ARTICLES
https://doi.org/10.1038/s41559-020-1271-x
nature ecology & evolution

Synthetic cross-phyla gene replacement and evolutionary assimilation of major enzymes

Troy E. Sandberg[1], Richard Szubin[1], Patrick V. Phaneuf[1] and Bernhard O. Palsson[1,2]

The ability of DNA to produce a functional protein even after transfer to a foreign host is of fundamental importance in both evolutionary biology and biotechnology, enabling horizontal gene transfer in the wild and heterologous expression in the lab. However, the influence of genetic particulars on DNA functionality in a new host is poorly understood, as are the evolutionary mechanisms of assimilation and refinement. Here, we describe an automation-enabled large-scale experiment wherein *Escherichia coli* strains were evolved in parallel after replacement of the genes *pgi* or *tpiA* with orthologous DNA from donor species spanning all domains of life, from humans to hyperthermophilic archaea. Via analysis of hundreds of clones evolved for 50,000+ cumulative generations across dozens of independent lineages, we show that orthogene-upregulating mutations can completely mitigate fitness defects that result from initial non-functionality, with coding sequence changes unnecessary. Gene target, donor species and genomic location of the swap all influenced outcomes—both the nature of adaptive mutations (often synonymous) and the frequency with which strains successfully evolved to assimilate the foreign DNA. Additionally, time series DNA sequencing and replay evolution experiments revealed transient copy number expansions, the contingency of lineage outcome on first-step mutations and the ability for strains to escape from suboptimal local fitness maxima. Overall, this study establishes the influence of various DNA and protein features on cross-species genetic interchangeability and evolutionary outcomes, with implications for both horizontal gene transfer and rational strain design.

Horizontal gene transfer (HGT), the non-reproductive transmission of genetic material that can transcend species boundaries, is possible due to shared mechanisms of the DNA decoding machinery inherited by all known life after descent from the last universal common ancestor[1]. In addition to shaping Earth's web of life, HGT has both clinical and industrial importance due to influencing the spread of antimicrobial resistance[2] and facilitating the engineering of organisms with desired phenotypes[3], respectively. Thus, understanding how HGT content assimilates into a new genome is of both basic and applied interest. The mechanistic basis for gene transfer has been increasingly uncovered[4] and replacement studies have established the impressive extent to which many genes retain cross-species functionality despite billions of years of phylogenetic divergence[5–7]. However, studies on the functionality of gene replacement immediately post-transfer are insufficient to reveal how organisms can adapt to utilize foreign DNA that initially provides no fitness benefit.

To date, several studies have paired gene swaps with evolution, providing an insight into the process of foreign gene assimilation. Lind et al.[8] replaced ribosomal genes in *Salmonella typhimurium* with microbial orthologues; within a few hundred generations of laboratory evolution, they found gene amplifications aimed at increasing orthogene copy number as a way to ameliorate fitness defects. Kacar et al.[9] replaced the essential elongation factor Tu 2 (*tufB)* in *Escherichia coli* with an ancestral variant; evolution showed that it was similarly selected for upregulation[9]. Bershtein et al.[10] replaced the essential *folA* gene in *E. coli* with orthologues from 35 close bacterial relatives and found that fitness defects were frequently evolutionarily compensated by protease-deactivating mutations that increased intracellular orthogene levels. Such studies establish that insufficient expression levels regularly hinder in vivo functionality of foreign genes, but the influence of particular variables is difficult or impossible to deconvolute from the existing data. Codon optimization was typically performed on the foreign genes before insertion, a critical change that limits the applicability of observed results to natural HGT. Essential genes were also the main targets for replacement, precluding any outcome where an initially non-functional swap could evolve functionality and vastly limiting the pool of organisms from which an orthogene could be taken without inducing lethality in the new host. Additional confounding factors complicated mutational interpretation, such as the use of gene targets involved in protein–protein interactions and complex formation, or differences in plasmid versus chromosomal insertion. Recent advances in automation[11] enable experimental evolution of a heretofore infeasible scale and data resolution, providing an empirical means to address unresolved details on cross-species gene functionality, potential for evolutionary assimilation and mutational mechanisms.

In this study, we designed and constructed eight distinct gene-swapped *E. coli* strains with orthologous donor DNA from four species spanning all domains of life: the γ- and α-proteobacteria *Vibrio cholerae* and *Brucella melitensis*; the hyperthermophilic archaeum *Pyrobaculum aerophilum*; and the mammalian eukaryote *Homo sapiens* (Fig. 1a). The glycolytic isomerase genes *pgi* and *tpiA* were selected for swapping for several reasons—they are non-essential but crucial for fitness, highly studied with known post-knockout evolutionary outcomes[12–14] and do not require cofactors or participate in protein complexes (Extended Data Fig. 1 and Supplementary Table 1). To further minimize potential confounding factors, strain construction involved scarless chromosomal replacement, from start to stop codon, with the coding sequence of the foreign orthologue not subjected to any codon optimization. Automated adaptive laboratory evolution (ALE) systems[15] allowed dozens of lineages to be evolved for growth rate improvements with

[1]Department of Bioengineering, University of California, San Diego, La Jolla, CA, USA. [2]Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Lyngby, Denmark. ✉e-mail: palsson@ucsd.edu

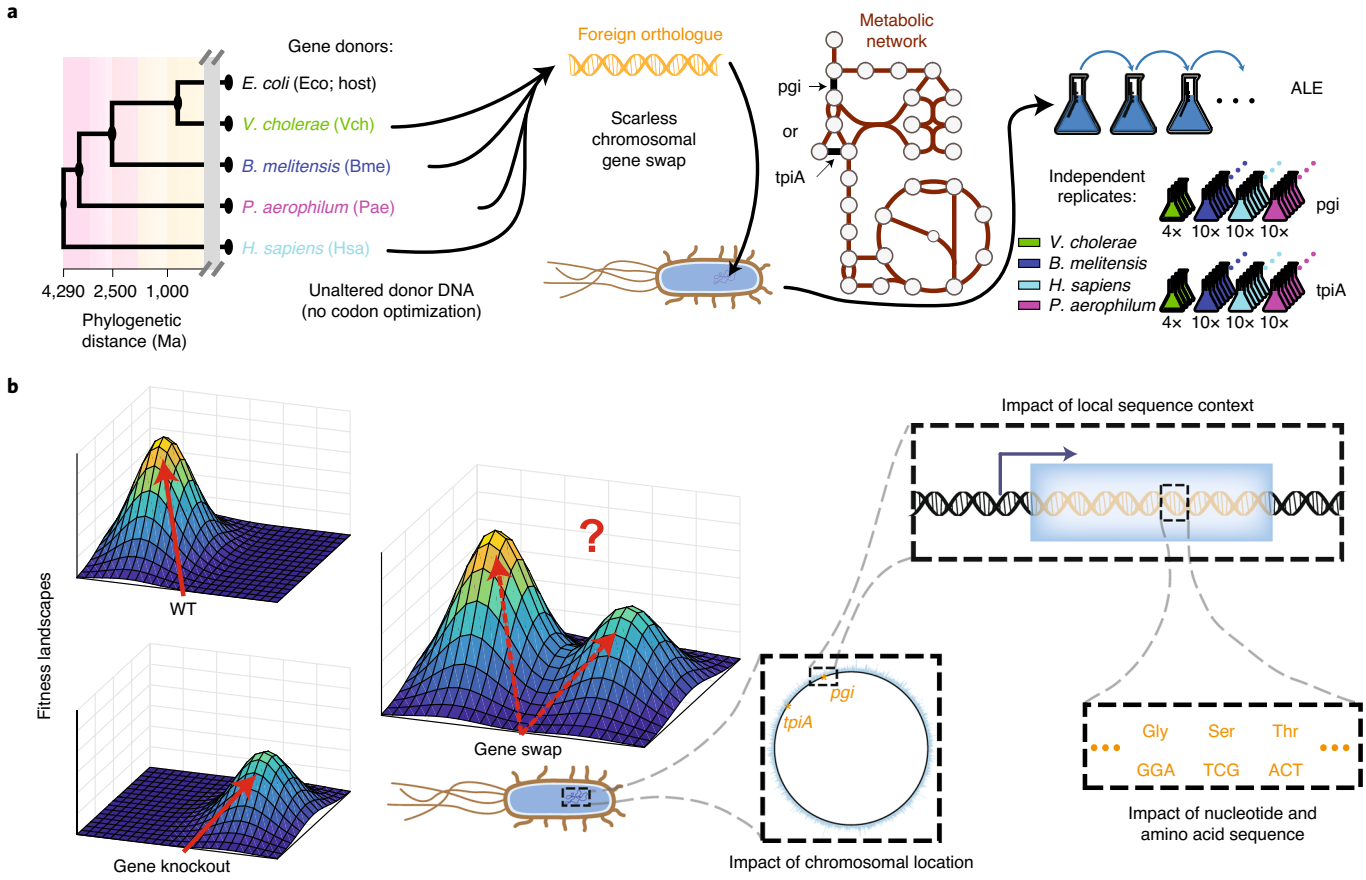


Fig. 1 | Gene swap strain construction and laboratory evolution. a, Schematic of the experimental workflow. Native *E. coli* glycolytic isomerases pgi and tpiA were replaced with the coding sequence of foreign orthologues and subjected to laboratory evolution for improved exponential phase growth rate. Ma, million years ago. **b**, Experimental evolution enabled interrogation of the fitness landscape after gene replacement. The spectrum of different swaps and the large replicate number provide an insight into various factors influencing adaptive outcomes.

real-time tracking of adaptive trajectories. Such high-throughput evolution studies essentially serve as empirical Monte Carlo sampling of the fitness landscape, providing an insight into the influence on adaptive outcomes of chromosomal location, local sequence context and nucleotide/amino acid features (Fig. 1b).

Results

Gene-swapped strains exhibited an initial physiology that was consistent with phylogenetic distance from the donor species: *V. cholerae* swaps did not show any phenotypic defects; *B. melitensis* swaps grew slower than wild-type (WT) but faster than full knockouts; while *P. aerophilum* and *H. sapiens* swaps had the same slow growth rate as knockouts. The intrinsic growth rate gap between WT and *pgi*- or *tpiA*-deficient strains, and their distinct evolutionary trajectories, enabled fitness-based classification of gene swap lineages into 'success' or 'failure' of orthogene assimilation (Fig. 2a). Across the 68 independently evolved gene swap lineages, we saw significantly different gene- and organism-specific outcomes (Fig. 2b and Extended Data Fig. 2). While only a single *tpiA* swap failure was found, *pgi* had less success with assimilation, with 40% of human swaps failing as well as 100% of archaeal swaps.

Evolved end point clones were isolated and whole-genome sequenced to determine the genetic mechanisms of adaptation; the mutational results provided striking reinforcement of the conclusions drawn from the initial physiology and adaptive outcomes. All 'failure' lineages (not reaching a growth rate above 0.75 per hour) lacked mutations in or around the foreign DNA, while every single 'successful' lineage acquired 1 or more mutations to this region, barring *V. cholerae* swaps, which did not require any orthogene changes to enable functionality in *E. coli* (Figs. 2c and 3a,b). Promoter and ribosome binding site mutations were the dominant adaptive mechanisms for *pgi* swaps, while *tpiA* swaps acquired the same synonymous L179L single-nucleotide polymorphism (SNP) in the upstream gene *yiiQ* more than 20 times independently. In most cases, fitness improvement to levels on par with evolved WT strains did not require any changes to the foreign coding sequence.

We performed several assays to verify that fitness improvement in 'successful' lineages was due to mutations increasing orthogene levels. Knocking out the orthologous gene from various end point strains caused significant drops in growth rate for successes but had no impact on failures (Extended Data Fig. 3a). Enzyme assays on these same end point strains and their unevolved post-swap ancestors were consistent with these results, revealing enzymatic activity levels that increased after evolution, but only in successes (Extended Data Fig. 3b). Finally, we expressed the *H. sapiens* orthogenes (the only donor species that led to failure and success lineages for both swapped genes) from inducible promoters inserted into the Δ*pgi* and Δ*tpiA* strains, demonstrating that growth rate increased with increasing induction level.

Failure lineages, in addition to never acquiring mutations within or proximal to the orthogenes (in any sequenced clones or populations), also had mutations highly characteristic of knockout control evolutions. For example, the *sthA* gene did not mutate in any

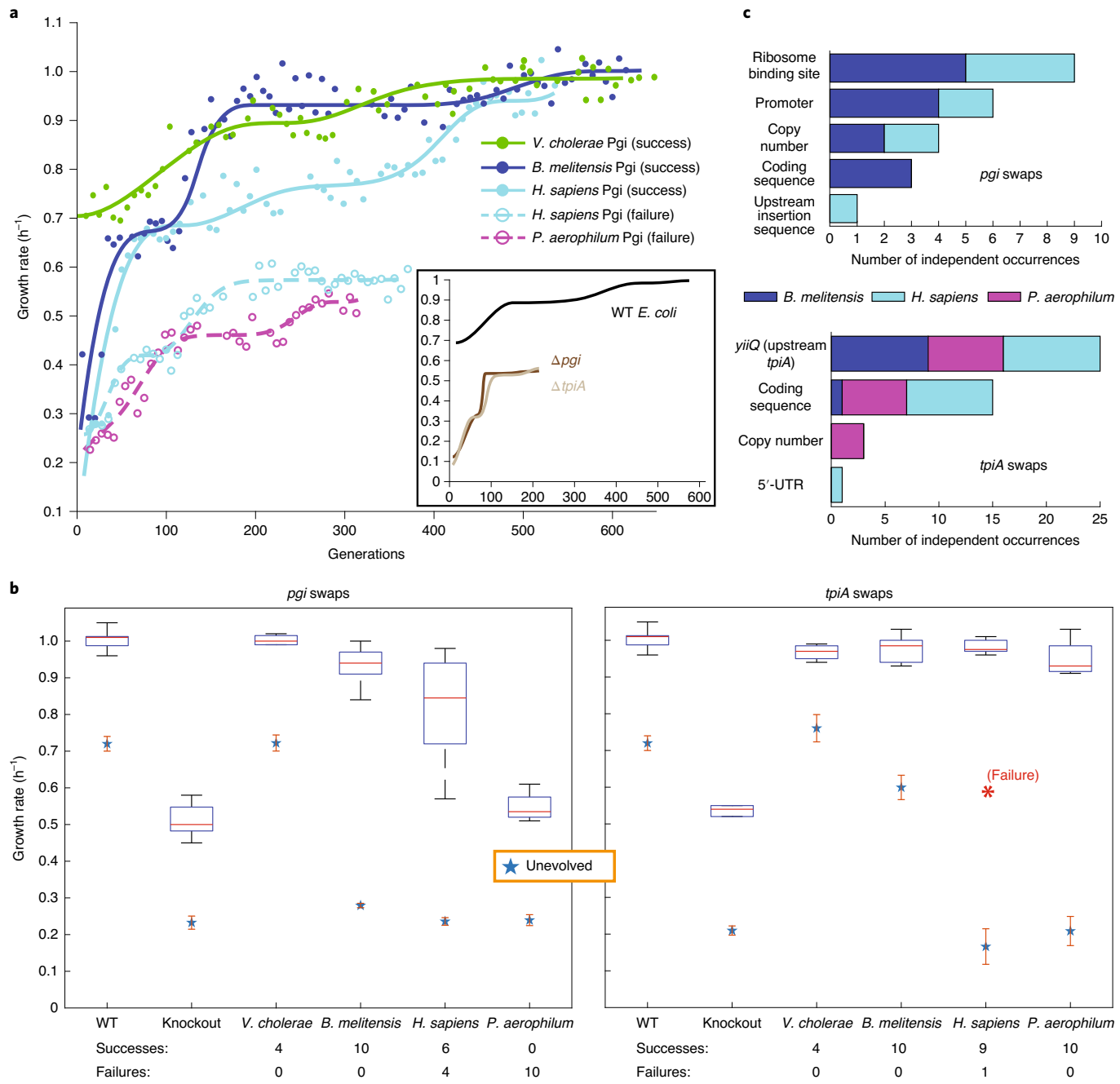
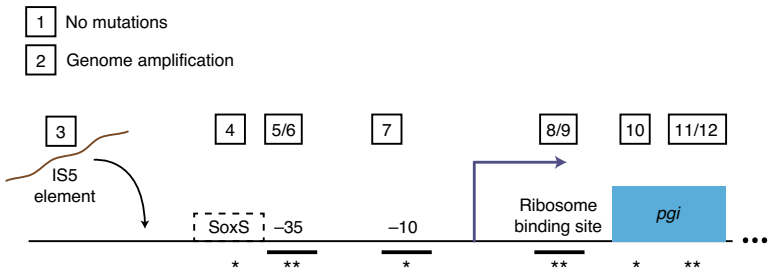


Fig. 2 | Evolutionary outcomes. a, Robotic systems allowed growth rate tracking over the course of evolution. (The points represent growth rates from successive culture tubes, with the lines being cubic interpolating splines fitted to the points.) Example trajectories demonstrate various outcomes; gene-swapped strains could evolve like WT ones, successfully escape from knockout-like fitness levels or fail to evolve any differently than knockout controls. **b**, Box plots (showing the first and third quartiles around the median red line, with whiskers to extreme values) of fitness outcomes for evolved population end points versus starting strains, with statistics on success and failure frequency. The error bars on the starting strains represent the s.d. from quadruplicate measurements. The red asterisk represents growth rate of the single *tpiA* failure lineage, excluded from the box plot of the other nine lineages. **c**, Mutations were found in or around orthogenes in all successful lineage end points. *V. cholerae* swaps did not require orthogene mutations to reach high growth rates, while failure lineages never acquired any.

successful *pgi* swaps but it mutated in two of the failures (one *H. sapiens pgi*, one *P. aerophilum pgi*) and in multiple Δ*pgi* controls[12]. Lineages mirroring WT control evolutions likewise shared characteristic mutations—three of four evolved *V. cholerae pgi* strains acquired *hns*/*tdk* insertion sequence element mutations, a causal mechanism only observed for WT glucose evolutions[16,17], highlighting the negligible influence of the *V. cholerae* swap. Considering only genes that mutated two or more times independently across the replicates, hierarchical clustering cleanly discriminated between *pgi* successes and failures (Fig. 3c). The single unsuccessful *H. sapiens tpiA* lineage likewise had unique adaptive signatures characteristic of Δ*tpiA* evolutions—*ptsG*, *galR* and *nemR* all mutated[13]. The large-scale nature of our study resulted in many genes targeted repeatedly for alteration, from which structural insights can be drawn and cross-study comparisons made that reveal mutational hotspots (Extended Data Fig. 4a and Supplementary Table 2).

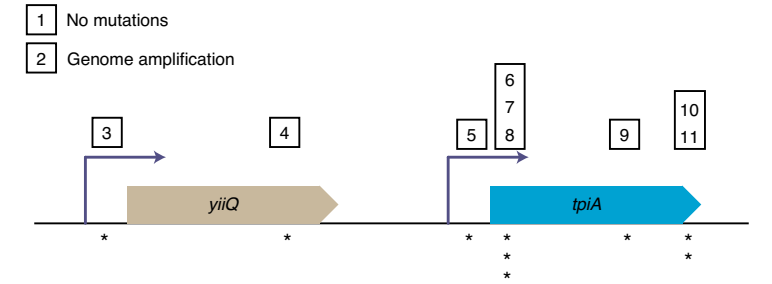
a



Number of occurences in end point ALE replicates

		V. cholerae	*B. melitensis*	*P. aerophilum*	*H. sapiens*
1	No mutations	4		10	4
2	DNA amplification		2		2
3	IS5 (–113 *pgi*)				1
4	T>C (–78 *pgi*)		1		
5	C>T (–69 *pgi*)				1
6	A>G (–68 *pgi*)		1		
7	C>T (–50 *pgi*)		2		1
8	A>G (–13 *pgi*)		1		
9	A>G (–12 *pgi*)		4		4
10	G>T (A5A)		1		
11	C>A (V12V)		1		
12	G>A (A13T)		1		

b



		V. cholerae	*B. melitensis*	*P. aerophilum*	*H. sapiens*
1	No mutations	4			1
2	DNA amplification			3	
3	A>C (–60 *yiiQ*)		2		
4	C>T (L179L)		7	7	9
5	+A (–42 *tpiA*)				1
6	C>T (P3L)		1		
7	C>A (P3H)				5
8	C>A (P3P)				3
9	G>A (G198R)			1	
10	C>A (P226Q)			4	
11	C>G (P226R)			1	

c

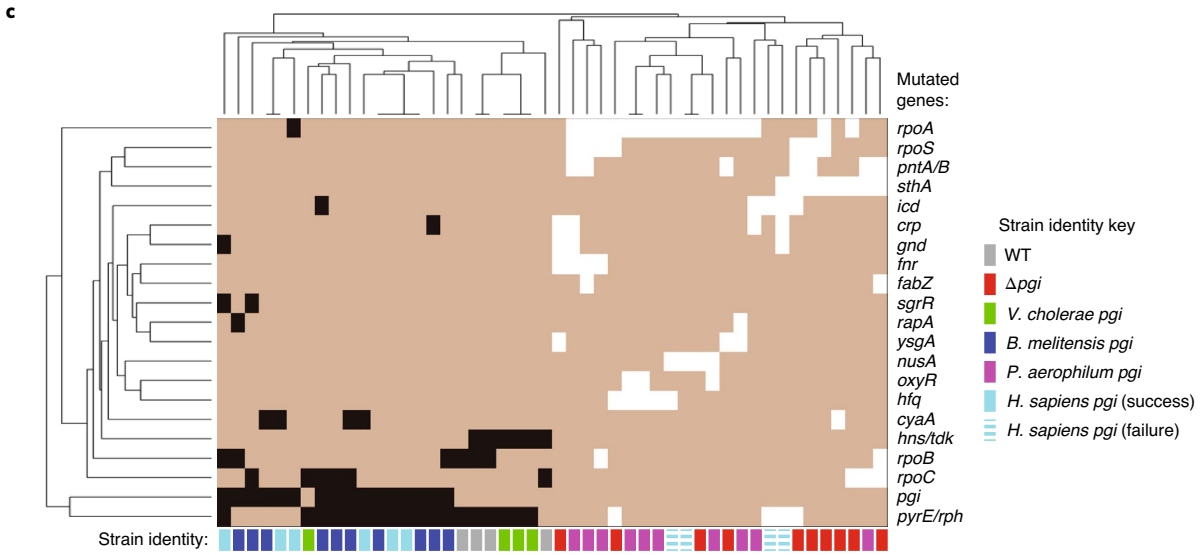


Fig. 3 | Evolved strain mutations. a, *pgi*-proximal mutations found in gene-swapped end point ALE strains. **b**, *tpiA*-proximal mutations found in gene-swapped end point ALE strains. **c**, Hierarchical clustering of strains based on mutations (indicated by the black or white squares) found in evolved end points, considering only genes that mutated independently two or more times. WT and successful lineages clustered together based on shared characteristic gene alterations (left), as did knockout and failure lineages (right). All *V. cholerae pgi* and *B. melitensis pgi* strains were successful, while all *P. aerophilum pgi* strains were failures.

Given the significant fitness defect resulting from absent or low-level *pgi* or *tpiA* flux, orthogene mutations (Fig. 2c) would be expected to improve fitness in one of two ways: by increasing expression level of the gene product through regulatory changes; or by enhancing specific enzyme activity through coding sequence alterations. While many of the observed mutations can be easily interpreted as expression-increasing (promoter/ribosome binding site SNPs and copy number expansions for upregulation), others require more detailed analysis. We found that the widespread *yiiQ* L179L SNP achieved orthogene upregulation by creating a new promoter 179 base pairs upstream of the *tpiA* start codon. Analysis with promoter prediction tools[18] revealed the increased chance for RNA polymerase binding that the C→T mutation creates (Fig. 4a). This mechanism was documented in a previous study[19] and demonstrates the strong impact of local sequence context on adaptive outcomes.

Coding sequence changes to the orthogenes were less common than *cis*-regulatory alterations and fell into two types: C-terminal

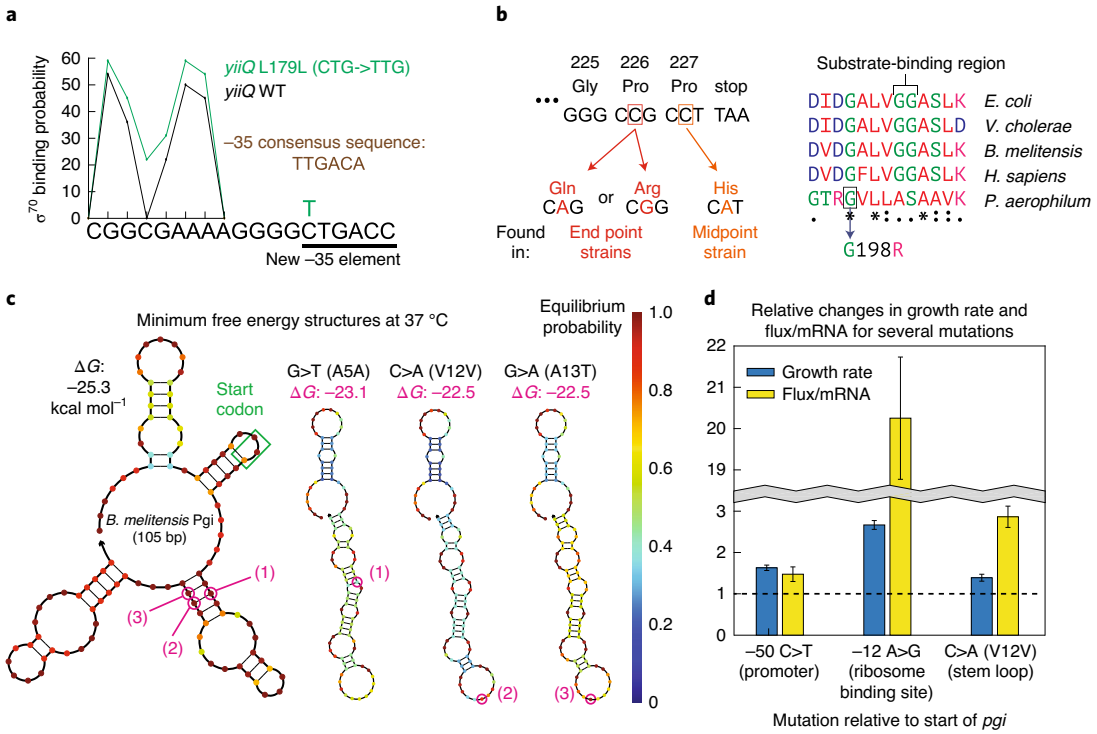


Fig. 4 | Mutational mechanisms of orthogene assimilation. a, A promoter-creating synonymous SNP in the gene upstream of *tpiA* occurred independently more than 20 times. **b**, Missense SNPs in the archaeal *tpiA* ameliorated polyproline ribosome stalling (left); in one instance, it probably modified enzyme properties (right). **c**, N-terminal SNPs, frequently synonymous, targeted prohibitively strong stem loops in mRNA secondary structure for destabilization. **d**, Several upregulating mutations had different impacts on the ratio of orthogene enzyme flux to mRNA level. Each pair of bars reflects growth, enzyme and qPCR assays performed on two different strains—one lacking the mutation in question and one genetically identical except for the addition of the listed mutation. The error bars represent the s.d. from quadruplicate measurements.

missense SNPs in the archaeal *tpiA*; and several N-terminal, mostly synonymous, SNPs across a variety of swapped strains. *P. aerophilum tpiA* swaps repeatedly acquired SNPs to the penultimate amino acid, consistent with a mechanism of increasing expression by reducing polyproline ribosomal stalling[20]. This mechanism explains the occurrence of growth-improved clones stemming from changes to either of the final two proline residues (Fig. 4b), although it is possible that the prolines hinder folding of the protein within *E. coli* rather than causing ribosomal stalling. The archaeal *tpiA* swap was also characterized by the only observed orthogene mutation that probably altered enzyme activity rather than expression level—residue 198 was adjacent to the conserved substrate-binding region of the enzyme.

Our observed N-terminal orthogene SNPs were frequently synonymous; although earlier experimental evolution studies did not determine the particular causal mechanism for fitness improvement[21,22], recent work identified the secondary structure of messenger RNA as the target for such SNPs in the case of a gene knockout that selected for native promiscuous enzyme upregulation[23]. In this study, we found that this mechanism of expression increase extended to orthogene assimilation, with both synonymous and non-synonymous SNPs targeting tightly bound stem loops in mRNA secondary structure for destabilization. This opens up the transcript for increased ribosomal read-through, most strikingly in the cases of *B. melitensis pgi* (Fig. 4c) and *H. sapiens tpiA* (Extended Data Fig. 4b), leading to more protein per mRNA. We validated this mechanism empirically with pairwise characterizations of *pgi* expression level and enzymatic activity in strains genetically identical except for single mutations of interest, thus enabling causal establishment (Fig. 4d). Moreover, we surveyed approximately 1,000 whole-genome sequenced *E. coli* strains and found that SNPs preferentially accumulated in the strongest stem-loop region (Extended Data Fig. 5a). Synonymous SNPs of this type may be underappreciated drivers of adaptation rather than neutral signatures of drift, potentially overlooked in previous evolution experiments or even contributing to speciation (Extended Data Fig. 5b).

We next probed the evolutionary dynamics governing adaptive outcomes by sequencing multiple midpoint strains from every ALE lineage and performed 'continuation ALEs' for a number of failure end points as well as 'replay ALEs' for selected clones of interest isolated from various lineages. We found that the first mutation fixed in an evolving population strongly constrained ultimate lineage outcome, exemplified by the single *tpiA* failure lineage (Fig. 5a). In ten out of ten Δ*tpiA* controls, the first adaptive step was knockout of *ptsG*, mirrored by the *H. sapiens tpiA* failure, but additional evolutionary time allowed this lineage to undo the *ptsG* nonsense SNP it had acquired and ultimately achieve success. Had the failure lineage knocked out *ptsG* via some other mechanism, such as a frameshifting indel, open reading frame restoration would have been prevented and escape from 'failure' might not have been possible. However, some lineages in the initial ALE achieved success despite first acquiring mutations characteristic of knockout (Extended Data Fig. 6); added ALE time for failures normally enabled success, although not for any archaeal *pgi* swaps (Extended Data Fig. 7).

Our midpoint sequencing also revealed that, for *tpiA* swaps, orthogene copy number expansions were much more widespread than indicated by end point clones. While the chromosomal location of *pgi* left it little flanking homology with which to facilitate genome amplifications (Extended Data Fig. 8a), *tpiA* fell between

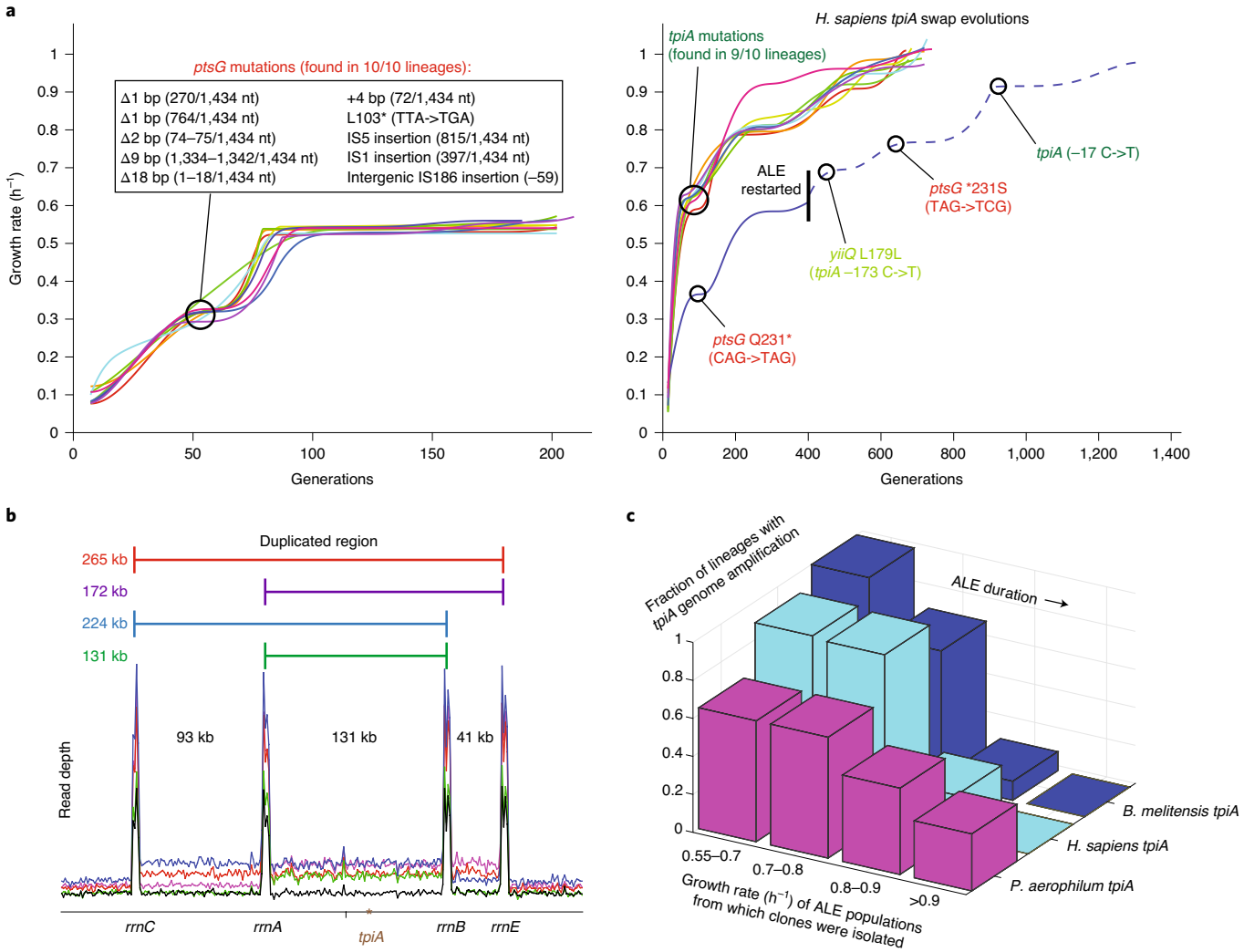


Fig. 5 | Adaptive dynamics. a, Knockout of *ptsG* was highly characteristic of Δ*tpiA* control evolutions. The single failure *tpiA* swap lineage took an initial step down this suboptimal knockout-like adaptive path, but when provided additional evolutionary time ultimately restored *ptsG* and successfully assimilated the orthogene. Asterisks represent stop codons in the coding sequence. nt, nucleotide. **b**, Copy number increases in *tpiA* came in four distinct types driven by homology between flanking *rrn* operons. **c**, Increases in *tpiA* copy number were widespread at the start of evolution but decreased precipitously in frequency as the experiment progressed.

several ribosomal RNA (*rrn*) operons that could cross over in different pairings; this led to four distinct types of copy number expansions (Fig. 5b). The increased frequency with which such homology-facilitated expansions occur is a probable reason for the greater success rate of *pgi* versus *tpiA* swaps, further emphasizing the importance of a gene's chromosomal location on adaptive outcomes. The extent of the amplified region had drastic changes in relative fitness for different growth environments (Extended Data Fig. 8b); more than 80% of clones isolated across the *B. melitensis*/*H. sapiens*/*P. aerophilum tpiA* lineages after their first jump in growth rate contained such expansions. As the ALE experiment progressed, these *tpiA*-amplified strains were ultimately outcompeted by more cellular resource parsimonious, that is, efficient, methods of upregulation (for example, *yiiQ* L179L and mRNA stem- loop SNPs), leaving only three *P. aerophilum tpiA* end points with a persistent amplification (Fig. 5c). Replay ALEs confirmed that amplified regions of the genome could further increase in copy number, remain stable or collapse back to single copy as evolution progressed (Extended Data Fig. 9).

Discussion

Taken together, our laboratory evolution study of orthogene assimilation and refinement reveals mechanisms underlying cross-species gene functionality and adaptive outcomes, such as mRNA stem-loop formation, *cis*-regulation and chromosomal copy number stability. Adjustments to enzyme abundance were much more common than alterations to specific enzyme activity, emphasizing that systems biology may be key to understanding adaptive evolutionary processes. Adaptive synonymous mutations were also surprisingly widespread—one replay lineage even acquired two successive synonymous orthogene SNPs as its only coding sequence alterations. Calculations of adaptive protein substitution rates are known to be sensitive to even weak selection for synonymous mutations[24]; thus, our results raise concerns about the applicability of K_a/K_s in evolutionary biology and the accuracy of existing genetic clock studies.

The collection of hundreds of evolved, whole-genome sequenced strains generated in this study also highlights several interesting or unexpected features. First, real-time tracking of culture growth rates with automated systems, at least in cases where individual

mutations can cause large fitness jumps, reliably enables the isolation of strains identical except for single genetic alterations. We captured all types of 'quantum steps' in evolution, that is, the smallest possible units of genetic change—from SNPs to indels to genome rearrangements. Strain pairs differing by such quantum steps in genome sequence space allow the validation of mutational mechanisms without necessitating the creation of knock-ins (Fig. 4d). Importantly, genome rearrangements often underlie adaptive events but cannot be perfectly reproduced with current genome engineering techniques[25]; yet our strain collection enabled us to assign causal influence to several of these rearrangements (Extended Data Fig. 8b). Second, we see a shocking extent of mutational reproducibility, with the same exact DNA base pair change occurring independently more than 20 times across 3 distinct strains (Fig. 3b). This demonstrates that evolutionary outcomes can be (probabilistically) predicted to the single-base-pair level, something not observed in higher organisms to date[26]. Finally, the non-orthogene mutations appearing in our strains fell predominantly within genes of the RNA polymerase complex, where we catalogued more than 90 unique mutations. Although RNA polymerase mutations are a common occurrence in evolution experiments[27], the fact that identical SNPs appear repeatedly after both our gene swaps and five entirely distinct metabolic gene knockouts[28] (Supplementary Table 2) is unexpected. With evolution-generated strain collections, such as the one in this study, paired with new analytical techniques[29], we are on the path to elucidating the role of RNA polymerase as a master regulator of transcriptional networks.

Methods

Strain design and engineering. DNA sequences for gene replacement were ordered from Gene Universal. For the human genes, the coding sequence (introns removed) of the annotated main isoform was used. Strains were constructed in two different ways—*pgi* swaps with a modified gene gorging protocol[30] as depicted in Extended Data Fig. 10 and *tpiA* swaps with a similar method but using CRISPR-induced double-stranded DNA breaks on the native *E. coli* sequence as the method of counterselection, thus not requiring an antibiotic cassette. Orthologous gene knockouts (Extended Data Fig. 3a) were generated via P1 phage transduction[31] from Δ*pgi* and Δ*tpiA* strains. Strain construction was checked for compositional and locational accuracy with both whole-genome and Sanger sequencing.

Protein similarity scores (Supplementary Table 1) were obtained using EMBOSS Needle pairwise sequence alignment[32]. The codon adaptation index (Supplementary Table 1) was calculated with the tool from Biologics International Corp (https://www.biologicscorp.com/tools/CAICalculator).

ALE. Strains were maintained in exponential phase growth and evolved via batch culture serial propagation of 100 µl volumes into 15 ml (working volume) tubes of 4 g l^{-1} glucose M9 minimal medium kept at 37 °C and aerated via magnetic stirring, exactly as described previously[33]. Cultures were propagated for 30 d or until the measured population growth rate reached within 10% of the maximum known to occur for the growth conditions, so that fitness trajectories would be dominated by sequential selective sweeps of large-effect beneficial mutations[34]. Stopped cultures were designated 'end points' and, due to the stochastic timing with which populations experienced jumps in growth rate, the total evolution time in generations varied across the cultures. Each independent ALE replicate was started from a unique pre-culture inoculated with a clone isolated from a lysogeny broth plate so as to prevent standing variation from influencing mutational independence across replicates.

DNA sequencing (DNA-seq) and analysis. Genomic DNA was isolated using bead agitation in 96-well plates as outlined previously[35]. Paired-end whole-genome DNA-seq libraries were generated with a Kapa HyperPlus Library Prep Kit (Kapa Biosystems) and run on an Illumina HiSeq 4000 platform with a HiSeq SBS Kit (150 base-pair (bp) reads). The generated DNA-seq FASTQ files were quality-controlled with AfterQC v.0.9.7 (ref. [36]) then processed with the breseq computational pipeline[37] according to standard procedures (https://barricklab.org/twiki/pub/Lab/ToolsBacterialGenomeResequencing/documentation/) and aligned to the *E. coli* genome (National Center for Biotechnology Information accession no. NC_000913.3) to identify mutations. Genome amplifications were determined with a custom script that identified discontinuities in read depth; all read depth coverage plots and marginal mutation calls were manually inspected.

mRNA structural analysis. Structures for mRNA transcripts were evaluated using the NUPACK computational tool[38] version 3.2.2. Annotated transcription start sites served as the 5′ end for evaluated structures; a range of transcript lengths in 5 bp intervals were analysed to ensure the robustness of results to the chosen region.

Strain characterization. Strains were assayed for growth rate via serial propagation and optical density (OD) sampling conditions identical to the ALE experiments, with given values averaged over a minimum of four independent growth tubes. Strains were assayed for *pgi* expression level via quantitative PCR with reverse transcription as follows: total RNA was purified with the QIAGEN RNeasy Mini Kit, assessed for quality on an Agilent 2100 Bioanalyzer model G2939A and quantified with a NanoDrop Spectrophotometer (Thermo Fisher Scientific). RNA was converted to complementary DNA with the LunaScript RT Kit (New England Biolabs), then quantified with the Luna Universal One-Step RT–qPCR Kit (New England Biolabs). A panel of five housekeeping genes were assayed along with *pgi* to allow for normalization. Strains were assayed for enzymatic flux with a Phosphoglucose Isomerase Activity Colorimetric Assay Kit (catalogue no. K775; BioVision) according to the manufacturer's protocols. Flux/mRNA values (Fig. 4d) were obtained by taking the ratio of the colorimetric assay results to the qPCR results. For both expression level and flux measurements, cultures were first flash-frozen (both biological and technical duplicates) in liquid nitrogen at the same OD in mid-exponential growth phase.

Statistical analysis. The quantitative values presented (that is, growth rates, qPCR mRNA levels, enzyme flux assays) are averages from quadruplicate measurements, with the error bars representing the s.d.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The genome sequence data that support the findings of this study are available from ALEdb (https://aledb.org) under project name 'SvNS'.

Code availability

AfterQC, the software used to trim and filter DNA-seq reads, is available at https://github.com/OpenGene/AfterQC. Breseq, the software used to identify mutations, is available at https://github.com/barricklab/breseq. Co, the software used to edit genome references (https://github.com/SBRG/svns_refseq), is available at https://github.com/biosustain/co.

Received: 17 June 2019; Accepted: 10 July 2020;
Published online: 10 August 2020

References

1. Soucy, S. M., Huang, J. & Gogarten, J. P. Horizontal gene transfer: building the web of life. *Nat. Rev. Genet.* **16**, 472–482 (2015).
2. Palmer, K. L., Kos, V. N. & Gilmore, M. S. Horizontal gene transfer and the genomics of enterococcal antibiotic resistance. *Curr. Opin. Microbiol.* **13**, 632–639 (2010).
3. Potvin, G., Ahmad, A. & Zhang, Z. Bioprocess engineering aspects of heterologous protein production in *Pichia pastoris*: a review. *Biochem. Eng. J.* **64**, 91–105 (2012).
4. Chen, J. et al. Genome hypermobility by lateral transduction. *Science* **362**, 207–212 (2018).
5. Kachroo, A. H. et al. Evolution. Systematic humanization of yeast genes reveals conserved functions and genetic modularity. *Science* **348**, 921–925 (2015).
6. Kachroo, A. H. et al. Systematic bacterialization of yeast genes identifies a near-universally swappable pathway. *eLife* **6**, e25093 (2017).
7. Kacar, B., Garmendia, E., Tuncbag, N., Andersson, D. I. & Hughes, D. Functional constraints on replacing an essential gene with its ancient and modern homologs. *mBio* **8**, e01276-17 (2017).
8. Lind, P. A., Tobin, C., Berg, O. G., Kurland, C. G. & Andersson, D. I. Compensatory gene amplification restores fitness after inter-species gene replacements. *Mol. Microbiol.* **75**, 1078–1089 (2010).
9. Kacar, B., Ge, X., Sanyal, S. & Gaucher, E. A. Experimental evolution of *Escherichia coli* harboring an ancient translation protein. *J. Mol. Evol.* **84**, 69–84 (2017).
10. Bershtein, S. et al. Protein homeostasis imposes a barrier on functional integration of horizontally transferred genes in bacteria. *PLoS Genet.* **11**, e1005612 (2015).
11. Sandberg, T. E., Salazar, M. J., Weng, L. L., Palsson, B. O. & Feist, A. M. The emergence of adaptive laboratory evolution as an efficient tool for biological discovery and industrial biotechnology. *Metab. Eng.* **56**, 1–16 (2019).
12. Charusanti, P. et al. Genetic basis of growth adaptation of *Escherichia coli* after deletion of *pgi*, a major metabolic gene. *PLoS Genet.* **6**, e1001186 (2010).
13. McCloskey, D. et al. Adaptation to the coupling of glycolysis to toxic methylglyoxal production in *tpiA* deletion strains of *Escherichia coli* requires synchronized and counterintuitive genetic changes. *Metab. Eng.* **48**, 82–93 (2018).

14. McCloskey, D. et al. Multiple optimal phenotypes overcome redox and glycolytic intermediate metabolite imbalances in *Escherichia coli pgi* knockout evolutions. *Appl. Environ. Microbiol.* **84**, e00823-18 (2018).
15. Sandberg, T. E. et al. Evolution of *Escherichia coli* to 42 °C and subsequent genetic engineering reveals adaptive mechanisms and novel mutations. *Mol. Biol. Evol.* **31**, 2647–2662 (2014).
16. LaCroix, R. A. et al. Use of adaptive laboratory evolution to discover key mutations enabling rapid growth of *Escherichia coli* K-12 MG1655 on glucose minimal medium. *Appl. Environ. Microbiol.* **81**, 17–30 (2015).
17. Sandberg, T. E. et al. Evolution of *E. coli* on [U-^{13}C]glucose reveals a negligible isotopic influence on metabolism and physiology. *PLoS ONE* **11**, e0151130 (2016).
18. de Avila e Silva, S.,& Notari, D. L., Neis, F. A., Ribeiro, H. G. & Echeverrigaray, S. BacPP: a web-based tool for Gram-negative bacterial promoter prediction. *Genet. Mol. Res.* **15**, gmr7973 (2016).
19. Kershner, J. P. et al. A synonymous mutation upstream of the gene encoding a weak-link enzyme causes an ultrasensitive response in growth rate. *J. Bacteriol.* **198**, 2853–2863 (2016).
20. Peil, L. et al. Distinct XPPX sequence motifs induce ribosome stalling, which is rescued by the translation elongation factor EF-P. *Proc. Natl Acad. Sci. USA* **110**, 15265–15270 (2013).
21. Bailey, S. F., Hinz, A. & Kassen, R. Adaptive synonymous mutations in an experimentally evolved *Pseudomonas fluorescens* population. *Nat. Commun.* **5**, 4076 (2014).
22. Agashe, D. et al. Large-effect beneficial synonymous mutations mediate rapid and parallel adaptation in a bacterium. *Mol. Biol. Evol.* **33**, 1542–1553 (2016).
23. Kristofich, J. et al. Synonymous mutations make dramatic contributions to fitness when growth is limited by a weak-link enzyme. *PLoS Genet.* **14**, e1007615 (2018).
24. Matsumoto, T., John, A., Baeza-Centurion, P., Li, B. & Akashi, H. Codon usage selection can bias estimation of the fraction of adaptive amino acid fixations. *Mol. Biol. Evol.* **33**, 1580–1589 (2016).
25. Leon, D., D'Alton, S., Quandt, E. M. & Barrick, J. E. Innovation in an *E. coli* evolution experiment is contingent on maintaining adaptive potential until competition subsides. *PLoS Genet.* **14**, e1007348 (2018).
26. Concha, C. et al. Interplay between developmental flexibility and determinism in the evolution of mimetic *Heliconius* wing patterns. *Curr. Biol.* **29**, 3996–4009.e4 (2019).
27. Conrad, T. M. et al. RNA polymerase mutants found through adaptive evolution reprogram *Escherichia coli* for optimal growth in minimal media. *Proc. Natl Acad. Sci. USA* **107**, 20500–20505 (2010).
28. Wytock, T. P. et al. Experimental evolution of diverse *Escherichia coli* metabolic mutants identifies genetic loci for convergent adaptation of growth rate. *PLoS Genet.* **14**, e1007284 (2018).
29. Sastry, A. V. et al. The *Escherichia coli* transcriptome mostly consists of independently regulated modules. *Nat. Commun.* **10**, 5536 (2019).
30. Herring, C. D., Glasner, J. D. & Blattner, F. R. Gene replacement without selection: regulated suppression of amber mutations in *Escherichia coli*. *Gene* **311**, 153–163 (2003).
31. Thomason, L. C., Costantino, N. & Court, D. L. *E. coli* genome manipulation by P1 transduction. *Curr. Protoc. Mol. Biol.* **79**, 1.17.1–1.17.8 (2007).
32. Li, W. et al. The EMBL-EBI bioinformatics web and programmatic tools framework. *Nucleic Acids Res.* **43**, W580–W584 (2015).
33. Sandberg, T. E., Lloyd, C. J., Palsson, B. O. & Feist, A. M. Laboratory evolution to alternating substrate environments yields distinct phenotypic and genetic adaptive strategies. *Appl. Environ. Microbiol.* **83**, e00410-17 (2017).
34. Lenski, R. E. Experimental evolution and the dynamics of adaptation and genome evolution in microbial populations. *ISME J.* **11**, 2181–2194 (2017).
35. Marotz, C. et al. DNA extraction for streamlined metagenomics of diverse environmental samples. *Biotechniques* **62**, 290–293 (2017).
36. Chen, S. et al. AfterQC: automatic filtering, trimming, error removing and quality control for fastq data. *BMC Bioinform.* **18**, 80 (2017).
37. Deatherage, D. E. & Barrick, J. E. Identification of mutations in laboratory-evolved microbes from next-generation sequencing data using *breseq*. *Methods Mol. Biol.* **1151**, 165–188 (2014).
38. Zadeh, J. N. et al. NUPACK: analysis and design of nucleic acid systems. *J. Comput. Chem.* **32**, 170–173 (2011).

Acknowledgements

This work was supported by the Novo Nordisk Foundation (grant no. NNF10CC1016517) and in part by the National Institutes of Health (grant no. R01GM057089). T.E.S. was supported in part by the National Science Foundation Graduate Research Fellowship grant no. DGE-1144086. We thank E. Brunk, E. Catoiu, M. Omar Din, C. Olson, M. Wu and Y. Hutchison for useful advice and discussions. We thank A. Feist for making automated evolution machines available for the experiments performed in this study.

Author contributions

T.E.S. and B.O.P. conceived the project and wrote the manuscript. T.E.S. and R.S. designed and constructed the strains. P.V.P. assisted with genome sequencing. T.E.S., R.S., P.V.P. and B.O.P. aided in data analysis.

Competing interests

The authors declare no competing interests.

Additional information

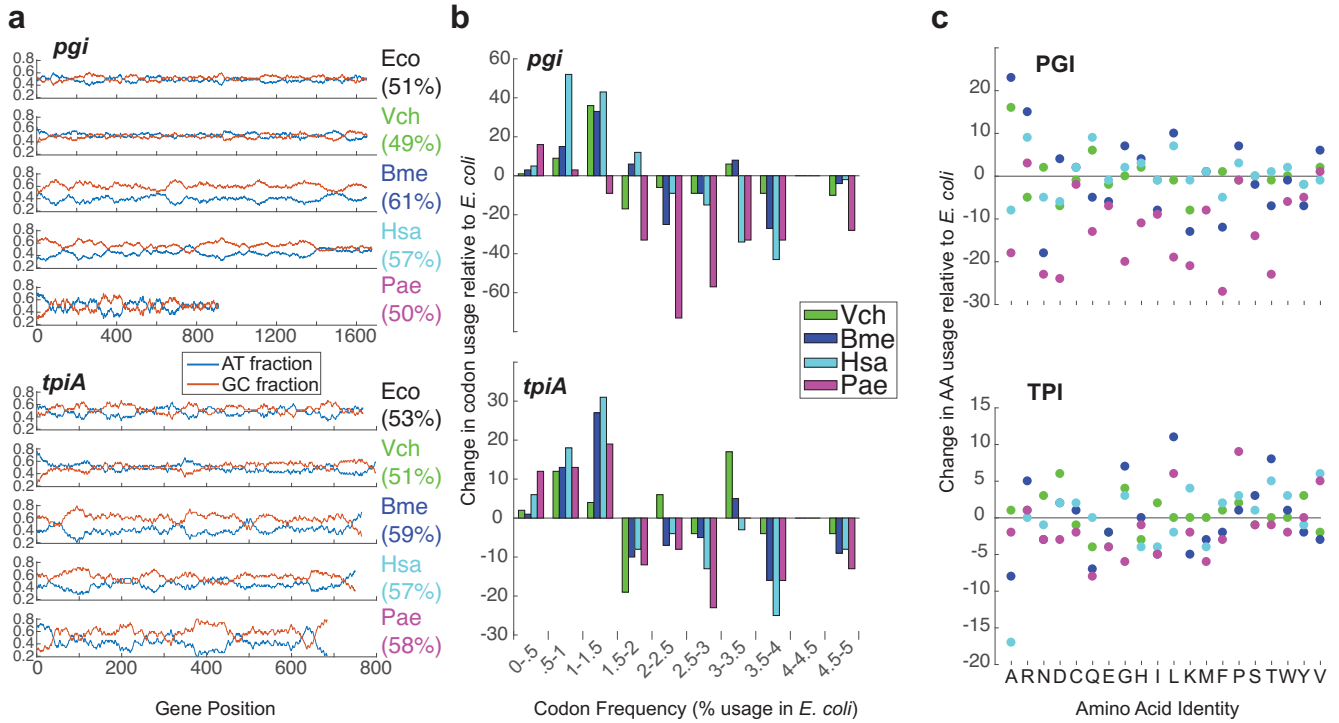
Extended data is available for this paper at https://doi.org/10.1038/s41559-020-1271-x.

Supplementary information is available for this paper at https://doi.org/10.1038/s41559-020-1271-x.

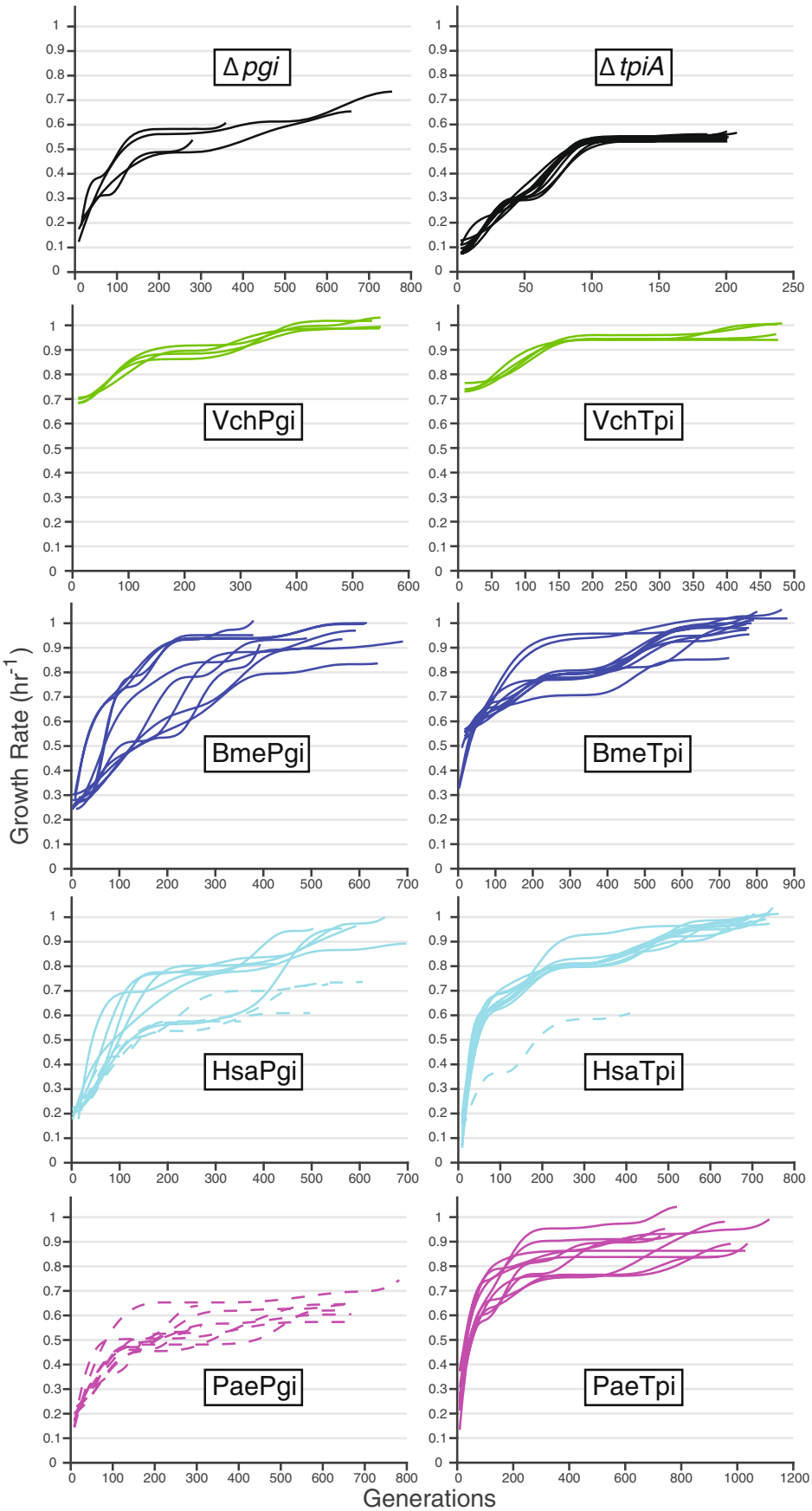
Correspondence and requests for materials should be addressed to B.O.P.

Reprints and permissions information is available at www.nature.com/reprints.

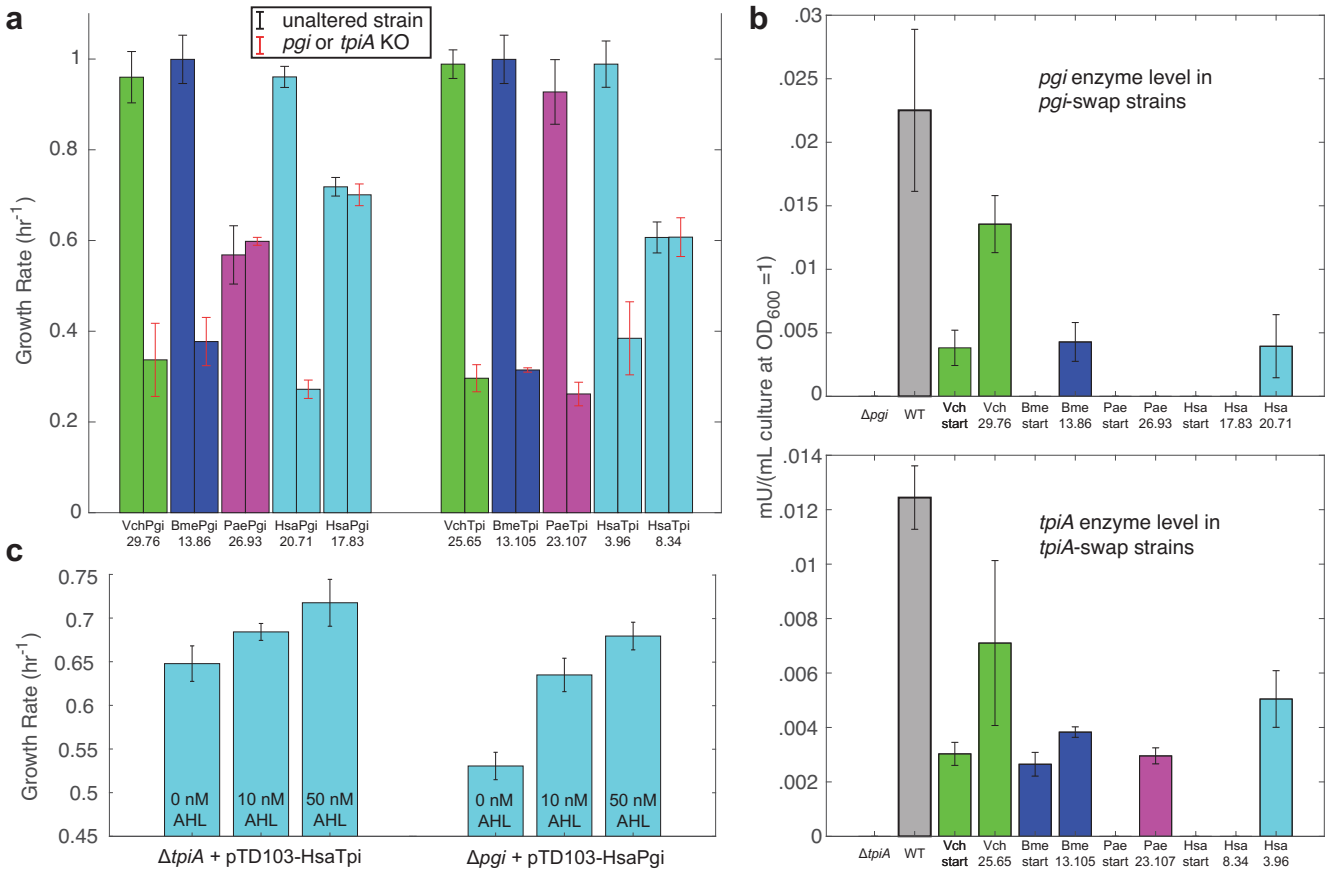
Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



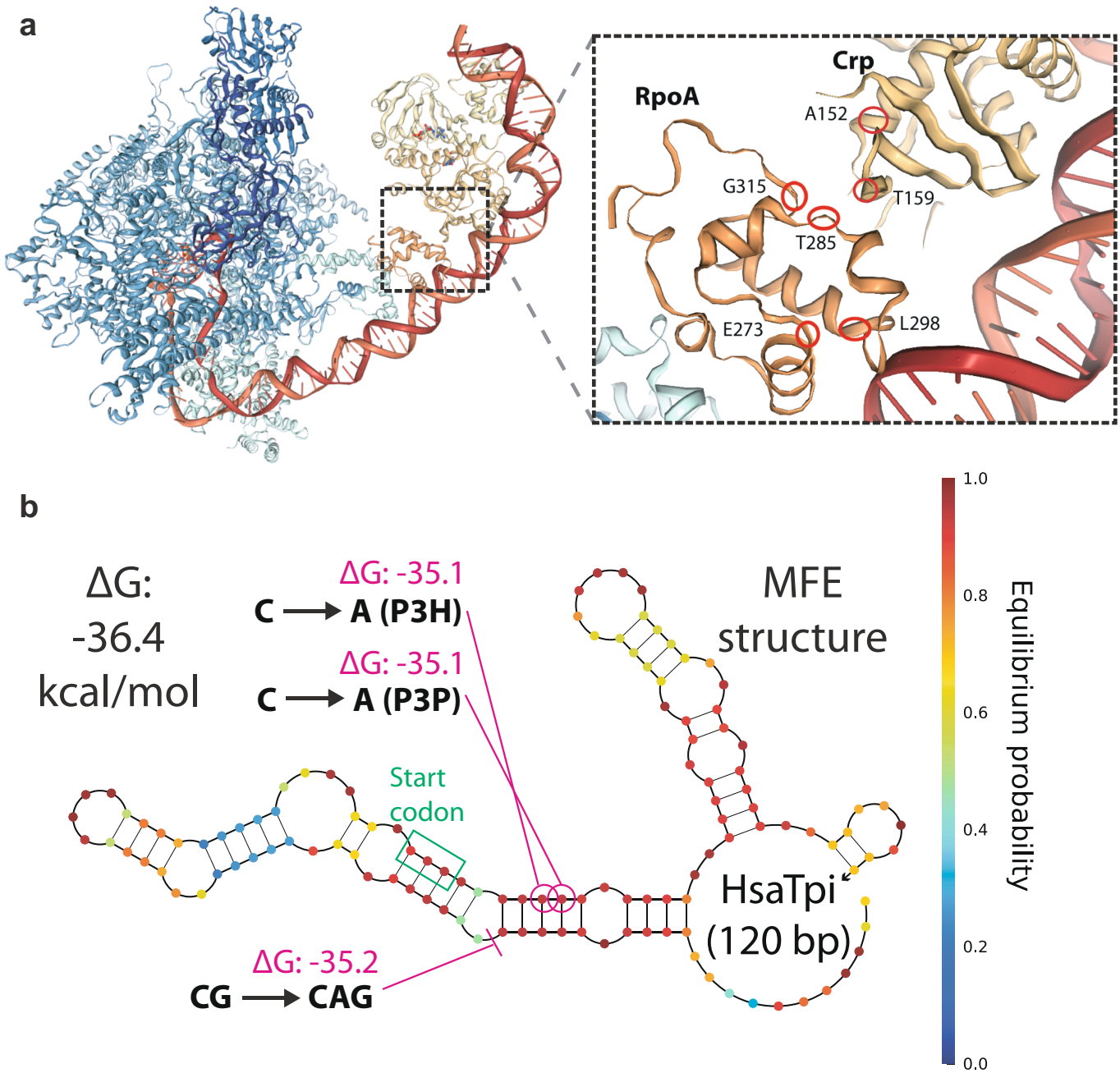
Extended Data Fig. 1 | Orthogene properties. a, GC content of native and donor gene sequences (GC% total in parentheses). **b**, Histogram of the change in codon usage resulting from replacement of native *E. coli* sequences with foreign versions. **c**, Change in protein's amino acid usage resulting from replacement of native sequences with foreign versions.



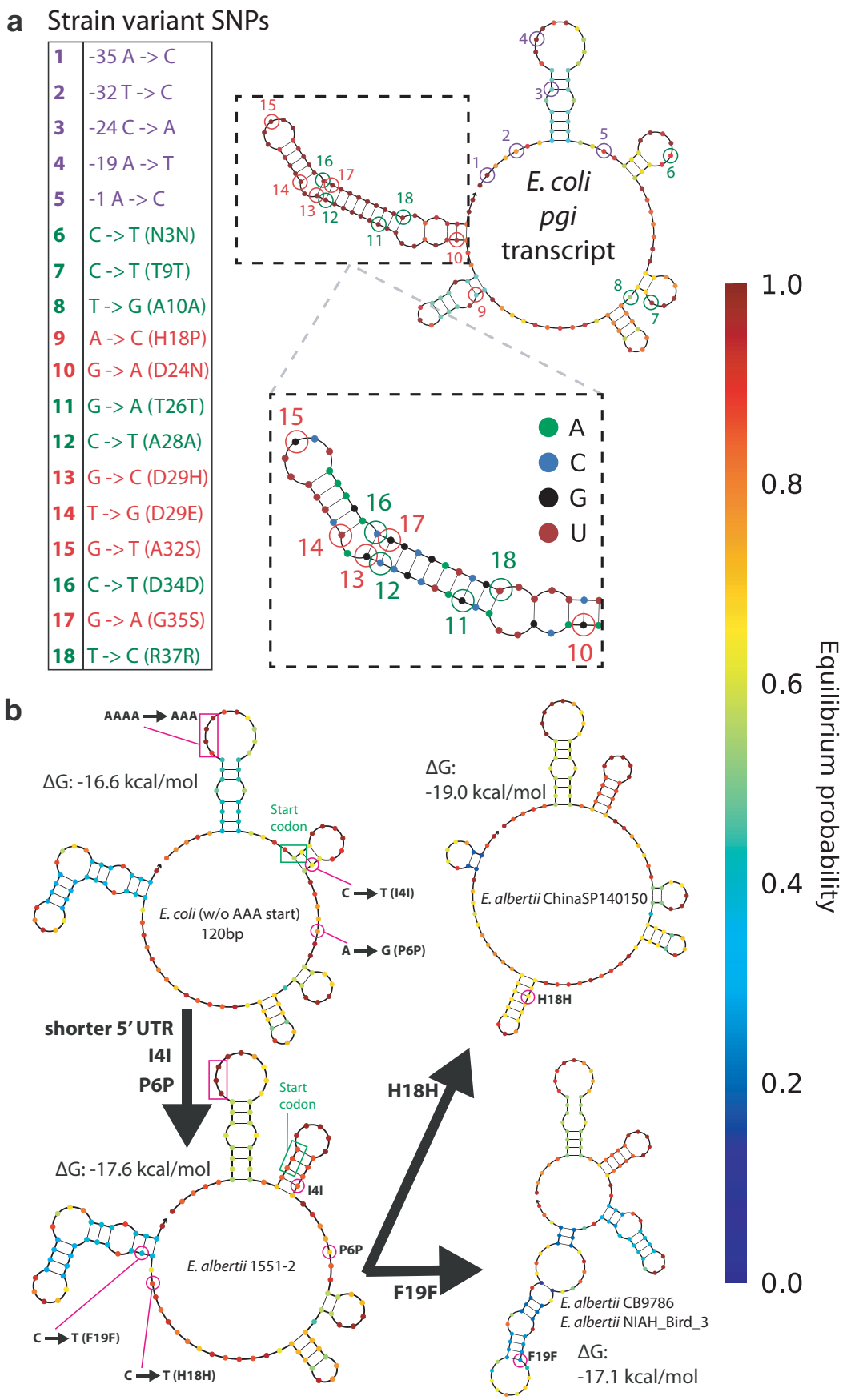
Extended Data Fig. 2 | Evolutionary trajectories. Fitness improvements over the course of evolution for knockout controls and gene-swapped strains, with failure lineages indicated by dotted lines. Δ*tpiA* controls were increased from four to ten to provide more comparison lineages for the single HsaTpi failure.



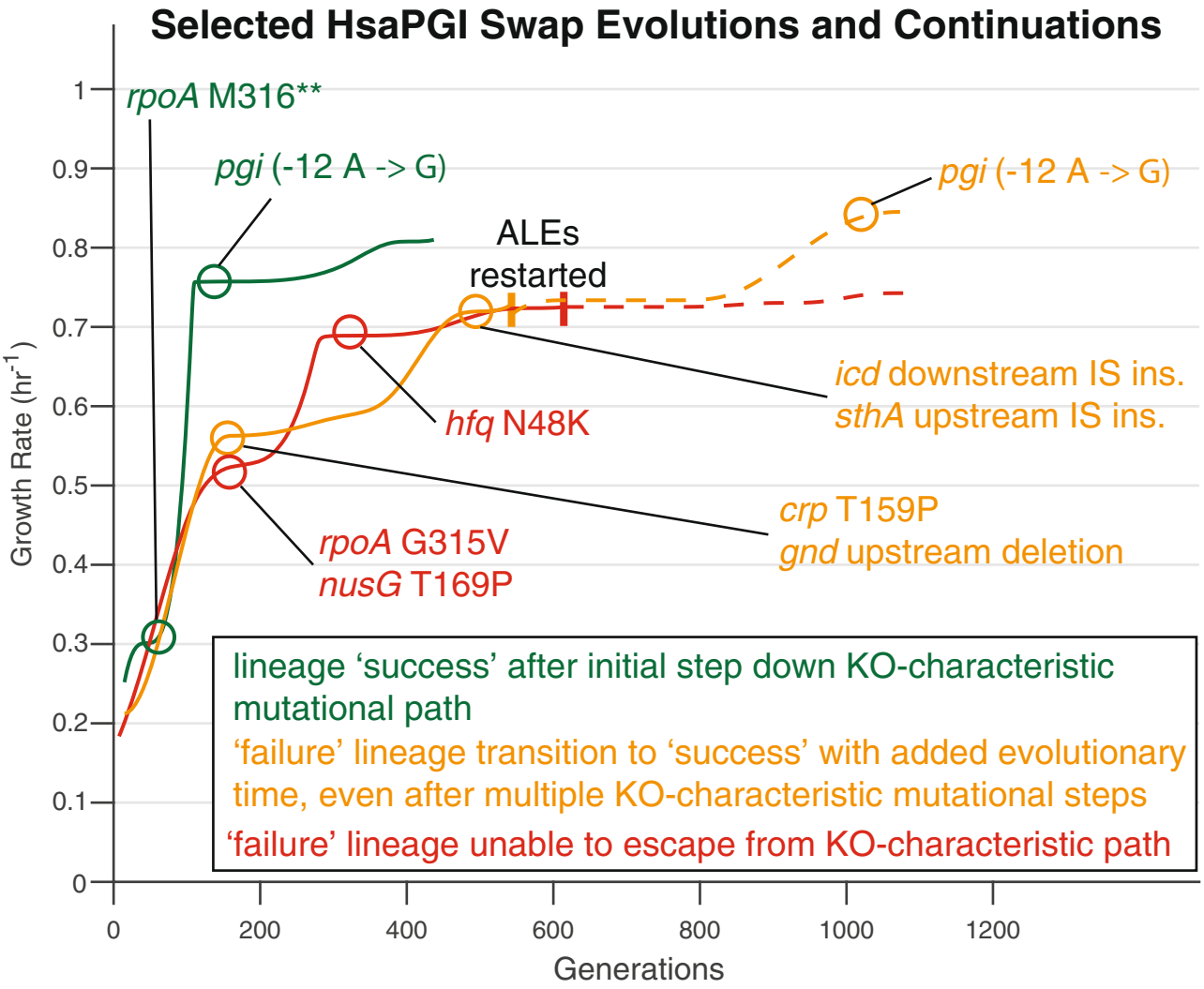
Extended Data Fig. 3 | Orthogene impact on strain fitness. a, Growth rates of various *pgi*- and *tpiA*-swap evolved strains before and after knockout of the orthogene. **b**, Enzyme activity levels of various strains, determined by colorimetric assay. No bar indicates an activity level below detection of the assay. **c**, Growth rates at various AHL concentrations of *pgi* or *tpiA* knockout strains containing plasmids with AHL-inducible expression of the *H. sapiens pgi* or *tpiA*. Growth rates higher than knockout levels even with no AHL induction may be due to leaky plasmid expression. In all panels strain names correspond with those given in the Supplementary Dataset containing DNA sequencing data, and error bars represent standard deviation from quadruplicate measurements. Orthogene-assimilation failures: PaePgi 26.93, HsaPgi 17.83, HsaTpi 8.34. Orthogene-assimilation successes: VchPgi 29.76, BmePgi 13.86, HsaPgi 20.71, VchTpi 25.65, BmeTpi 13.105, PaeTpi 23.107, HsaTpi 3.96.



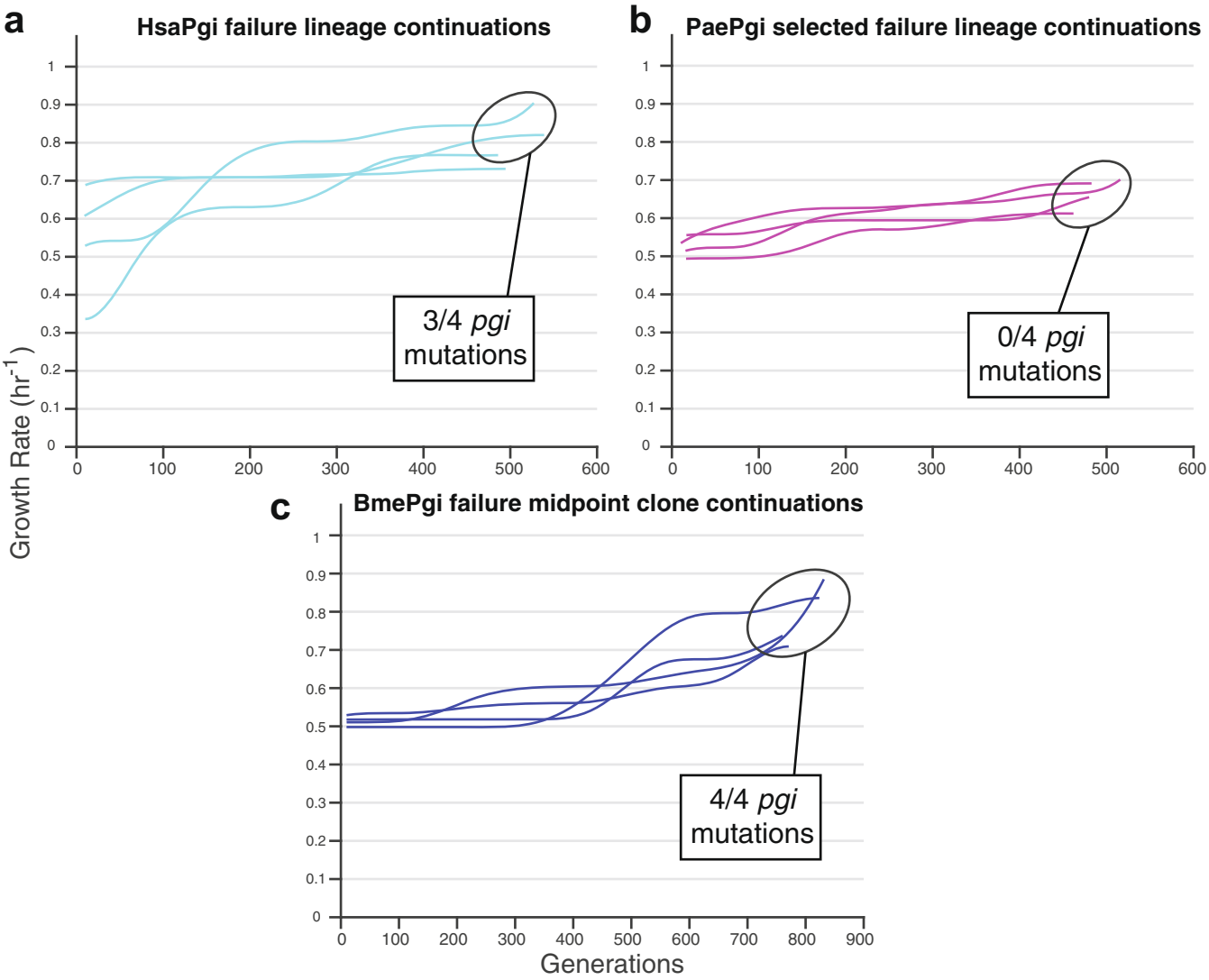
Extended Data Fig. 4 | Recurring mutations highlight regions under selection. a, Mutations to *crp* and the C-terminus of *rpoA* were highly characteristic of *pgi* failures and knockout controls. Mapping to the cryoEM structure of the transcription activation complex (PDB ID: 6B6H) reveals that these characteristic mutations cluster in the same spatial region. **b**, Minimum free energy (MFE) structure at 37 °C of *tpiA* transcript for HsaTpi swap, with observed ALE endpoint mutations. The coding sequence changes destabilize G-C rungs of the strongest stem-loop, while the 5′-UTR +A insertion destabilizes this same stem-loop via increased stabilization of a stem-loop-adjacent unstructured region.



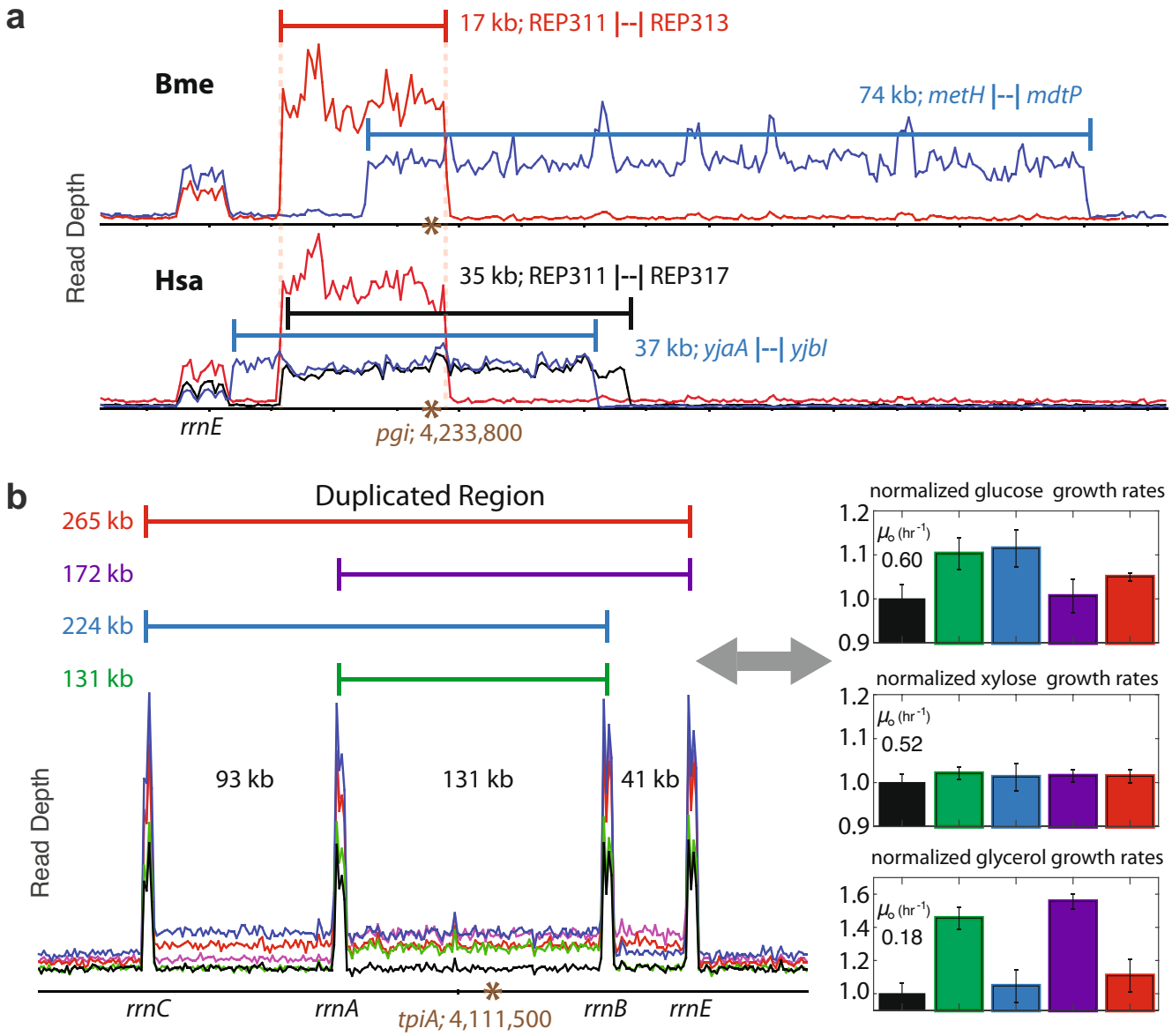
Extended Data Fig. 5 | Various *pgi* transcripts and mutations. a, SNP accumulation in *pgi* across 924 *Escherichia coli* strain variants with whole genome sequences available. **b**, The only *pgi* differences in *E. coli* and various *E. albertii* strains within the first 120 bp of transcript are minor 5'-UTR changes and a number of synonymous mutations.



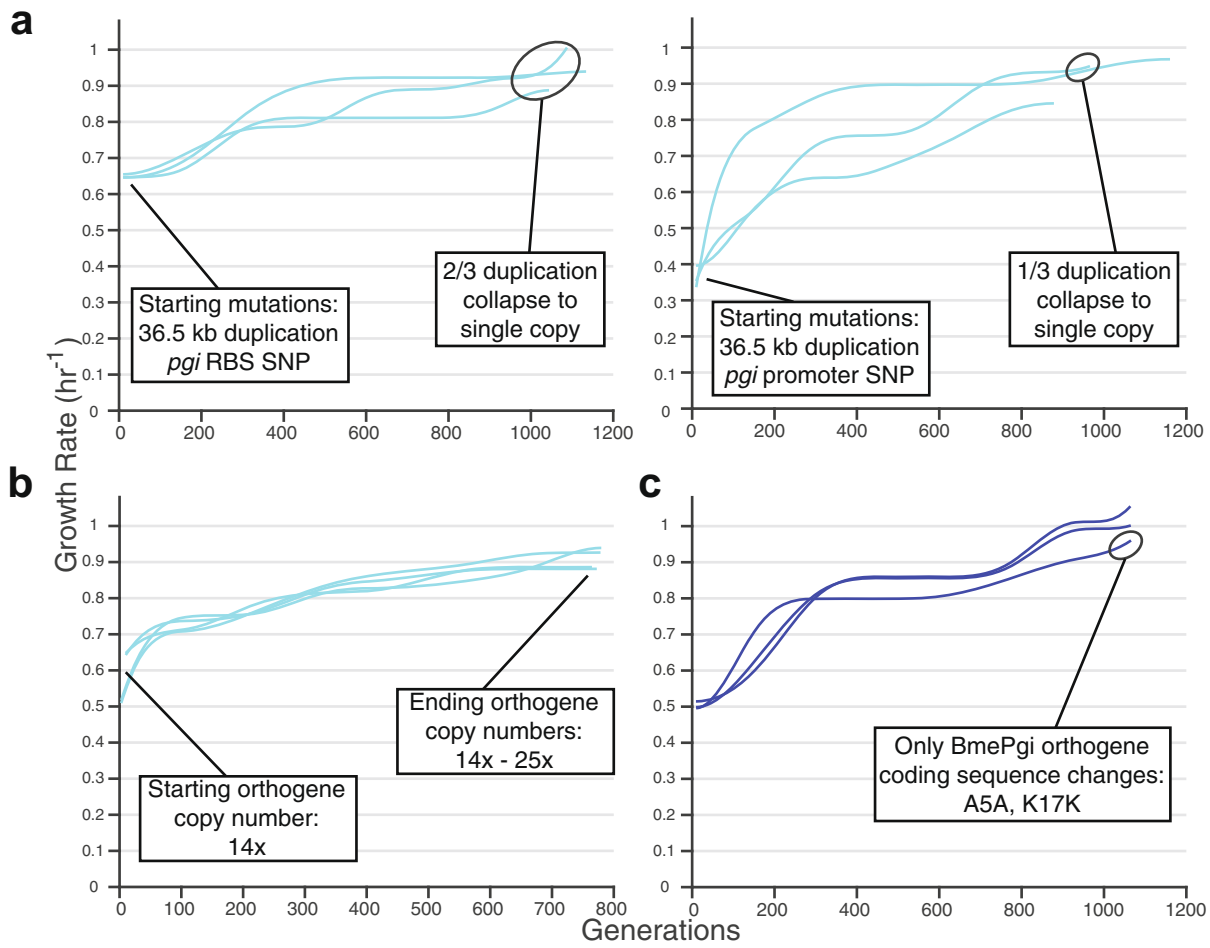
Extended Data Fig. 6 | Orthogene assimilation in the presence of knockout-characteristic mutations. Even in the presence of one or more Δ*pgi*-characteristic mutations a lineage could still successfully assimilate the orthogene, with added evolutionary time facilitating but not guaranteeing this outcome.



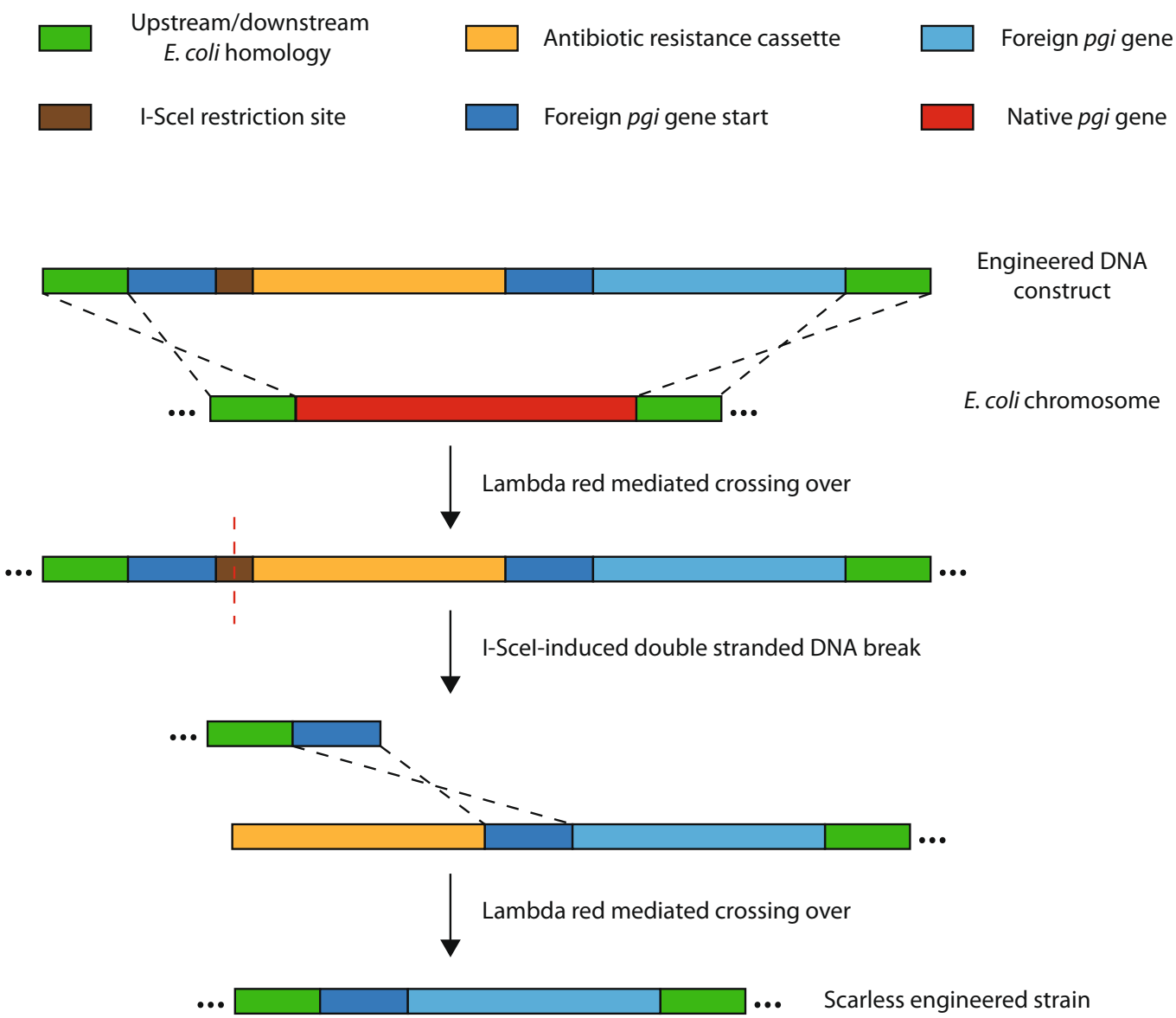
Extended Data Fig. 7 | Continuation ALEs. a, The only four failure HsaPgi lineages were given additional evolutionary time and reached success in most cases. **b**, Additional evolutionary time did not enable success for any of four continued PaePgi failure lineages. **c**, A single BmePgi midpoint clone was isolated with two knockout-characteristic mutations (to *rpoA* and *nusA*) and no orthogene changes. Independent lineages founded from this clone all eventually acquired orthogene mutations and higher growth rates.



Extended Data Fig. 8 | Orthogene copy number expansions. a, All orthogene copy number expansions found in *pgi* swap ALE strains. Homology between flanking genes or repetitive extragenic palindromic (REP) sequences facilitated the expansions. **b**, Relative growth rates on various carbon sources for HsaTpi strains genetically identical except for the size of *tpiA* duplication. Inclusion of the *rrnB-rrnE* region hindered growth on glucose, while the *rrnC-rrnA* expansion hindered glycerol growth. Error bars represent standard deviation from quadruplicate measurements.



Extended Data Fig. 9 | Replay ALEs. a, Two distinct HsaPgi strains with the same *gnd*- containing genome duplication but different orthogene mutations were split into lineages and evolved, which could lead to collapse of the *gnd* duplication. **b**, An HsaPgi strain that had acquired a genomic copy number increase of the orthogene was used to found four lineages. Evolution resulted in copy number remaining stable or increasing. **c**, A BmePgi strain with a synonymous orthogene SNP (A5A) was used to found three lineages. Evolution enabled further orthogene-upregulating mutations to increase growth rate, and one lineage acquired a second synonymous SNP as the only coding sequence changes to the foreign DNA.



Extended Data Fig. 10 | Method of scarless strain construction. The *pgi* swap strains were constructed as shown. The *tpiA* swap strains used a construct lacking the foreign gene start homology, I-SceI site, and antibiotic cassette; double stranded DNA breaks were induced by CRISPR-targeting to native gene sequence, obviating the need for antibiotics.

nature research

Corresponding author(s): Bernhard O. Palsson

Last updated by author(s): Jun 24, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☒	☐	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	The process control software running the automated experimental evolution machines was custom written in MATLAB R2017a.
Data analysis	AfterQC, the software used to trim and filter DNAseq reads, is available at https://github.com/OpenGene/AfterQC. Breseq, the software used to identify mutations, is available at https://github.com/barricklab/breseq. Co, the software used to edit genome references (https://github.com/SBRG/svns_refseq), is available at https://github.com/biosustain/co.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The genome sequence data that support the findings of this study are available from ALEdb (https://aledb.org) under project name 'SvNS.'

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☐ Behavioural & social sciences ☒ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	Experimental laboratory evolution of a number of engineered E. coli strains. Evolved to select for improved exponential phase growth on glucose minimal media at 37C.
Research sample	The model organism E. coli K12 MG1655 was used, with heterologous engineering performed using orthologous DNA from a range of species.
Sampling strategy	Sample sizes of 10 independent evolutionary replicates for each strain exhibiting initial growth rate lower than wild-type, and 4 trajectories for strains which did not exhibit fitness defects, were somewhat arbitrarily chosen given constraints on the automated evolution machine system throughput.
Data collection	Real-time growth rate data were collected over the course of evolution by the automated system. Frozen stocks of evolving populations were saved whenever a jump in growth rate of the population became apparent.
Timing and spatial scale	Evolution experiments were run for 30 days so as to select for the most significant large-effect beneficial mutations. Depending on strain growth rates and evolutionary trajectories this corresponded with ~300-1000 generations.
Data exclusions	No data were excluded from the analyses.
Reproducibility	The stochastic nature makes experimental evolution inherently unreproducible, though many reproducible adaptive mutations were observed.
Randomization	Randomization not relevant to microbial experimental evolution.
Blinding	Blinding not relevant to microbial experimental evolution.

Did the study involve field work? ☐ Yes ☒ No

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging