haloferax mediterranei strain R-4	Chromosome: NC_017941.2
	Plasmid pHM100: NC_017942.1
	Plasmid pHM300: NC_017943.1
	Plasmid pHM500: NC_017944.1
Haloferax volcanii strain DS2	Chromosome: NC_013967.1
	Plasmid pHV1: NC_013968.1
	Plasmid pHV2: NC_013965.1
	Plasmid pHV3: NC_013964.1
	Plasmid pHV4: NC_013966.1

Analyzing genomic proximity between HGT fragments and between HGT fragments and mutations

Pairwise genomic distances (in bp) between all HGT-acquired fragments from the other-*Bacillus* populations and between HGT fragments and mutations were calculated. For each case, three different distance thresholds were considered: 1%, 0.5% and 0.25% of half the genome of *B. subtilis* 168 (when half a genome is the maximal distance between two positions in a circular genome), corresponding to 21078bp, 10539bp and 5269bp, respectively. For each threshold in each of the two analyses, all pairwise distances between the examined genetic alterations (either foreign DNA fragments alone or foreign DNA fragments and mutations) that were equal or smaller than the threshold were counted and divided by the total number of pairwise distances. This ratio was defined as the proximity score of the examined alterations in the other-*Bacillus* populations. The p-value of the hypothesis that either fragments are clustered next to one another or that mutations and HGT fragments are clustered, was set to be the probability of obtaining a score similar to or higher than the observed proximity score, in a distribution of scores calculated from 1000 randomizations of the locations of the genetic alterations examined (See Figure S5 for illustration). When examining clustering between HGT fragments, the locations of the foreign DNA fragments were randomized, whereas in the clustering between HGT fragments and mutations, we randomized the location of mutations. The randomizations preserved the real number of foreign DNA events and their size distribution, as well as the real number of mutations.

In order to explore the possibility that clustering occurs in a clone-dependent manner, a further step was added. All pairwise distances were divided into those within the same clone, and those between different clones. A proximity score was calculated for the two groups of distances with the same threshold. In order to check if there is enrichment for shorter distances within clones, the within clones' proximity score was compared to a distribution of 1000 such scores in which the clone identity of each fragment (in the case of clustering

between HGT fragments) or each mutation (in the case of clustering between HGT fragments and mutations) was randomized (preserving the real number of foreign DNA fragments and mutations in each clone). A p-value was obtained by counting the frequency of randomized scores that are equal to or larger than that observed within the clone proximity score.

Competition-based fitness measurement

A 5.6kb fragment comprising the HGT fragment of interest (4.3kb) plus an additional ~500bp from each side (coordinated 3,811,990-3,817,611 on *Bacillus subtilis* 168 genome) was PCR amplified from the RS-D-2 donor genome. 400ng of PCR product was transformed to *Bacillus subtilis* 168 competent cells (the ancestral strain used in the experimental evolution experiment) together with 4ng of integrative plasmid containing a Spectynomycin resistance gene (pDG1731 plasmid, BGSC cat# ECE119). The transformed cells were then plated on LB plates containing 100µg/ml Spectynomycin (Sigma cat# S4014) and incubated overnight at 30°C. A total of 41 transformation reactions were done, and all resulting colonies (~4100 colonies, many of which did not integrated any part of the amplified fragment) were scraped and pooled together. This pool served as generation zero of the competition experiment. Competition was carried out by serial dilution as follows: Cells were grown in 1.2ml of LB + 0.8M NaCl or LB medium, both containing 5µg/ml erythromycin. Cultures were grown at 30°C under shaking conditions until reaching the stationary phase and then diluted by a factor of 1:120 into fresh media (~ 7 generations per dilution). This procedure was repeated daily for 10 days (a total of 70 generations). The competition experiment under LB + 0.8M NaCl medium was performed in six independent replicates, and the competition in LB medium was carried out in three replicates. Each day, samples from each replicate were frozen in 30% glycerol. To calculate the fitness of a foreign-DNA-containing cell, genomic DNA from samples taken at generations 0, 42 and 70 of the competition experiment was purified as described in the “Genomic DNA extraction” section. The genomic region of interest (~5.6kb)

was PCR amplified and cleaned using SPRI-beads. PCR amplicons were sequenced as previously described (Blecher-Gonen et al. 2013) with the following modifications: ~2000ng of DNA was sheared using the Vovaris E220X sonicator (Covaris). 1000ng of sheared DNA was used for library preparation as follow: End repair was performed at 20°C for 30 min and then purified using 0.75X Agencourt Ampure XP beads cleanup. A-bases were added to both 3' ends, the product was purified with 2.2X Ampure XP beads and adapters were ligated (25°C for 15 min). The ligation product was purified with 0.75X Ampure XP beads and amplified by PCR for 8 cycles. Libraries were quantified by Qubit and qPCR analysis using Illumina primers. Libraries were pooled and sequenced on 1 lane of NextSeq (Illumina) using the high output PE150_V2 kit. Sequencing reads were aligned to the reference Wild-Type sequence using Bowtie2 (Langmead and Salzberg 2012). Bam files were generated from the alignment using SamTools (Li et al. 2009). Bam files were analyzed using a custom pipeline for generating a frequency matrix of the known variable positions in the segment, across all samples.

The competition sample with the lowest increase in donor variants frequency (sample 2) was compared with the ancestor sample in order to calculate a confidence interval for the difference in donor proportions between the two samples for each given position of variation within the fragment. Since this test requires the assumption of normal distribution, only SNPs with coverage high enough to maintain: N* p > 10 and N * (1-p)> 10 where N is the total read count and p is the proportion of donor variants, were used for analysis (92% of the SNPs). The confidence interval at each SNP was calculated as follows:

$$\partial = p_2 - p_1 \pm z_{0.025} * \sqrt{\frac{p_2 * (1 - p_2)}{n_2} * \frac{p_1 * (1 - p_1)}{n_1}}$$

Where ∂ is the confidence interval, $p_1$and p_2 are the proportions of donor variants in the ancestor and the competition sample respectively, and n_1 and $n_2$are the total number of reads in the ancestor and the competition samples respectively.

Determining the percentage of competent cells in the population

Cells were grown overnight in 3ml of LB medium at 30°C until reaching the stationary phase. Cells were then diluted (1:120) into 1.2ml of fresh LB medium containing 0.8M NaCl and ~ 2.4µg genomic DNA extracted from *Bacillus subtilis* strain 168 containing two genes conferring resistance to one of two antibiotics, Phleomycin or Chloramphenicol (a kind gift from Avigdor Eldar, Tel Aviv university, Israel). In parallel, the same overnight culture was diluted into a similar medium that did not contain DNA. Cultures were then incubated for 24h at 30°C. To score transformants, 250µl of each culture was plated on LB plates containing 5ng/µl Chloramphenicol (Sigma cat# C0378). In order to estimate the total number of cells in each culture, cells were serially diluted and plated on LB plates. Plates were incubated overnight at 30°C and the number of colonies was counted. The ratio of Chloramphenicol-resistant cells in the population was calculated as the ratio between the number of colonies grown on Chloramphenicol-containing plates to that of colonies grown on LB (both corrected for plating and dilution factors). In order to deduce the percentage of competent cells in the populations, this ratio was multiplied by 100 (to obtain percentage) and then multiplied by 10^3, which approximates the ratio between the antibiotic-resistance gene length and the total genome size. In order to estimate the change in competence level during evolution, the same procedure was done for the ancestor and populations taken from different time points during evolution.

References

- Blecher-Gonen R, Barnett-Itzhaki Z, Jaitin D, Amann-Zalcenstein D, Lara-Astiaso D, Amit I. 2013. High-throughput chromatin immunoprecipitation for genome-wide mapping of in vivo

protein-DNA interactions and epigenomic states. Nat Protoc. 8(3):539–54. doi:10.1038/nprot.2013.023. http://www.ncbi.nlm.nih.gov/pubmed/23429716.

- Charlebois RL, Lam WL, Cline SW, Doolittle WF. 1987. Characterization of pHV2 from *Halobacterium volcanii* and its use in demonstrating transformation of an archaebacterium. Proc Natl Acad Sci U S A. 84(23):8530–4. doi:10.1073/pnas.84.23.8530. http://www.ncbi.nlm.nih.gov/pubmed/2825193.
- Cohan FM, Roberts MS, King EC. 1991. The potential for genetic exchange by transformation within a natural population of *Bacillus subtilis*. Evolution. 45(6):1393–1421. doi:10.1111/j.1558-5646.1991.tb02644.x. http://www.ncbi.nlm.nih.gov/pubmed/28563825.
- Dunlap CA, Bowman MJ, Zeigler DR. 2020. Promotion of *Bacillus subtilis* subsp. inaquosorum, *Bacillus subtilis subsp. spizizenii* and *Bacillus subtilis subsp. stercoris* to species status. Antonie Van Leeuwenhoek. 113(1):1–12. doi:10.1007/s10482-019-01354-9. http://www.ncbi.nlm.nih.gov/pubmed/31721032.
- Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, Miller CA, Mardis ER, Ding L, Wilson RK. 2012. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. 22(3):568–76. doi:10.1101/gr.129684.111. http://www.ncbi.nlm.nih.gov/pubmed/22300766.
- Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods. 9(4):357–9. doi:10.1038/nmeth.1923. http://www.ncbi.nlm.nih.gov/pubmed/22388286%0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3322381.
- Li H, Durbin R. 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics. 26(5):589–95. doi:10.1093/bioinformatics/btp698. http://www.ncbi.nlm.nih.gov/pubmed/20080505.
- Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 25(16):2078–9. doi:10.1093/bioinformatics/btp352.

http://www.ncbi.nlm.nih.gov/pubmed/19505943.

- Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal. 17:10–12.
- Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 26(6):841–2. doi:10.1093/bioinformatics/btq033. http://www.ncbi.nlm.nih.gov/pubmed/20110278.
- Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. 2011. Integrative genomics viewer. Nat Biotechnol. 29(1):24–6. doi:10.1038/nbt.1754. http://www.ncbi.nlm.nih.gov/pubmed/21221095.
- Stefanic P, Decorosi F, Viti C, Petito J, Cohan FM, Mandic-Mulec I. 2012. The quorum sensing diversity within and between ecotypes of *Bacillus subtilis*. Environ Microbiol. 14(6):1378–89. doi:10.1111/j.1462-2920.2012.02717.x. http://www.ncbi.nlm.nih.gov/pubmed/22390407.
- Tarasov A, Vilella AJ, Cuppen E, Nijman IJ, Prins P. 2015. Sambamba: fast processing of NGS alignment formats. Bioinformatics. 31(12):2032–4. doi:10.1093/bioinformatics/btv098. http://www.ncbi.nlm.nih.gov/pubmed/25697820.