Profiling the initial burst of beneficial genetic diversity to anticipate evolution of a cell population

Daniel E. Deatherage and Jeffrey E. Barrick*

Department of Molecular Biosciences, Center for Systems and Synthetic Biology,

The University of Texas at Austin, Austin, Texas 78712, U.S.A.

* Email: jbarrick@cm.utexas.edu

Data Availability Statement: DNA sequence files are available from the NCBI Sequence Read Archive (accession number PRJNA601748). Code and processed data files are available on GitHub (https://github.com/barricklab/adaptome-capture).

Funding: This work was supported by the Cancer Prevention & Research Institute of Texas (CPRIT) (RP130124), the National Institutes of Health (R00-GM087550), the National Science Foundation (CBET-1554179 and DEB-1813069), and the NSF BEACON Center for the Study of Evolution in Action (DBI-0939454). D.E.D. acknowledges support from a University of Texas at Austin CPRIT research traineeship (RP101501). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: J.E.B. is the owner of Evolvomics LLC. D.E.D. has been a paid consultant for Evolvomics LLC.

Abstract

Clonal populations of cells continuously evolve new genetic diversity, but it takes a significant amount of time for the progeny of a single cell with a new beneficial mutation to outstrip both its ancestor and competitors to fully dominate a population. If genotypes with these driver mutations can be discovered earlier—while they are still extremely rare—it may be possible to anticipate the future evolution of these populations. For example, one could diagnose the likely course of incipient diseases, such as cancer and bacterial infections, and better judge which treatments will be effective, by tracking rare drug-resistant variants. To test this approach, we replayed the first 500 generations of a >70,000-generation *Escherichia coli* experiment and examined the trajectories of new mutations in eight genes known to be under positive selection in this environment in six populations. By employing a deep sequencing procedure using molecular indexes and target enrichment we were able to track 236 beneficial mutations at frequencies as low as 0.01% and infer selection coefficients for 180 of these. Distinct molecular signatures of selection on protein structure and function were evident for the three genes in which beneficial mutations were most common (*nadR*, *pykF*, and *topA*). We detected mutations hundreds of generations before they became dominant and tracked beneficial alleles in genes that were not mutated in the long-term experiment until thousands of generations had passed. Therefore, this targeted adaptome sequencing approach can function as an early warning system to inform interventions that aim to prevent undesirable evolution.

Introduction

New genetic variation naturally arises in lineages of cells and organisms during genome replication and repair. These *de novo* mutations are the main drivers of adaptive evolution in many populations, particularly those with little or no recombination or standing genetic variation. In large laboratory populations of asexual microbes, numerous lineages with different beneficial mutations arise and contend within a population before any one outcompetes the ancestor and its competitors [1–3]. This 'clonal interference' leads to heterogeneous populations with many lineages simultaneously adapting via distinct sets of mutations, though often these mutations are in a small subset of genes that are under the strongest selection [4–6].

In human cancers and chronic microbial infections, single cells clonally expand in a similar fashion by evolving driver mutations that lead to disease progression and drug resistance. Both solid tumors and blood cancers have been shown to be genetically heterogeneous [7–9]. *De novo* mutations within these cell populations are responsible for neoplastic progression [10], differences in responses to chemotherapy [11], and relapse [12]. Populations of *Pseudomonas aeruginosa* and other bacteria that persistently infect the lungs of cystic fibrosis patients become increasingly invasive and antibiotic resistant over time [13–15]. In these cases, there are also specific genetic loci that are repeatedly mutated in different individuals. Better predicting the future evolution of each of these types of cell populations and others would inform treatment decisions and improve medical outcomes.

Cells used in biomanufacturing are also prone to evolving unwanted genetic heterogeneity [16,17]. Typically, these cells have been heavily engineered to optimize the titer of a product of interest at the expense of rapid cellular replication [18,19]. Therefore, there are strong selective pressures for 'escape mutations' that cause production to decline. Usually escape mutations

directly inactivate one or more key genes in the engineered pathway. The resulting nonproducing cells can become dominant during the many cell divisions that are necessary to scale these processes up to large bioreactors [20–22]. The ability to predict the future evolution of nonproducing cells before attempting scale-up could guide strain design decisions and thereby improve the efficiency of industrial processes.

Evolution experiments conducted in controlled laboratory environments reproduce key aspects of microbial evolution that are observed in chronic infections and bioreactors [23,24]. In theory, profiling rare mutations in the earliest stages of clonal interference using high-throughput DNA sequencing should allow one to anticipate the future evolution of these populations. However, these studies have generally been limited by sequencing depth and per-base error rates to reliably identifying mutations that are present in at least one sample at a frequency above ~1-10% when they have already succeeded in becoming dominant [1,3,25,26]. Theory and simulations predict that many more highly beneficial mutations evolve in these populations but never reach such high frequencies before they are driven extinct [4,6], and recent studies that track the evolution of barcoded lineages of microbes show that this is the case [27,28].

Here, we used methods for selectively increasing sequencing depth and lowering sequencing error rates to deeply profile the initial burst of rare beneficial mutations in laboratory populations of *E. coli*. We directly identified diverse beneficial mutations in six genes when they were orders of magnitude lower in frequency and hundreds of generations earlier than could be accomplished by standard metagenomic sequencing methods. By comparing our results to the long history of a >70,000-generation *E. coli* evolution experiment that used the same ancestral strains and nearly identical culture conditions [29], we evaluate the potential of this type of targeted adaptome analysis for anticipating the future evolution of cell populations.

Results

Replaying the beginning of a long-term evolution experiment

We tracked new mutations in nine replicate *E. coli* populations that were propagated via daily serial transfers in glucose-limited minimal medium for 500 generations. Our experiment used the same *E. coli* strains as the Lenski long-term evolution experiment (LTEE) and similar growth conditions (see **Methods**). Each population was inoculated with a 50/50 mixture of the two neutrally marked LTEE ancestor strains to visualize the initial selective sweep [30]. Most populations maintained a roughly equal representation of descendants of both ancestral strains through the first 300 generations (**Fig. 1**). These dynamics are in agreement with what has previously been observed in studies of the LTEE, where few mutations reach a high frequency in the first few hundred generations of evolution [3].

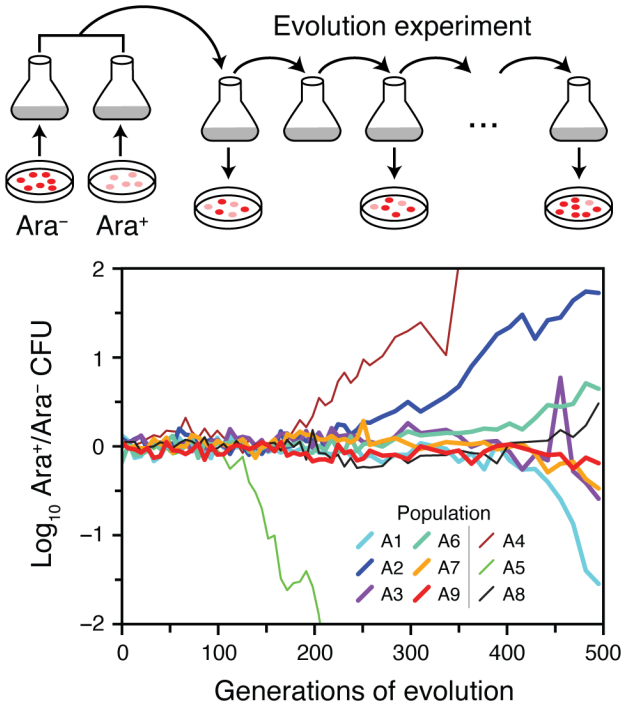

Figure 1. Replaying the first selective sweep of a long-term evolution experiment. Nine *E. coli* populations were initiated from equal mixtures of two variants of the ancestral strain that differ in a neutral genetic marker for arabinose utilization (Ara). We observed the evolutionary dynamics of these populations over ~500 generations of regrowth from 75 daily 1:100 serial transfers by periodically plating dilutions of each population on indicator agar. The ratio of Ara^+ cells (pink colonies) to Ara^- cells (red colonies) diverges from 1:1 when descendants of one ancestor type accumulate enough of a fitness advantage due to *de novo* beneficial mutations that they take over. We focused further analysis on six of the nine populations (thick lines).

Reconstructing the trajectories of new beneficial mutations

We next performed deep sequencing of eight genes at ~25 generation increments over all 500 generations of the evolution experiment for four of the nine populations. These eight genes (*nadR*, *pykF*, *topA*, *spoT*, *fabR*, *ybaL*, *hslU*, and *iclR*) are known to be targets of selection in the LTEE [3,31]. Illumina libraries containing molecular indexes [32] were prepared for sequencing and enriched for the regions of interest using solution based hybridization [33]. Consensus sequence reads were generated based on groups of reads with identical molecular indexes and aligned to the *E. coli* genome to predict mutations, including using split-read mapping to identify transposon insertions and large deletions (**Fig. 2A**). The enrichment procedure was effective. In the sample with the median number of total consensus reads, the average coverage depth across each of the eight genes of interest exceeded 5,000 (**Fig. 2B**). After analyzing patterns in mutation frequencies over time to eliminate other systematic biases (see Methods), we were able to track the evolution and competition of 180 *de novo* mutations, including when many were present in less than 0.1% of the cells in a population (**Fig. 2C**, **Fig. 3**).

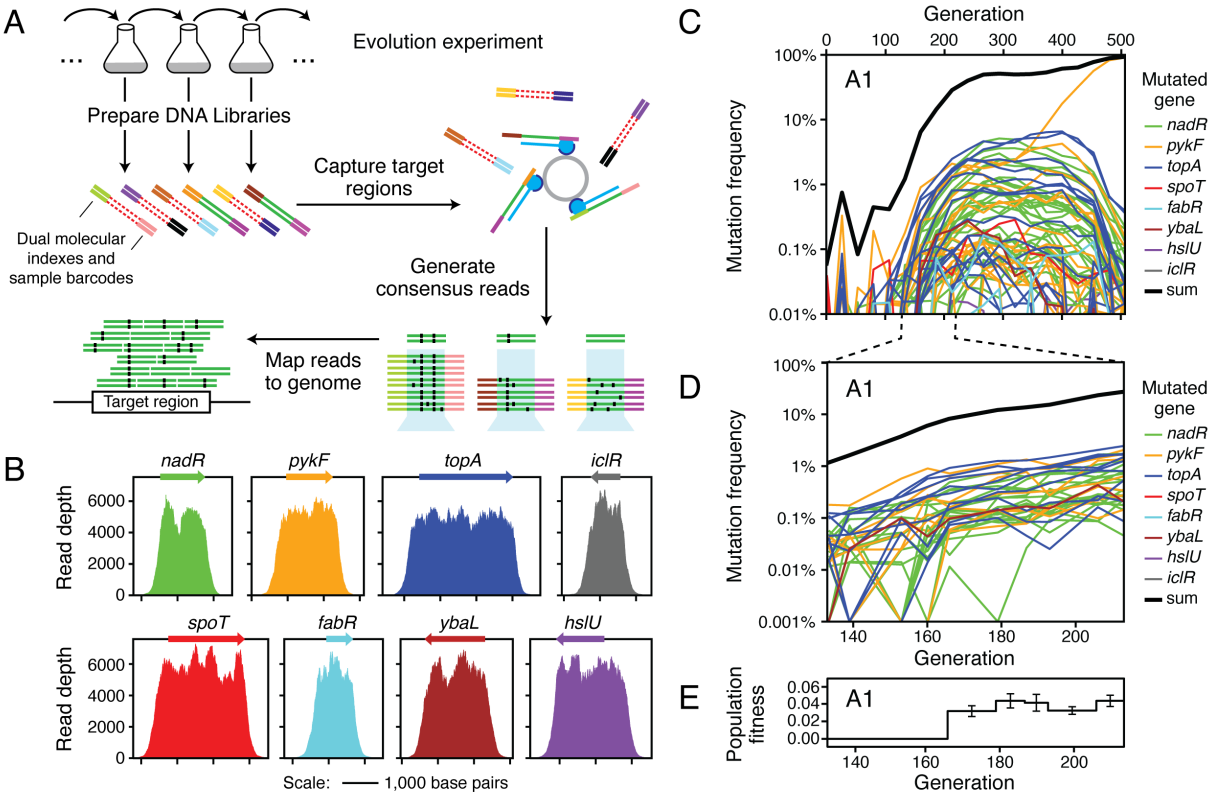

Figure 2. Profiling many beneficial mutations in the first selective sweep by deep sequencing. (A) Schematic of the deep sequencing approach. Genomic DNA is directly isolated from the *E. coli* populations and prepared for Illumina sequencing with unique molecular indexes (colored ends attached to red/green double stranded DNA). DNA fragments matching the targeted genome regions (green centers) are captured by probes (blue) bound to magnetic beads and other sequences are washed away (red centers). Reads with the same unique molecular index, which were amplified from the same original genomic DNA fragment, are used to construct a consensus read to eliminate sequencing errors. Consensus reads are mapped to the reference genome to call sequence variants. (B) Enrichment of reads mapping to eight genes known to be early targets of selection in this environment from the long-term evolution experiment. The final coverage depth of consensus reads in and around these genes is shown for a typical sample (population A7 at 500 generations). (C) Frequency trajectories for mutations in the eight targeted genes as well as the sum total frequency in population A1 over the complete time course of the evolution experiment. When a mutation was not detected at a time point, its trajectory is shown as passing through a frequency of 0.0001% (outside of the graphed region). (D) Mutation frequency trajectories for population A1 during the window from 133 to 213 generations when mutations were first reaching detectable frequencies as they outcompeted the ancestral genotype. At time points when a mutation was not detected, its frequency is shown as 0.001% (at the bottom of the plot). (E) Estimate of average population fitness between the time points in the window when mutations were first detected. The frequency trajectories of the beneficial mutations in the initial sweep shown in D were used to jointly estimate population fitness and the individual selection coefficients of each mutation. Error bars are 95% confidence intervals on fitness estimations.

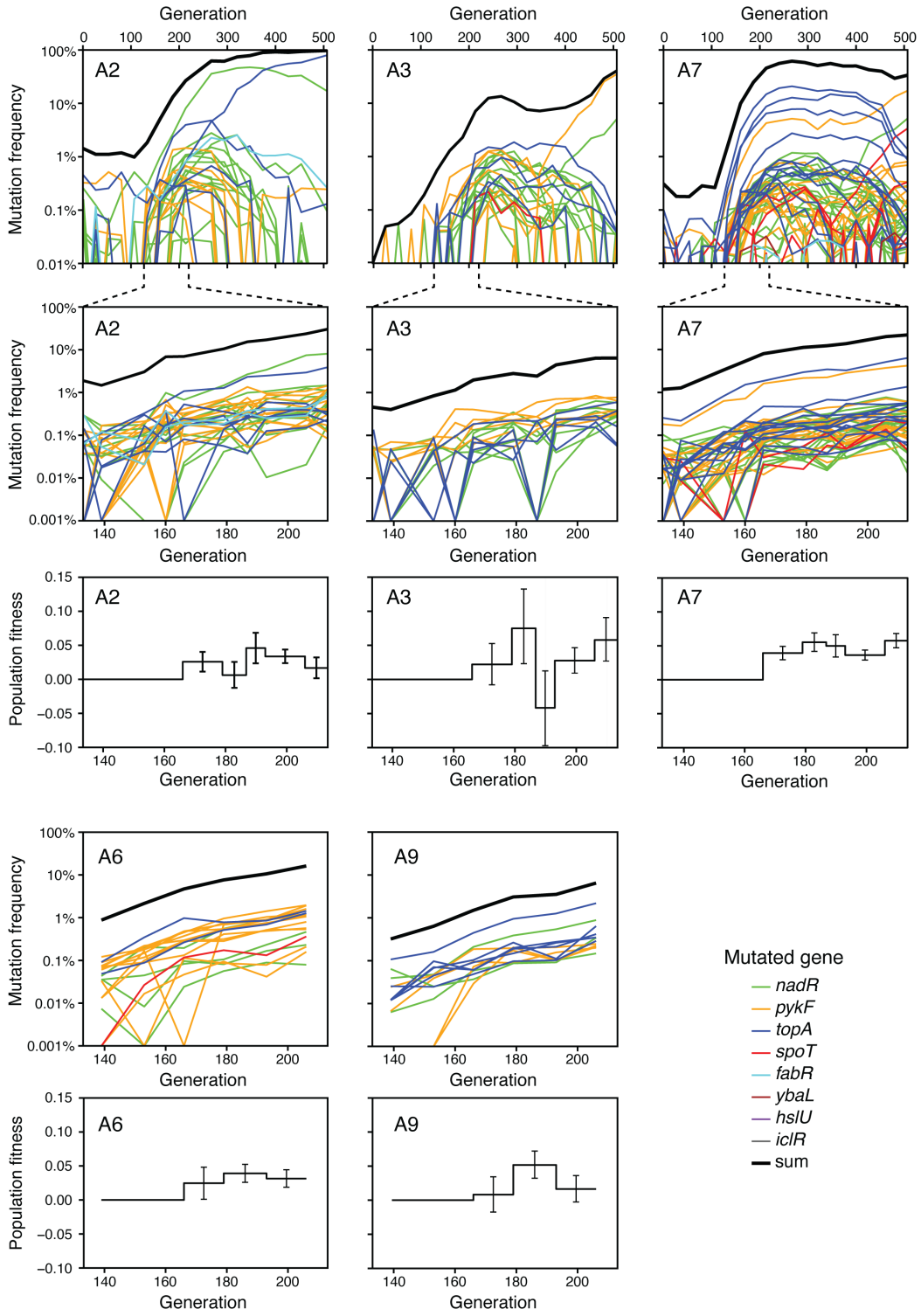

Figure 3. Frequency trajectories of mutations in the remaining populations. The same plots described in Figure 2C-E for population A1 are shown for populations A2, A3, and A7 (top three sets of panels). For populations A6 and A9, sequencing was only performed at time points during the selective sweep window so only the plots corresponding to Figure 2D-E are shown (bottom two sets of panels).

Mutation trajectories in all four populations exhibited a burst of genetic diversity in the targeted genes followed by loss of this diversity. The initial dynamics are expected to be largely driven by new genotypes that each evolve a single beneficial mutation very early in the experiment. If their descendants escape stochastic loss, they will gradually increase in frequency over the first few hundred generations as they outcompete the ancestral genotype. Once the population becomes dominated by these first-step mutants, their frequency trajectories plateau because of clonal interference: they are now mainly competing against one another and are relatively evenly matched. In populations A1, A2, and A7, the total frequencies of the mutations we identified sums to 50-62% at generation 270, indicating that each population is mostly composed of genotypes with a single mutation in one of the focal genes. We recovered less of the initial beneficial mutation diversity in population A3 where this sum was only 13%.

After around 300 generations, there is a steady decline in the frequencies of most mutations in the eight targeted genes. At this point, new more-fit genotypes that have evolved from the single-step mutants begin to exert their influence and displace them. Many of the most successful second-step genotypes are descended from cells that already have a mutation in one of the targeted genes. The original mutations serve as markers for the further expansion of these subpopulations after a period during which their frequencies stagnate or decrease, but the new beneficial mutations responsible for this further increase in fitness are outside of the genomic regions we surveilled. The converse situation, in which a beneficial mutation in one of the eight focal genes appears in a cell with an untracked beneficial mutation elsewhere in the genome, also occurs in a few cases. Most strikingly, a new mutation in *pykF* that only appears after 300 generations in population A3 rapidly increases in frequency and becomes dominant, strongly suggesting that it appeared in a genetic background with a prior, unknown beneficial mutation.

180

181 **Selection coefficients can be inferred from initial mutation trajectories**

182 We next sought to calculate the fitness benefits of individual mutations by tracking how rapidly
183 their frequencies rose early in the experiment when they were largely competing versus the
184 ancestral genotype because all new mutations in the population were still rare. To that end, we
185 performed additional sequencing on six populations (including the four already sequenced at 25
186 generation increments) at ~13-generation increments in a time window from 133 to 213
187 generations (**Fig. 2D**, **Fig. 3**). We were able to track a total of 161 mutations during this time,
188 including 56 that were not identified in the complete time courses. More than 95% of these
189 mutations occurred in just three of the targeted genes: *nadR*, *pykF*, and *topA* **(Fig. 4A)**.

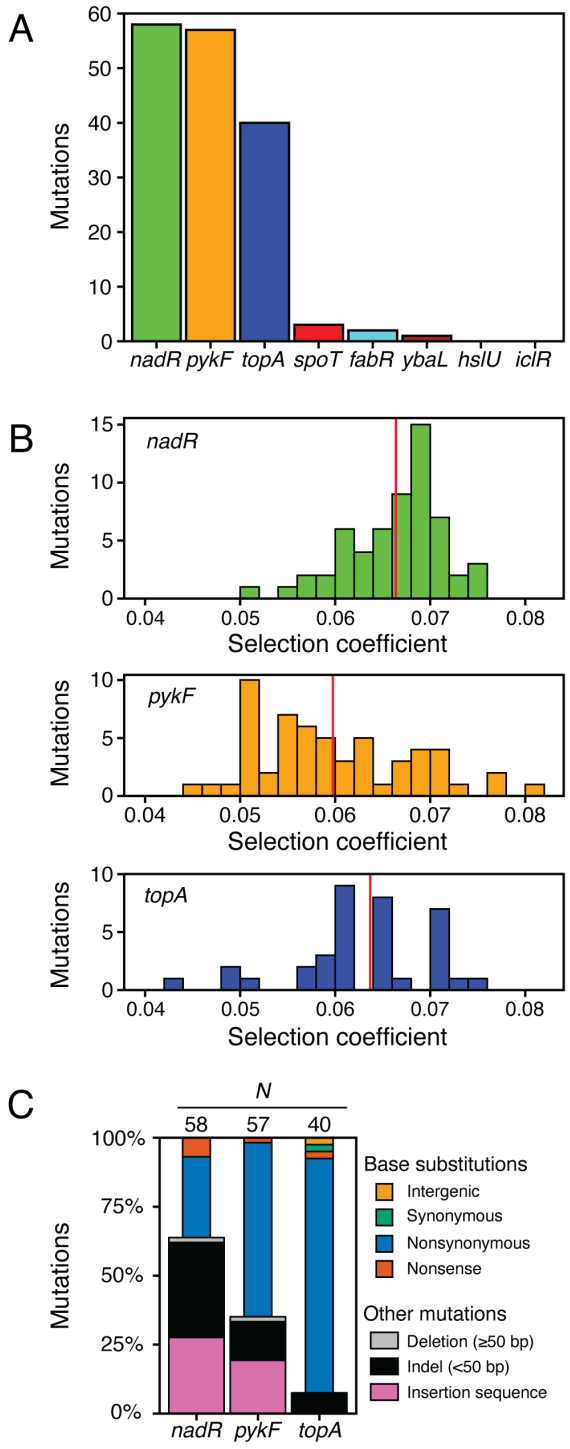

Figure 4. Characteristics of beneficial mutations in the initial selective sweep. (A) Total number of beneficial mutations in each targeted gene identified in the window time courses from 133 and 213 generations for all six *E. coli* populations that were profiled. (B) Distribution of selection coefficients of beneficial mutations determined from the window time courses in the three genes that were the dominant targets of selection. Vertical red lines show the mean of each distribution. (C) Spectrum of beneficial mutation types in the three genes that were the dominant targets of selection in the window time courses.

We were able to estimate a selection coefficient for each of the 161 mutations predicted in the window time courses by fitting a binomial logistic model to the counts of reads supporting the variant versus reference sequences over time. In all populations, there is a slight deceleration in the rate at which the frequencies of the new mutations increase at generation 166 and later that coincides with the onset of clonal interference (**Fig. 2E**, **Fig. 3**), At this point, genotypes with beneficial mutations begin to make up an appreciable fraction of the population and compete against one another rather than effectively only versus their ancestor. After correcting for this increase in overall population fitness (see Methods), the mean selection coefficient that we inferred for the tracked mutations in all six populations was 6.32% with a standard deviation of 0.74%. Although the distributions of selection coefficients estimated for mutations in *nadR*, *pykF*, and *topA* overlap **(Fig. 4B)**, there was a significant stratification among these genes. Mutations in *nadR* were 0.27% more beneficial than mutations in *topA*, on average, and this difference was significant ($p = 0.024$, Kolmogorov-Smirnov test). In turn, mutations in *topA* were 0.39% more beneficial than those in *pykF* ($p = 0.00035$, Kolmogorov-Smirnov test). The six mutations in other genes (*spoT*, *yijC,* and *ybaL*) for which we were able to estimate selection coefficients were roughly as beneficial as mutations in *nadR*, *pykF*, and *topA*.

Beneficial mutations reveal different signatures of selection on gene function

Of the 236 mutations that we were able to track in complete or window time courses, 218 were in the *nadR*, *pykF*, or *topA* genes. The large sets of beneficial mutations in these genes gave us the statistical power to test for several signatures of molecular evolution to predict what types of changes in the function of each gene improved *E. coli* fitness in this environment. Each of the three genes exhibited a distinct spectrum of beneficial mutations (**Fig. 4C**). In some cases,

different types of mutations were also unevenly distributed throughout the sequences of these three commonly hit genes and had noticeably different effects on bacterial fitness (**Fig. 5A**).

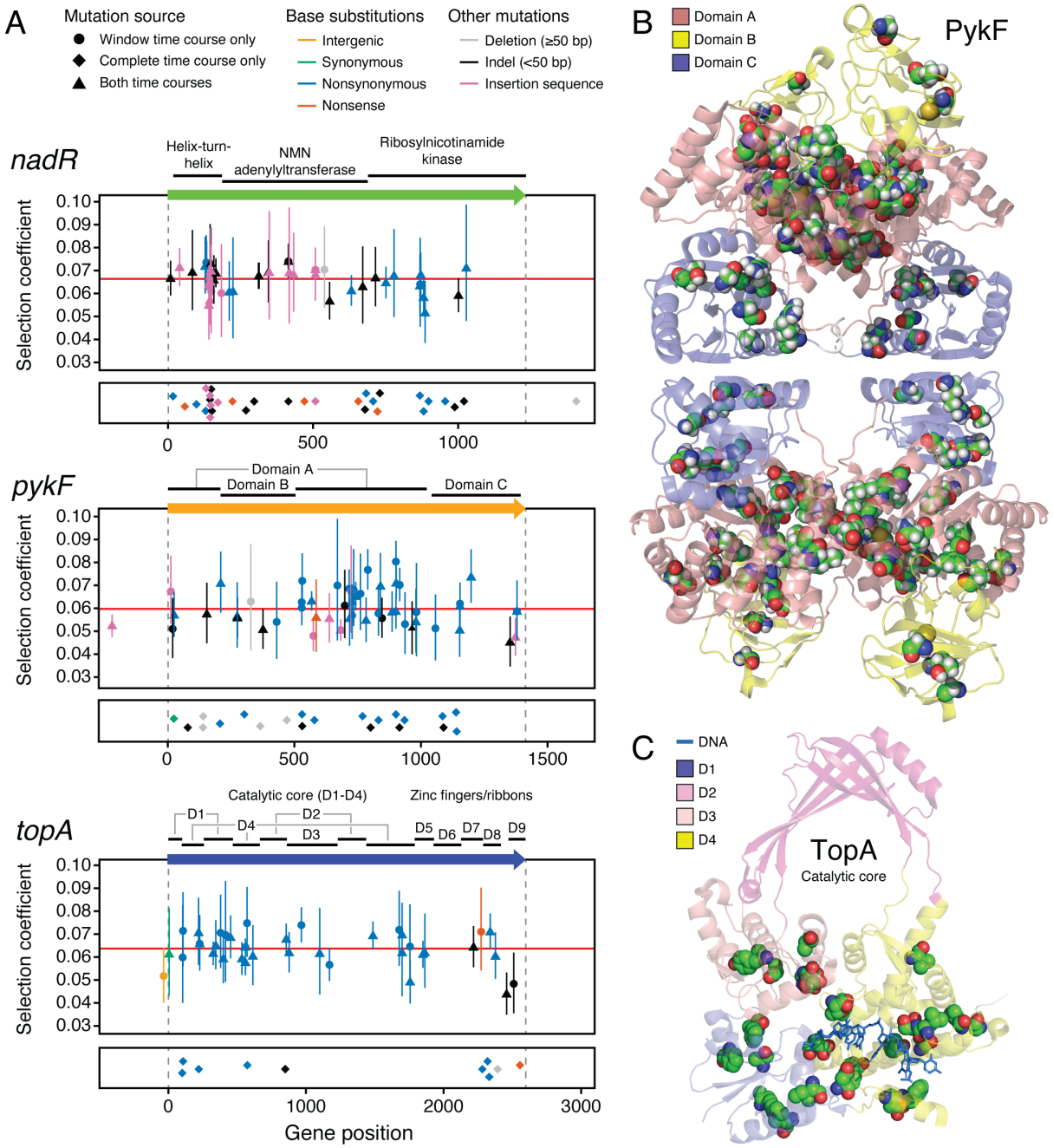

Figure 5. Mutations in the three genes that were the dominant targets of selection. (A) Locations and properties of all mutations found in each of the three genes that were the dominant targets of selection during the evolution experiment. Colors represent the type of mutation. Symbols indicate whether each mutation was detected in the window time course, the complete time course, or both. The reading frame of each gene is shown above each panel with protein domains labeled. Vertical dashed grey lines represent the start and end of each gene. Error bars

are 95% confidence limits on selection coefficients determined for the mutations detected in the window time courses. Horizontal red lines represent the average selection coefficient for all mutations in a gene. Mutations that were only detected in the complete time course are shown in the band below each graph because they do not have estimated selection coefficients. (B) Structural context of mutations in PykF. Sites of nonsymonymous mutations are highlighted by showing space-filling models of the substituted amino acid residues. All four subunits of the PykF homotetramer are shown. (C) Structural context of mutations in the catalytic core of TopA. Sites of nonsymonymous mutations are highlighted by showing space-filling models of the substituted amino acid residues. Only domains D1-D4 are present in the structure. The DNA strand interacting with TopA is shown as a stick model.

The *E. coli nadR* gene has three distinct functions related to NAD biosynthesis: (1) the N-terminal domain is a helix-turn-helix that binds to DNA so that it can act as a negative transcriptional regulator of NAD salvage and transport pathways; (2) the internal domain is an NMN adenylyltransferase [34]; and (3) the C-terminal domain is predicted to have ribosylnicotinamide kinase activity [35]. Large deletions, frameshifts from small insertions or deletions (indels), insertions of transposable insertion sequence (IS) elements, and base substitutions creating stop codons dominate the *nadR* mutational spectrum (**Fig. 4C**). These disruptive mutations, which are expected to result in complete loss of gene function, are significantly overrepresented versus nonsynonymous base substitutions in the first two domains of the gene compared to the remainder (11.4 odds ratio, $p = 4.2 \times 10^{-6}$, one-tailed Fisher's exact test) (**Fig. 5A**). Yet, there is no evidence of a greater selection coefficient for disruptive mutations compared to nonsynonymous mutations overall ($p = 0.19$, one-tailed Kolmogorov-Smirnov test). These results suggest that complete inactivation of *nadR* yields the maximum benefit possible for a mutation in this gene, through disrupting all three of its distinct functions may not be necessary for achieving this full benefit. Consistent with this prediction, deletion of *nadR* is highly beneficial in the LTEE environment [36].

Pyruvate kinase 1 (*pykF*) catalyzes the final step of glycolysis, transferring a phosphate group from phophoenolpyruvate (PEP) to ADP to generate pyruvate and ATP. It is a key enzyme in

regulating glycolytic flux [37,38]. We observed an intermediate representation of disruptive mutations in *pykF*, fewer than in *nadR* but more than in *topA* (**Fig. 4C**). Interestingly, nonsynonymous base substitutions in *pykF* tend to have a larger selection coefficient than disruptive mutations ($p = 0.00390$, Kolmogorov-Smirnov test) (**Fig. 5A**). This finding is in agreement with a recent study of various *pykF* alleles that arose in the LTEE which found that nearly all *pykF* point mutations were more beneficial than deletion of the *pykF* gene, both in the ancestor and in evolved genetic backgrounds [39]. PykF forms a homotetramer in which each polypeptide is folds into three structural domains [40,41]. The central domain C forms the active site at the interface with domain B and the binding site for the allosteric effector fructose 1,6-bisphosphate at the interface with domain A. The nonsynonymous mutations that we observed are more concentrated than expected in domain C versus the other structural domains ($p = 0.0050$, one-tailed binomial test) (**Fig. 5B**). Overall, these results suggest that complete inactivation of *pykF* is highly beneficial in the environment of our evolution experiment, but mutations that alter its activity—likely in ways that reduce glycolytic flux—are even more so. It has been suggested that reducing *pykF* activity is beneficial in the similar glucose-limited conditions of the LTEE because this allows more PEP to be used for import of glucose into cells by the phosphotransfer system [42].

DNA topoisomerase I (*topA*) relaxes negative supercoiling introduced into the chromosome by replication and transcription [43]. The mutations we observed in *topA* are almost exclusively single-base substitutions (**Fig. 4C**), suggesting that modulating the activity of this enzyme provides a fitness benefit. Indeed, complete loss of *topA* function is lethal to *E. coli* without compensatory mutations in DNA gyrase [44,45]. The structure of *E. coli* TopA consists of four N-terminal domains (D1-D4) that make up the catalytic core and five C-terminal zinc finger and

ribbon domains (D5-D9) [46]. The few out-of-frame indels and the large deletion that we observe truncate TopA within domains D7-D9, which interact with single-stranded DNA and RNA polymerase but are not critical for catalysis. Considering only the catalytic core, we find that nonsynonymous mutations are concentrated in domains D1 and D4 versus D2 and D3 (p = 0.0060, one-tailed binomial test) (**Fig. 5C**). D1 and D4 together form the ssDNA binding groove leading to the active site, and D1 also forms part of the active site at its interface with D3 [47]. Several base substitutions in *topA* have been shown to increase positive supercoiling in evolved LTEE strains [48,49]. The exact reason that this change in supercoiling is beneficial is unknown, but it may be linked to increasing the expression of ribosomal RNAs [48], altering gene regulation responses to starvation or stress [49], and/or increasing gene expression divergently transcribed operons [50].

Discussion

We examined bacterial evolution during the initial stages of clonal competition when there is a burst of beneficial genetic diversity as many new subpopulations with different mutations evolve and begin to displace the ancestral genotype. We focused on eight genes known to accumulate adaptive mutations in the >70,000 generation Lenski long-term evolution experiment (LTEE) with *E. coli* that used nearly the same environment as our experiments. The only difference was that we added four times as much of the limiting nutrient (glucose). By combining Illumina sequencing of reads that incorporate molecular indexes for error correction, hybridization-based capture of DNA encoding these genes, and dense temporal sampling, we were able to identify more beneficial mutations and track them at much lower frequencies than is possible with standard metagenomic sequencing. We detected a total of 236 mutations in the focal genes: 180

304 in the complete time courses of four populations and 161 in the window time courses of these
305 populations and two others, with 105 mutations overlapping between the two sets.

306 By densely sampling and deeply sequencing *E. coli* populations, we were able to characterize
307 many beneficial mutations that never reach the detection limits of standard Illumina sequencing
308 before they become casualties of clonal interference. Only 13 of the 180 mutations we detected
309 in the complete time courses ever achieved a frequency of 5% or more, which can be reliably
310 distinguished from noise by standard metagenomic sequencing, and only seven were this
311 common for 100 or more generations, such that they were likely to be detected by a typical time-
312 sampling scheme. Considering both the complete and window time courses we characterized 177
313 and 27 mutations that never reached 1% or 0.1% thresholds, respectively, at any of our sampled
314 time points. Our success in recovering rare variants meant that we discovered more examples of
315 beneficial mutations in the three commonly mutated genes (*topA*, *pykF*, and *nadR*) than have
316 been reported in many prior studies of the evolution of the twelve LTEE populations
317 [3,31,36,42,51,52]. These large sets of mutations enabled us to identify distinct molecular
318 signatures of adaptation in each protein.

319 We profiled evolution driven by mutations in eight genes known to be targets of selection in
320 the LTEE. Mutations in four of these (*topA*, *pykF*, *spoT*, and *fabR*) reach high frequencies within
321 the first 1,000 generations of the LTEE in multiple populations [3,52]. Mutations in the other
322 four (*hslU*, *nadR*, *ybaL*, and *iclR*) are also common in the LTEE, but they typically occur later
323 (often within the first 2,000 to 10,000 generations) [3,31]. Nearly all mutations in these genes in
324 our evolution experiment were in *topA*, *pykF*, and *nadR*, but we also found multiple mutations
325 that were similarly beneficial in *spoT*, *fabR*, and *ybaL*. Mutations in *nadR* were more widespread
326 than expected in our experiment and may be more likely to completely disrupt its function than

beneficial alleles that evolve in the LTEE [51]. Mutations in *spoT* and *fabR* were rarer than expected from the LTEE. One possible explanation for these slight differences is the increased concentration of glucose in our experiment compared to the LTEE. These minor deviations are also reminiscent of how changing a different aspect of the environment (temperature) re-focuses the mutations of largest benefit that succeed early onto different subsets of genes, nearly all of which eventually accumulate beneficial mutations later in the LTEE environment, in related evolution experiments [53,54]. Despite these subtle differences, we were still able to account for majority of the genetic variation present in three of four of the four populations that we profiled over the entire 500 generations by analyzing evolution in the eight candidate genes.

We also wanted to understand to what extent we gained early warning of driver mutations by deeply profiling evolution in genes we expected to be under strong selection. In general, we were able to begin tracking most mutations when they were above a frequency of 0.01%. This level of profiling enabled us to first detect mutations an average of 75, 152, and 290 generations before they surpassed frequencies of 0.1%, 1%, and 5%, respectively. Under the conditions of our experiment these intervals take roughly 11, 23, and 44 days, respectively; so, even though we made these predictions retrospectively, there would have been sufficient time to complete the DNA isolation, library preparation, sequencing, and analysis steps quickly enough for this approach to give early warning of specific genetic variants driving evolution of these populations. The amount of lead time becomes disproportionately longer at higher frequencies due to clonal interference between beneficial mutations. The chances and timescales of earlier detection are expected to increase even more when there are ecological interactions or spatial structure that further slow the takeover of new variants, as has been demonstrated and discussed in other microbial evolution experiments [26,55,56].

A further prediction is that the genes in which we observe early, but unsuccessful beneficial mutations will sustain mutations again and again until they are successful in a population's evolutionary future. This prediction is limited by the nature of epistatic interactions. In the LTEE and other microbial evolution experiments, diminishing returns epistasis dominates between beneficial mutations in different genes [57–61]. That is, mutations in one gene that improve the fitness of the ancestor tend to still be beneficial to evolved genotypes containing beneficial mutations in other genes, just less so than when those other mutations are not present. Subpopulations with mutations in both *nadR* and *pykF* evolve by 20,000 generations in all 12 LTEE populations, and cells that also contain a mutation in *topA* are found in six of the LTEE populations at this point [31]. By this time, mutations in *ybaL* and *spoT* are also found in nine and six LTEE populations, respectively. So, for five of the six genes in which we detected multiple mutations in the initial burst phase, it is likely that nearly all of them would have eventually accumulated beneficial mutations if we continued our experiment.

The genes in which we did not detect multiple mutations (*fabR*, *iclR*, and *hslU*) likely represent other scenarios. Mutations in *fabR* transiently appear within the first 2,000 generations of the LTEE [52]. They interact unfavorably with beneficial mutations in *spoT* and other genes, such that a *fabR* mutation essentially precludes further adaptation by mutating the other set of genes and vice-versa [52,62]. So, *fabR* mutations are unlikely to re-emerge in the future of these populations. On the other hand, mutations in *iclR* and *hslU* appear to either require the presence of mutations in other genes to become highly beneficial or may not be able to experience any mutations that are beneficial enough to make them competitive early on in the LTEE. Of the 12 LTEE populations, 11 have cells mutations in *iclR* and 11 have mutations in *hslU* by 20,000 generations, which makes them more common than mutations in *spoT* and *ybaL* in the long run.

373 The nature of epistasis and the limits that it imposes on predicting the future evolution of a
374 cell population could be further probed using our approach in several ways. One could repeat the
375 evolution experiment beginning with genotypes containing different first-step beneficial
376 mutations as starting points to more finely map the fitness landscape. One could also interrogate
377 the diverse collections of cells containing different beneficial alleles that we have evolved, by
378 taking the 150-generation populations and further evolving them under different conditions to
379 map genotype by environment effects, for example. Such experiments might also reveal latent
380 beneficial mutation in other genes (e.g., *iclR* and *hslU*) that were able to outcompete the ancestor
381 in our experiment but remained below the detection limit because they were not as beneficial as
382 mutations in *topA*, *pykF*, and *nadR* in this environment. There is precedent for changes in the
383 environment deflecting selection to different subsets of the same genes. In an offshoot of the
384 LTEE that began with a clone that had *spoT*, *topA*, and *pykF* mutations, selection was focused on
385 *hslU*, *iclR*, or *nadR* depending on changes in temperature [54].
386 Alternative and complementary methods exist for deeply profiling the evolutionary
387 possibilities inherent in the fitness landscape of a cell, i.e., its evolvome. We tracked spontaneous
388 beneficial mutations within targeted genome regions, or a portion of what one could more
389 specifically describe as the adaptome [63]. Amplicon sequencing can also capture mutations in a
390 subset of the genome with deep coverage. We used hybridization-based enrichment, which did
391 not require any experimental optimization for different targets and is less likely to introduce
392 biases in inferring the frequencies of mutations, like IS insertions, that change amplicon sizes
393 [64]. With enough input DNA and enough sequencing, our approach could be scaled to more
394 genes or the whole genome. Tracking the frequencies of barcoded cells and their progeny has
395 been used to characterize the statistical properties of much larger collections of naturally

396 occurring beneficial mutations and when they are much rarer within populations [27,28].
397 However, one must barcode individuals in the population to apply this method, which may be
398 difficult in certain cell types or in clinical samples, and additional genome sequencing after an
399 experiment is completed is required to discover the identities of the beneficial mutations linked
400 to barcodes. Other methods such as deep-mutational scanning [65] or CRISPR-enabled trackable
401 genome engineering [66] can simultaneously interrogate large libraries of mutants to map
402 evolvomes. However, since they artificially construct variant libraries, they do not necessarily
403 provide information about which genetic variants are accessible by spontaneous mutations and
404 would therefore be expected to contribute the most to a cell's adaptome.

405 Exhaustively mapping paths that clonal evolution is likely to follow is of particular interest
406 and utility in systems that evolve repeatedly from a defined starting point. These range from
407 bioreactors that are seeded with the same strain in different production runs to lung infections in
408 cystic fibrosis patients that start from similar, but not identical, opportunistic pathogens. The
409 ability to identify mutations in key genes while they are still very rare may also be used to
410 improve the early detection and predicting drug resistance in other human infections and cancer.
411 The evolutionary dynamics will be more complex in many of these systems, but they may also
412 unfold more slowly. For example, biofilm formation and the necessity of invading already
413 colonized niches will slow the dynamics of competition. This potentially makes the therapeutic
414 window for detecting incipient evolution by profiling the adaptome even greater.
415

416 ## Materials and Methods

417 ### Evolution experiment

418 Strains and growth conditions are derived from the Lenski long-term evolution experiment
419 [29,67]. Nine clonal isolates of *E. coli* B strain REL606 and nine of strain REL607 were grown
420 overnight at 37°C with orbital shaking over a one-inch diameter at 120 RPM in 10 mL of Davis
421 Minimal (DM) media containing 100 µg/L glucose (DM100). This is a slightly higher
422 concentration of glucose than the 25 µg/L glucose (DM25) used in the LTEE. Day 0 cultures
423 containing 10 mL of fresh DM100 were inoculated with 50 µL of one REL606 culture and 50 µL
424 of one REL607 culture for overnight growth in the same conditions. The remaining culture
425 volume was archived at −80°C with 2 mL dimethyl sulfoxide (DMSO) added as cryoprotectant.
426 Daily transfer of 100 µL of overnight culture to 10 mL of fresh DM100 and archival of the
427 remaining culture volume in the same way continued through 75 daily transfers. Periodically 1
428 µL of culture was diluted 10,000-fold in sterile saline and plated on tetrazolium arabinose (TA)
429 agar to allow growth of ~200 colonies. REL606 and REL607 differ by a mutation in an
430 arabinose utilization gene that makes REL606 (Ara^-) colonies red and REL607 (Ara^+) colonies
431 pink [29]. The ratio of red to pink colonies was used to monitor these populations for selective
432 sweeps [62,68].
433
434 **DNA isolation and library preparation**
435 Genomic DNA (gDNA) was isolated from frozen population samples by first thawing each 15
436 mL conical tube on ice. Of the ~12 mL total volume of culture plus cryoprotectant, 1.2 mL was
437 transferred to a 2 mL cryovial and refrozen. The remaining ~10.8 mL was centrifuged at 6,500 ×
438 g at 4°C for 15 minutes. The resulting cell pellets were transferred with a volume of remaining
439 solution to 1.7 mL Eppendorf tubes. Then, gDNA was isolated using the PureLink Genomic

DNA Mini kit (Life Technologies). For each sample, 1 µg of gDNA was randomly fragmented on a Covaris S2 focused-ultrasonicator.

Illumina libraries were constructed using the Kappa Biosystems LTP Library Preparation Kit with the following modifications. End-repaired, fragmented DNA was T-tailed (rather than A-tailed) in a 50 µl reaction including 10 mM dTTP and 5 units of Klenow fragment, exo$^-$ (New England Biolabs). Illumina adapters containing 12-base molecular indexes were ligated to the T-tailed fragments as previously described [32], except full-length adapter sequences containing unique external sample barcodes were directly ligated to the T-tailed dsDNA inserts to reduce the risk of cross-contamination between samples. The full list of DNA sequence adaptors used is provided in **Table S1**.

Probe design and target capture

Oligonucleotide probes consisting of 60-base xGen Lockdown probes (Integrated DNA Technologies) were designed to tile across each of the eight genes of interest including upstream promoter elements. Probes for each gene were compared to the entire *E. coli* B strain REL606 reference genome (GenBank: NC_012967.1) [69] using BLASTN [70]. The starting positions of all probes in a set were shifted by one base at a time until every probe had only a single significant predicted binding location (match with E-value $< 2\times10^{-5}$). The sequences of the final set of 242 probes are provided in **Table S2**.

Capture was performed using a SeqCap EZ Exome Enrichment kit v3.0 (NimbleGen) with several modifications to the protocol. First, 18 to 20 population samples with unique barcodes were pooled together in a single capture reaction that contained a total of 1 µg of library DNA from all samples, 1 mmol of a universal blocking oligo, and 1 mmol of a degenerate barcode

463 blocking oligo. The sequences of these blocking oligos are provided in **Table S3**. Second, after
464 hybridization for 72 h, DNA fragments hybridized to the biotinylated probes were recovered
465 using MyOne Streptavidin C1 Dynabeads (Life Technologies). Third, a final 8-cycle PCR step
466 was performed with HiFi Hotstart DNA Polymerase (Kappa Biosystems).
467
468 **Sequencing and read processing**
469 Paired-end 101- or 125-base sequencing of the final libraries was performed on an Illumina
470 HiSeq 2000 at the University of Texas at Austin the Genome Sequencing and Analysis Facility
471 (GSAF). Read sequences have been deposited into the NCBI Sequence Read Archive
472 (PRJNA601748). Raw reads were used to generate Consensus Sequence Reads (CSR) using
473 custom Python scripts that carried out the following steps. First, the beginning of each read was
474 evaluated for the presence of the expected 5′-end tag, consisting of the random twelve-base
475 molecular index (MI) followed by four fixed bases (5′-N_{12}CAGT). Read pairs lacking the correct
476 5′-end tag on either read were discarded. For remaining read pairs, the MIs from each read were
477 concatenated to create a 24-base dual-MI that uniquely identifies the original DNA fragment that
478 was amplified and sequenced. To group all reads corresponding to the same initial DNA
479 molecule, a FASTA file of all dual-MIs was used as input into the *ustacks* program from the
480 Stacks software pipeline (Version 1.48) [71] with the following options: a single read was
481 sufficient to seed a stack, a single mismatch within the 24 base MI was allowed in assigning a
482 read to a stack, secondary reads and haplotypes were disabled, and stacks with high coverage
483 were preserved. Then, CSRs were generated for all MI groups sequenced at least twice by taking
484 the straight consensus of all reads that were merged into that stack. If no base exceeded 50%
485 frequency at a given position in this set of reads, then that base was set as unknown (N).

Variant calling

We used the *breseq* pipeline [52,72,73] (version 0.26.0) to call single-nucleotide variants (SNVs) and structural variants (SVs) from the CSRs. We divided the genome sequence of the ancestral *E. coli* REL606 strain into two types of reference regions for mapping in this analysis. The eight regions of the genome tiled with probes—extended with hundreds of bases of flanking sequence on both sides—were input as "targeted" sequences, and the remainder of the genome with the identical eight regions masked to degenerate N bases was supplied as a "junction-only" reference (to which reads are mapped without variant calling). All 116 samples were analyzed using *breseq* in polymorphism prediction mode with all bias, minimum allele frequency, and read-count filters disabled. Evidence items in the Genome Diff (GD) files for all samples were combined using the *gdtools* utility program to generate a single merged GD file with each piece of evidence listed a single time, regardless of how many times it was detected in different samples. We then re-ran *breseq* using the same parameters except that this GD file was supplied as an input user-evidence file to force output of variant and reference information for these putative variants in every sample. Then, we extracted the number of variant reads supporting each putative variant allele and the total number of reads at that reference location from the GD file output by *breseq* and performed subsequent statistical tests and fitting steps in R (version 3.2.2) [74]. Scripts and data files for this analysis are available in GitHub (https://github.com/barricklab/adaptome-capture).

Since this original analysis was conducted at the level of *breseq* evidence (i.e., single columns of read pileups on the reference genome or instances of new sequence junctions), we next merged sets of observations that were consistent with a single mutational event when they

also had frequency trajectories that tracked together. To identify these candidates for merging, we analyzed each of the six window (generation 133 to 213) and four complete (generation 0 to 500) time courses separately. We only considered mutations that exceeded a threshold frequency of 0.03% at some time during each time course as candidates for merging. Read alignment (RA) evidence items were merged when they were located within 6 base pairs of one another and within a normalized Canberra distance of 0.1 between the vectors of their frequency observations across all of the time points in a dataset. All RA evidence pairs of this kind were found to co-occur in the same sequencing reads. For these cases, the read counts for the first linked mutation were used to represent the entire event. For example, if a deletion of three base pairs was predicted by missing bases at positions x, y, and z; then the frequency of missing the first base (x) was assigned to the entire three-base deletion mutation. For new junction (JC) evidence we performed the same merging procedure but allowed linked mutations to be within a larger window of 20 base pairs and within a normalized Canberra distance of 0.5. JC pairs passing these criteria were only merged if they were also consistent with an IS-element insertion in terms of their relative orientation and spacing. In this case the variant and total read counts were added together for the two different junctions, as the junctions on each side of the inserted IS element provide independent information for estimating the frequency of this type of mutation.

Time course filtering and selection coefficient estimation

After merging evidence of genetic variants into lists of putative mutations, we further eliminated putative evolved alleles from consideration using several filtering steps. For the complete time courses, we first required that non-zero frequencies be observed for a mutation in samples from at least two different time points. We next applied a filter to eliminate spurious variants that can

be recognized as arising from systematic sequencing or alignment errors because they do not exhibit the correlated changes in frequency over time expected for the frequency trajectories of real mutations [1]. Specifically, we required that the time-series of estimated frequencies for a mutation over all analyzed time points have an autocorrelation value ≥ 0.55.

For the window time courses, we further required that the estimated frequency of a putative mutation be $\geq 10^{-4}$ at each of the last three time points that were sequenced (generations 193, 206, and 213). Then, we fit a binomial logistic model with slope and x-intercept terms to the time courses of counts of variant and reference (total minus variant) observations for each mutation. We filtered out any mutations for which this fit had an AIC < 200, Bonferroni-corrected p-value for the slope differing from zero of > 0.05, or an x-intercept < -15. The slope fit from the frequency trajectory of each mutation is an estimate of the selection coefficient of each mutation, assuming the trajectories reflect competition purely against the ancestral strain. However, in the latter half of the window time courses we detected a significant deviation from linearity for all mutation trajectories, indicating that the overall population fitness had increased to a degree that it decreased the rate at which all newly evolved genotypes with beneficial mutations increased in frequency. The figures show the best models for a stepwise increase in population fitness between the sequenced time points that improved the fits for all mutations in each population considered separately. Because there was significant uncertainty in these estimates and the fitness trajectories are expected to be highly similar between different populations, we used a consensus model with one step-wise increase in fitness over time that best improved the fits for all mutations from all populations to correct the estimated selection coefficients for this effect.

Mutation statistics and plots

555 One *nadR* mutation from population A2 was a noticeable outlier in terms of its large apparent
556 fitness benefit of 9.2%. Given that the next-highest observed selection coefficient for a mutation
557 was 8.0%, it is likely that the lineage with this *nadR* mutation also sustained a secondary
558 beneficial mutation early enough that they rose to detectable frequencies together. Therefore, we
559 removed this mutation before analyzing or graphing the characteristics of the set of likely single-
560 step mutations. Graphs were generated in R using the ggplot2 package [75].
561
562 **Protein structure analysis**
563 Structural domains in NadR, PykF, and TopA were defined according to UniProt and papers
564 reporting x-ray crystal structures. Mutations in PykF were mapped onto Protein Data Bank
565 structure 4YNG [41]. Mutations in TopA were mapped onto Protein Data Bank structure 1MW8
566 [47]. Protein structures were visualized using Pymol v2.3.5 (Schrödinger LLC).
567

568 ## Acknowledgements
569 The authors acknowledge the Texas Advanced Computing Center (TACC) at The University of
570 Texas at Austin for providing high-performance computing resources.
571

572 ## Supporting Information
573 **Table S1** Adapter sequences used in DNA library preparation
574 **Table S2** Sequences of pulldown probes
575 **Table S3** Blocking oligos used to limit read-to-read binding during pulldown
576

577 ## Author contributions

Conceptualization: DED JEB.

Data Curation: DED JEB.

Funding Acquisition: DED JEB.

Investigation: DED.

Methodology: DED JEB.

Software – DED JEB.

Supervision – JEB.

Visualization: DED JEB.

Writing – Original Draft Preparation: DED JEB.

Writing – Review & Editing: DED JEB.

References

1. Lang GI, Rice DP, Hickman MJ, Sodergren E, Weinstock GM, Botstein D, et al. Pervasive genetic hitchhiking and clonal interference in forty evolving yeast populations. Nature. 2013;500: 571–574. doi:10.1038/nature12344

2. Maddamsetti R, Lenski RE, Barrick JE. Adaptation, clonal interference, and frequency-dependent interactions in a long-term evolution experiment with *Escherichia coli*. Genetics. 2015;200: 619–631. doi:10.1534/genetics.115.176677

3. Good BH, McDonald MJ, Barrick JE, Lenski RE, Desai MM. The dynamics of molecular evolution over 60,000 generations. Nature. 2017;551: 45–50. doi:10.1038/nature24287

4. Gerrish PJ, Lenski RE. The fate of competing beneficial mutations in an asexual population. Genetica. 1998;102–103: 127–144.

5. Park S-C, Krug J. Clonal interference in large populations. Proc Natl Acad Sci U S A. 2007;104: 18135–18140.

6. Desai MM, Walczak AM, Fisher DS. Genetic diversity and the structure of genealogies in rapidly adapting populations. Genetics. 2012;193: 565–585. doi:10.1534/genetics.112.147157

7. Merlo LMF, Pepper JW, Reid BJ, Maley CC. Cancer as an evolutionary and ecological process. Nat Rev Cancer. 2006;6: 924–935. doi:10.1038/nrc2013

8. Thomas RK, Nickerson E, Simons JF, Jänne PA, Tengs T, Yuza Y, et al. Sensitive mutation detection in heterogeneous cancer specimens by massively parallel picoliter

reactor sequencing. Nat Med. 2006;12: 852–855. doi:10.1038/nm1437

9. Marusyk A, Almendro V, Polyak K. Intra-tumour heterogeneity: A looking glass for cancer? Nat Rev Cancer. 2012;12: 323–334. doi:10.1038/nrc3261

10. Merlo LMF, Shah NA, Li X, Blount PL, Vaughan TL, Reid BJ, et al. A comprehensive survey of clonal diversity measures in Barrett's esophagus as biomarkers of progression to esophageal adenocarcinoma. Cancer Prev Res (Phila). 2010;3: 1388–1397. doi:10.1158/1940-6207.CAPR-10-0108

11. Landau DA, Carter SL, Stojanov P, McKenna A, Stevenson K, Lawrence MS, et al. Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. Cell. 2013;152: 714–26. doi:10.1016/j.cell.2013.01.019

12. Ding L, Ley TJ, Larson DE, Miller C a., Koboldt DC, Welch JS, et al. Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. Nature. 2012;481: 506–10. doi:10.1038/nature10738

13. Marvig RL, Sommer LM, Molin S, Johansen HK. Convergent evolution and adaptation of *Pseudomonas aeruginosa* within patients with cystic fibrosis. Nat Genet. 2015;47: 57–64. doi:10.1038/ng.3148

14. Winstanley C, O'Brien S, Brockhurst MA. *Pseudomonas aeruginosa* evolutionary adaptation and diversification in cystic fibrosis chronic lung infections. Trends Microbiol. 2016;24: 327–337. doi:10.1016/j.tim.2016.01.008

15. Stefani S, Campana S, Cariani L, Carnovale V, Colombo C, Lleo MM, et al. Relevance of multidrug-resistant *Pseudomonas aeruginosa* infections in cystic fibrosis. Int J Med Microbiol. 2017;307: 353–362. doi:10.1016/j.ijmm.2017.07.004

16. Renda BA, Hammerling MJ, Barrick JE. Engineering reduced evolutionary potential for synthetic biology. Mol Biosyst. 2014;10: 1668–1678. doi:10.1039/c3mb70606k

17. Rugbjerg P, Sommer MOA. Overcoming genetic heterogeneity in industrial fermentations. Nat Biotechnol. 2019;37: 869–876. doi:10.1038/s41587-019-0171-6

18. Lee SY, Kim HU. Systems strategies for developing industrial microbial strains. Nat Biotechnol. 2015;33: 1061–72. doi:10.1038/nbt.3365

19. Nielsen J, Keasling JD. Engineering cellular metabolism. Cell. 2016;164: 1185–1197. doi:10.1016/j.cell.2016.02.004

20. Sandoval CM, Ayson M, Moss N, Lieu B, Jackson P, Gaucher SP, et al. Use of pantothenate as a metabolic switch increases the genetic stability of farnesene producing *Saccharomyces cerevisiae*. Metab Eng. 2014;25: 215–226. doi:10.1016/j.ymben.2014.07.006

21. Rugbjerg P, Myling-Petersen N, Porse A, Sarup-Lytzen K, Sommer MOA. Diverse genetic error modes constrain large-scale bio-based production. Nat Commun. 2018;9: 787. doi:10.1038/s41467-018-03232-w

22. Zelder O, Hauer B. Environmentally directed mutations and their impact on industrial biotransformation and fermentation processes. Curr Opin Microbiol. 2000;3: 248–251.

649 doi:10.1016/S1369-5274(00)00084-9

650 23. Barrick JE, Lenski RE. Genome dynamics during experimental evolution. Nat Rev Genet.
651 2013;14: 827–839. doi:10.1038/nrg3564

652 24. Gresham D, Dunham MJ. The enduring utility of continuous culturing in experimental
653 evolution. Genomics. 2014;104: 399–405. doi:10.1016/j.ygeno.2014.09.015

654 25. Barrick JE, Lenski RE. Genome-wide mutational diversity in an evolving population of
655 *Escherichia coli*. Cold Spring Harb Symp Quant Biol. 2009;74: 119–129.

656 26. Traverse CC, Mayo-Smith LM, Poltak SR, Cooper VS. Tangled bank of experimentally
657 evolved *Burkholderia* biofilms reflects selection during chronic infections. Proc Natl Acad
658 Sci U S A. 2013;110: E250–E259. doi:10.1073/pnas.1207025110

659 27. Levy SF, Blundell JR, Venkataram S, Petrov DA, Fisher DS, Sherlock G. Quantitative
660 evolutionary dynamics using high-resolution lineage tracking. Nature. 2015;519: 181–
661 186. doi:10.1038/nature14279

662 28. Venkataram S, Dunn B, Li Y, Agarwala A, Chang J, Ebel ER, et al. Development of a
663 comprehensive genotype-to-fitness map of adaptation-driving mutations in yeast. Cell.
664 2016;166: 1585-1596.E22. doi:10.1016/j.cell.2016.08.002

665 29. Lenski RE, Rose MR, Simpson SC, Tadler SC. Long-term experimental evolution in
666 *Escherichia coli*. I. Adaptation and divergence during 2,000 generations. Am Nat.
667 1991;138: 1315–1341.

668 30. Hegreness M, Shoresh N, Hartl D, Kishony R. An equivalence principle for the
669 incorporation of favorable mutations in asexual populations. Science. 2006;311: 1615–
670 1617.

671 31. Tenaillon O, Barrick JE, Ribeck N, Deatherage DE, Blanchard JL, Dasgupta A, et al.
672 Tempo and mode of genome evolution in a 50,000-generation experiment. Nature.
673 2016;536: 165–170. doi:10.1038/nature18959

674 32. Schmitt MW, Kennedy SR, Salk JJ, Fox EJ, Hiatt JB, Loeb LA. Detection of ultra-rare
675 mutations by next-generation sequencing. Proc Natl Acad Sci U S A. 2012;109: 14508–
676 14513. doi:10.1073/pnas.1208715109

677 33. Bainbridge MN, Wang M, Burgess DL, Kovar C, Rodesch MJ, D'Ascenzo M, et al.
678 Whole exome capture in solution with 3 Gbp of data. Genome Biol. 2010;11: R62.
679 doi:10.1186/gb-2010-11-6-r62

680 34. Raffaelli N, Lorenzi T, Mariani PL, Emanuelli M, Amici A, Ruggieri S, et al. The
681 *Escherichia coli* NadR regulator is endowed with nicotinamide mononucleotide
682 adenylyltransferase activity. J Bacteriol. 1999;181: 5509–5511.

683 35. Kurnasov O V., Polanuyer BM, Ananta S, Sloutsky R, Tam A, Gerdes SY, et al.
684 Ribosylnicotinamide Kinase Domain of NadR Protein: Identification and Implications in
685 NAD Biosynthesis. J Bacteriol. 2003;185: 698–698. doi:10.1128/JB.185.2.698.2003

686 36. Barrick JE, Yu DS, Yoon SH, Jeong H, Oh TK, Schneider D, et al. Genome evolution and
687 adaptation in a long-term experiment with *Escherichia coli*. Nature. 2009;461: 1243–

1247.

37. Siddiquee KAZ, Arauzo-Bravo MJ, Shimizu K. Effect of a pyruvate kinase (*pykF*-gene) knockout mutation on the control of gene expression and metabolic fluxes in *Escherichia coli*. FEMS Microbiol Lett. 2004;235: 25–33. doi:10.1016/j.femsle.2004.04.004

38. Kochanowski K, Volkmer B, Gerosa L, Van Rijsewijk BRH, Schmidt A, Heinemann M. Functioning of a metabolic flux sensor in *Escherichia coli*. Proc Natl Acad Sci U S A. 2013;110: 1130–1135. doi:10.1073/pnas.1202582110

39. Peng F, Widmann S, Wünsche A, Duan K, Donovan KA, Dobson RCJ, et al. Effects of beneficial mutations in *pykF* gene vary over time and across replicate populations in a long-term experiment with bacteria. Mol Biol Evol. 2018;35: 202–210. doi:10.1093/molbev/msx279

40. Mattevi A, Valentini G, Rizzi M, Speranza ML, Bolognesi M, Coda A. Crystal structure of *Escherichia coli* pyruvate kinase type I: molecular basis of the allosteric transition. Structure. 1995;3: 729–741. doi:10.1016/S0969-2126(01)00207-6

41. Donovan KA, Atkinson SC, Kessans SA, Peng F, Cooper TF, Griffin MDW, et al. Grappling with anisotropic data, pseudo-merohedral twinning and pseudo-translational noncrystallographic symmetry: A case study involving pyruvate kinase. Acta Crystallogr Sect D Struct Biol. 2016;72: 512–519. doi:10.1107/S205979831600142X

42. Woods R, Schneider D, Winkworth CL, Riley MA, Lenski RE. Tests of parallel molecular evolution in a long-term experiment with *Escherichia coli*. Proc Natl Acad Sci U S A. 2006;103: 9107–9712. doi:10.1073/pnas.0602917103

43. Massé E, Drolet M. Relaxation of transcription-induced negative supercoiling is an essential function of *Escherichia coli* DNA topoisomerase I. J Biol Chem. 1999;274: 16654–16658. doi:10.1074/jbc.274.23.16654

44. Dinardo S, Voelkel KA, Sternglanz R, Reynolds AE, Wright A. *Escherichia coli* DNA topoisomerase I mutants have compensatory mutations in DNA gyrase genes. Cell. 1982;31: 43–51. doi:10.1016/0092-8674(82)90403-2

45. Pruss GJ, Manes SH, Drlica K. *Escherichia coli* DNA topoisomerase I mutants: Increased supercoiling is corrected by mutations near gyrase genes. Cell. 1982;31: 35–42. doi:10.1016/0092-8674(82)90402-0

46. Tan K, Zhou Q, Cheng B, Zhang Z, Joachimiak A, Tse-Dinh YC. Structural basis for suppression of hypernegative DNA supercoiling by *E. coli* topoisomerase I. Nucleic Acids Res. 2015;43: 11031–11046. doi:10.1093/nar/gkv1073

47. Perry K, Mondragón A. Structure of a complex between *E. coli* DNA topoisomerase I and single-stranded DNA. Structure. 2003;11: 1349–1358. doi:10.1016/j.str.2003.09.013

48. Crozat E, Philippe N, Lenski RE, Geiselmann J, Schneider D. Long-term experimental evolution in *Escherichia coli*. XII. DNA topology as a key target of selection. Genetics. 2005;169: 523–532.

49. Crozat E, Winkworth C, Gaffé J, Hallin PF, Riley MA, Lenski RE, et al. Parallel genetic and phenotypic evolution of DNA superhelicity in experimental populations of

Escherichia coli. Mol Biol Evol. 2010;27: 2113–2128.

50. Houdaigui B El, Forquet R, Hindré T, Schneider D, Nasser W, Reverchon S, et al. Bacterial genome architecture shapes global transcriptional regulation by DNA supercoiling. Nucleic Acids Res. 2019;47: 5648–5657. doi:10.1093/nar/gkz300

51. Ostrowski EA, Woods RJ, Lenski RE. The genetic basis of parallel and divergent phenotypic responses in evolving populations of *Escherichia coli*. Proc R Soc B. 2008;275: 277–284. doi:10.1098/rspb.2007.1244

52. Deatherage DE, Traverse CC, Wolf LN, Barrick JE. Detecting rare structural variation in evolving microbial populations from new sequence junctions using *breseq*. Front Genet. 2015;5: 468. doi:10.3389/fgene.2014.00468

53. Tenaillon O, Rodríguez-Verdugo A, Gaut RL, McDonald P, Bennett AF, Long AD, et al. The molecular diversity of adaptive convergence. Science. 2012;335: 457–461. doi:10.1126/science.1212986

54. Deatherage DE, Kepner JL, Bennett AF, Lenski RE, Barrick JE. Specificity of genome evolution in experimental populations of *Escherichia coli* evolved at different temperatures. Proc Natl Acad Sci U S A. 2017;114: E1904–E1912. doi:10.1073/pnas.1616132114

55. Frenkel EM, McDonald MJ, Van Dyken JD, Kosheleva K, Lang GI, Desai MM. Crowded growth leads to the spontaneous evolution of semistable coexistence in laboratory yeast populations. Proc Natl Acad Sci. 2015;112: 11306–11311. doi:10.1073/pnas.1506184112

56. Baym M, Lieberman TD, Kelsic ED, Chait R, Gross R, Yelin I, et al. Spatiotemporal microbial evolution on antibiotic landscapes. Science. 2016;353: 1147–1151. doi:10.1126/science.aag0822

57. Khan AI, Dinh DM, Schneider D, Lenski RE, Cooper TF. Negative epistasis between beneficial mutations in an evolving bacterial population. Science. 2011;332: 1193–1196.

58. Chou H-H, Chiu H-C, Delaney NF, Segrè D, Marx CJ. Diminishing returns epistasis among beneficial mutations decelerates adaptation. Science. 2011;332: 1190–1192.

59. Wiser MJ, Ribeck N, Lenski RE. Long-term dynamics of adaptation in asexual populations. Science. 2013;342: 1364–1367. doi:10.1126/science.1243357

60. Kryazhimskiy S, Rice DP, Jerison ER, Desai MM. Global epistasis makes adaptation predictable despite sequence-level stochasticity. Science. 2014;344: 1519–1522. doi:10.1126/science.1250939

61. Wei X, Zhang J. Patterns and mechanisms of diminishing returns from beneficial mutations. Mol Biol Evol. 2019;36: 1008–1021. doi:10.1093/molbev/msz035

62. Woods RJ, Barrick JE, Cooper TF, Shrestha U, Kauth MR, Lenski RE. Second-order selection for evolvability in a large *Escherichia coli* population. Science. 2011;331: 1433–1436.

63. Ryall B, Eydallin G, Ferenci T. Culture history and population heterogeneity as determinants of bacterial adaptation: the adaptomics of a single environmental transition.

Microbiol Mol Biol Rev. 2012;76: 597–625. doi:10.1128/MMBR.05028-11

64. Fischer S, Greipel L, Klockgether J, Dorda M, Wiehlmann L, Cramer N, et al. Multilocus amplicon sequencing of *Pseudomonas aeruginosa* cystic fibrosis airways isolates collected prior to and after early antipseudomonal chemotherapy. J Cyst Fibros. 2017;16: 346–352. doi:10.1016/j.jcf.2016.10.013

65. Fowler DM, Fields S. Deep mutational scanning: a new style of protein science. Nat Methods. 2014;11: 801–807. doi:10.1038/nmeth.3027

66. Garst AD, Bassalo MC, Pines G, Lynch SA, Halweg-Edwards AL, Liu R, et al. Genome-wide mapping of mutations at single-nucleotide resolution for protein, metabolic and genome engineering. Nat Biotechnol. 2017;35: 48–55. doi:10.1038/nbt.3718

67. Lenski RE, Travisano M. Dynamics of adaptation and diversification: a 10,000-generation experiment with bacterial populations. Proc Natl Acad Sci U S A. 1994;91: 6808–6814.

68. Hegreness M, Kishony R. Analysis of genetic systems using experimental evolution and whole-genome sequencing. Genome Biol. 2007;8: 201.

69. Jeong H, Barbe V, Lee CH, Vallenet D, Yu DS, Choi S-H, et al. Genome sequences of *Escherichia coli* B strains REL606 and BL21(DE3). J Mol Biol. 2009;394: 644–652.

70. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10: 421. doi:10.1186/1471-2105-10-421

71. Catchen J, Hohenlohe PA, Bassham S, Amores A, Cresko WA. Stacks: an analysis tool set for population genomics. Mol Ecol. 2013;22: 3124–3140. doi:10.1111/mec.12354

72. Deatherage DE, Barrick JE. Identification of mutations in laboratory-evolved microbes from next-generation sequencing data using *breseq*. Sun L, Shou W, editors. Methods Mol Biol. 2014;1151: 165–188. doi:10.1007/978-1-4939-0554-6_12

73. Barrick JE, Colburn G, Deatherage DE, Traverse CC, Strand MD, Borges JJ, et al. Identifying structural variation in haploid microbial genomes from short-read resequencing data using *breseq*. BMC Genomics. 2014;15: 1039. doi:10.1186/1471-2164-15-1039

74. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2016.

75. Wickham H. ggplot2: Elegant Graphics for Data Analysis. New York: Springer-Verlag; 2016.