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

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

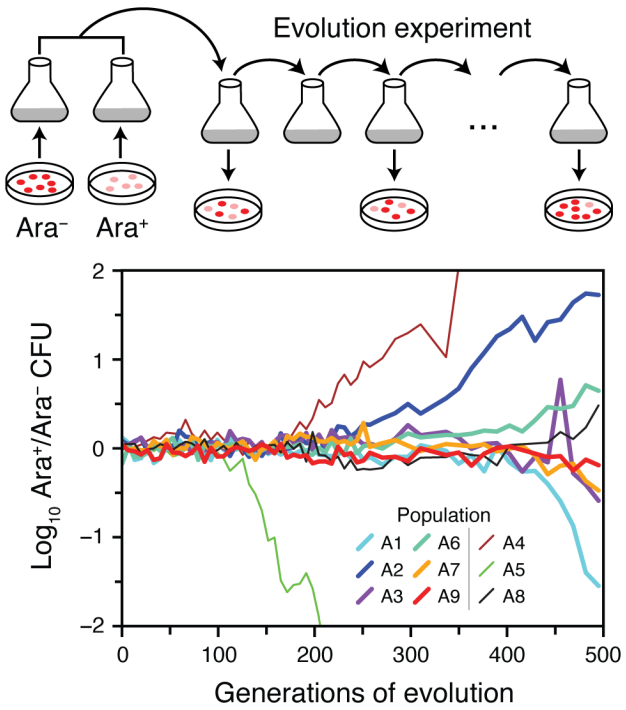
72 Evolution experiments conducted in controlled laboratory environments reproduce key
73 aspects of microbial evolution that are observed in chronic infections and bioreactors [23,24]. In
74 theory, profiling rare mutations in the earliest stages of clonal interference using high-throughput
75 DNA sequencing should allow one to anticipate the future evolution of these populations.
76 However, these studies have generally been limited by sequencing depth and per-base error rates
77 to reliably identifying mutations that are present in at least one sample at a frequency above ~1-
78 10% when they have already succeeded in becoming dominant [1,3,25,26]. Theory and
79 simulations predict that many more highly beneficial mutations evolve in these populations but
80 never reach such high frequencies before they are driven extinct [4,6], and recent studies that
81 track the evolution of barcoded lineages of microbes show that this is the case [27,28].

82 Here, we used methods for selectively increasing sequencing depth and lowering sequencing
83 error rates to deeply profile the initial burst of rare beneficial mutations in laboratory populations
84 of *E. coli*. We directly identified diverse beneficial mutations in six genes when they were orders
85 of magnitude lower in frequency and hundreds of generations earlier than could be accomplished
86 by standard metagenomic sequencing methods. By comparing our results to the long history of a
87 >70,000-generation *E. coli* evolution experiment that used the same ancestral strains and nearly
88 identical culture conditions [29], we evaluate the potential of this type of targeted adaptome
89 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].



104 **Figure 1. Replaying the first selective sweep of a long-term evolution experiment.** Nine *E.*
105 *coli* populations were initiated from equal mixtures of two variants of the ancestral strain that
106 differ in a neutral genetic marker for arabinose utilization (Ara). We observed the evolutionary
107 dynamics of these populations over ~500 generations of regrowth from 75 daily 1:100 serial
108 transfers by periodically plating dilutions of each population on indicator agar. The ratio of Ara^+
109 cells (pink colonies) to Ara^- cells (red colonies) diverges from 1:1 when descendants of one
110 ancestor type accumulate enough of a fitness advantage due to *de novo* beneficial mutations that
111 they take over. We focused further analysis on six of the nine populations (thick lines).
112

113 **Reconstructing the trajectories of new beneficial mutations**

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

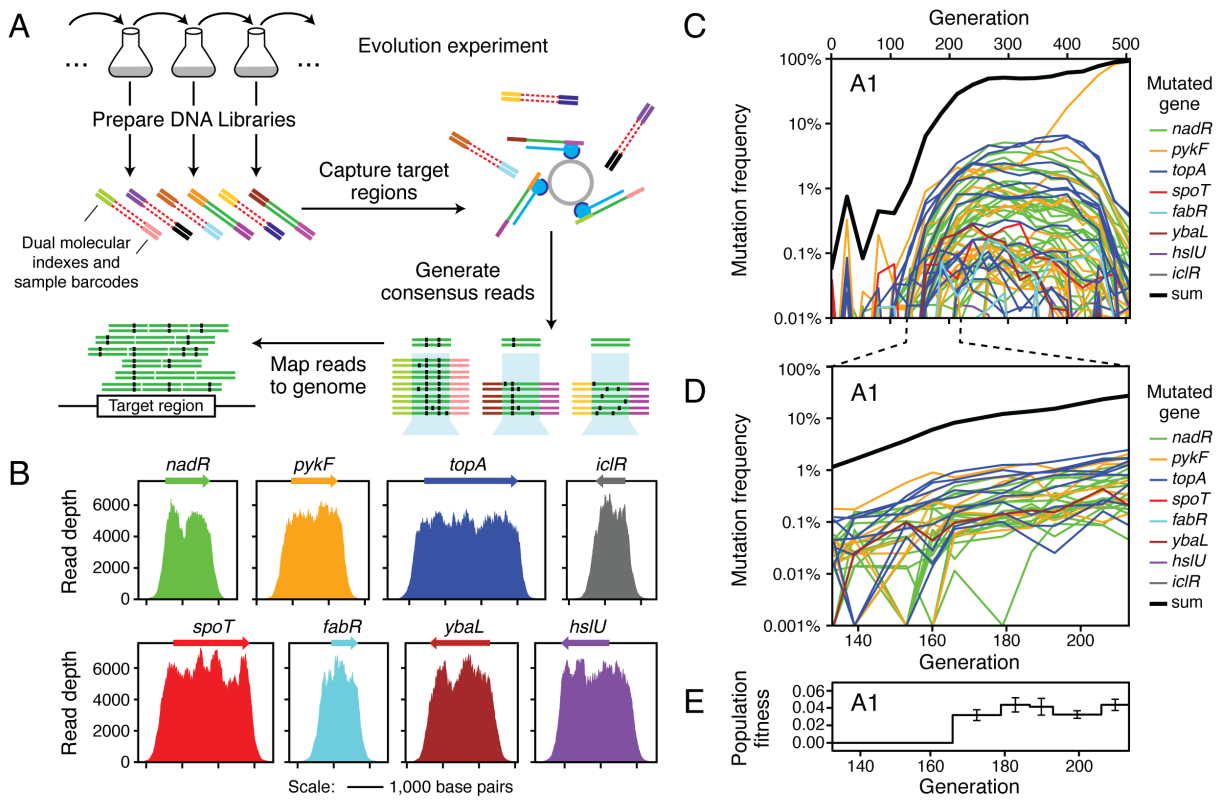


Figure 2. Profiling many beneficial mutations in the first selective sweep by deep sequencing. (A) Schematic of the deep sequencing approach. Genomic DNA is directly isolated from the *E. coli* populations and prepared for Illumina sequencing with unique molecular indexes (colored ends attached to red/green 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.

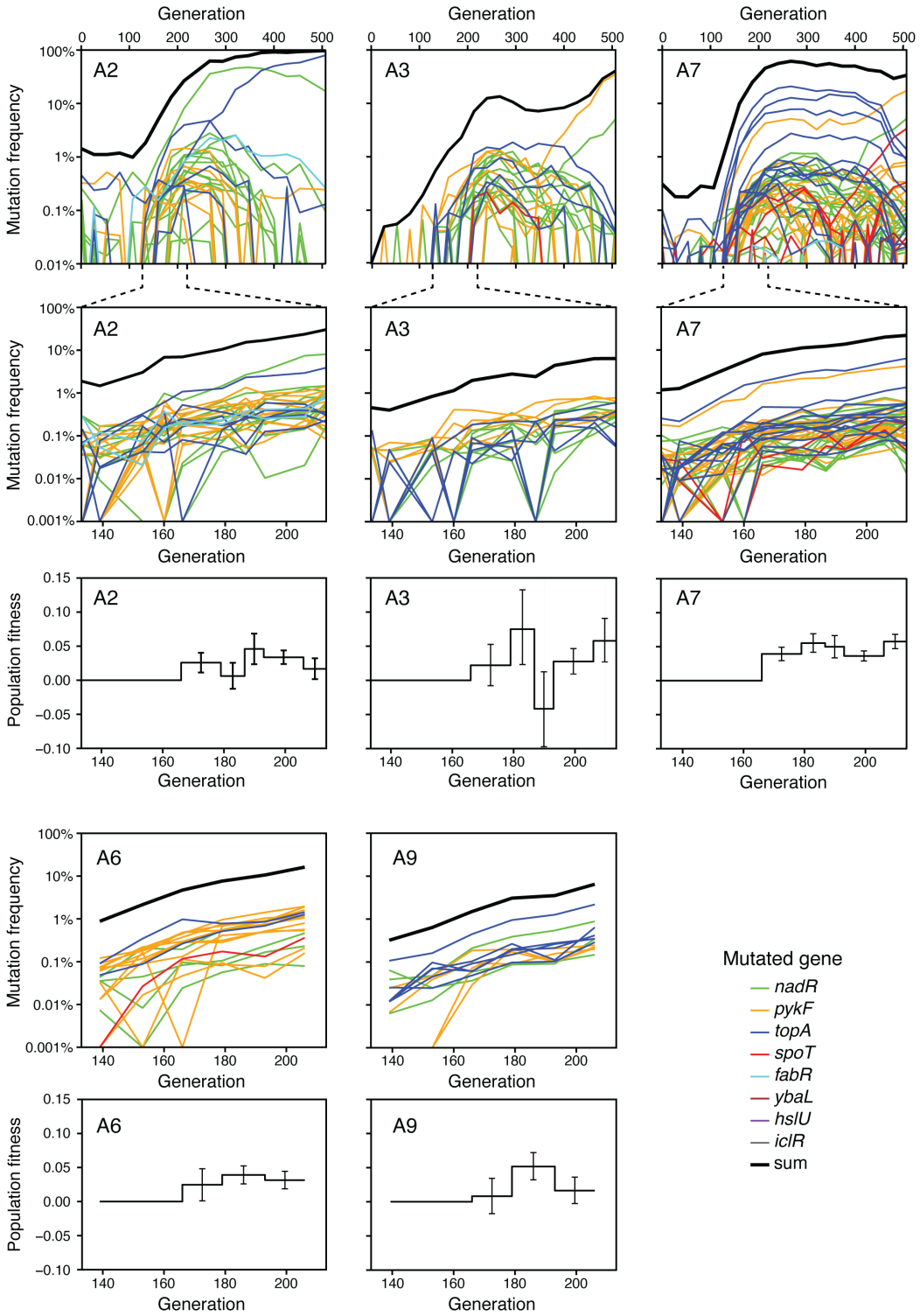


Figure 3. Frequency trajectories of mutations in the remaining populations. The same plots described in Figure 2C-E for population A1 are shown for populations A2, A3, and A7 (top three sets of panels). For populations A6 and A9, sequencing was only performed at time points during the selective sweep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

157 Mutation trajectories in all four populations exhibited a burst of genetic diversity in the
158 targeted genes followed by loss of this diversity. The initial dynamics are expected to be largely
159 driven by new genotypes that each evolve a single beneficial mutation very early in the
160 experiment. If their descendants escape stochastic loss, they will gradually increase in frequency
161 over the first few hundred generations as they outcompete the ancestral genotype. Once the
162 population becomes dominated by these first-step mutants, their frequency trajectories plateau
163 because of clonal interference: they are now mainly competing against one another and are
164 relatively evenly matched. In populations A1, A2, and A7, the total frequencies of the mutations
165 we identified sums to 50-62% at generation 270, indicating that each population is mostly
166 composed of genotypes with a single mutation in one of the focal genes. We recovered less of
167 the initial beneficial mutation diversity in population A3 where this sum was only 13%.
168 After around 300 generations, there is a steady decline in the frequencies of most mutations
169 in the eight targeted genes. At this point, new more-fit genotypes that have evolved from the
170 single-step mutants begin to exert their influence and displace them. Many of the most successful
171 second-step genotypes are descended from cells that already have a mutation in one of the
172 targeted genes. The original mutations serve as markers for the further expansion of these
173 subpopulations after a period during which their frequencies stagnate or decrease, but the new
174 beneficial mutations responsible for this further increase in fitness are outside of the genomic
175 regions we surveilled. The converse situation, in which a beneficial mutation in one of the eight
176 focal genes appears in a cell with an untracked beneficial mutation elsewhere in the genome, also
177 occurs in a few cases. Most strikingly, a new mutation in *pykF* that only appears after 300
178 generations in population A3 rapidly increases in frequency and becomes dominant, strongly
179 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)**.

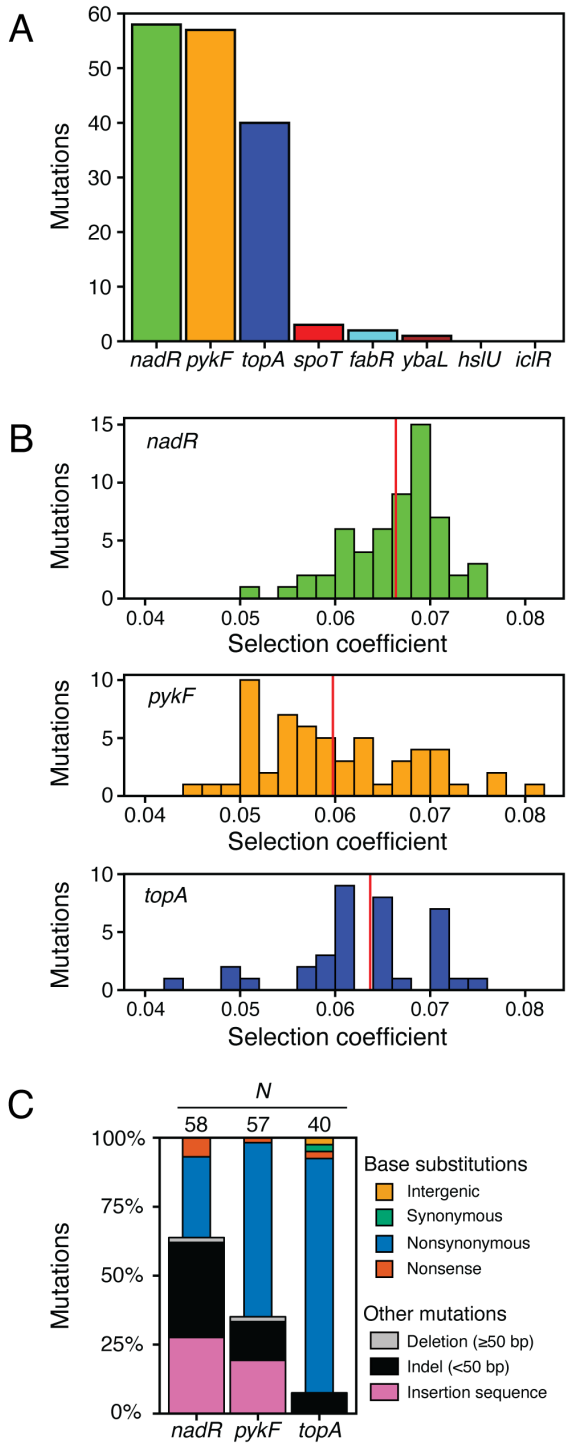


Figure 4. Characteristics of beneficial mutations in the initial selective sweep. (A) Total number of beneficial mutations in each targeted gene identified in the window time courses from 133 and 213 generations for all six *E. coli* populations that were profiled. (B) Distribution of selection coefficients 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.

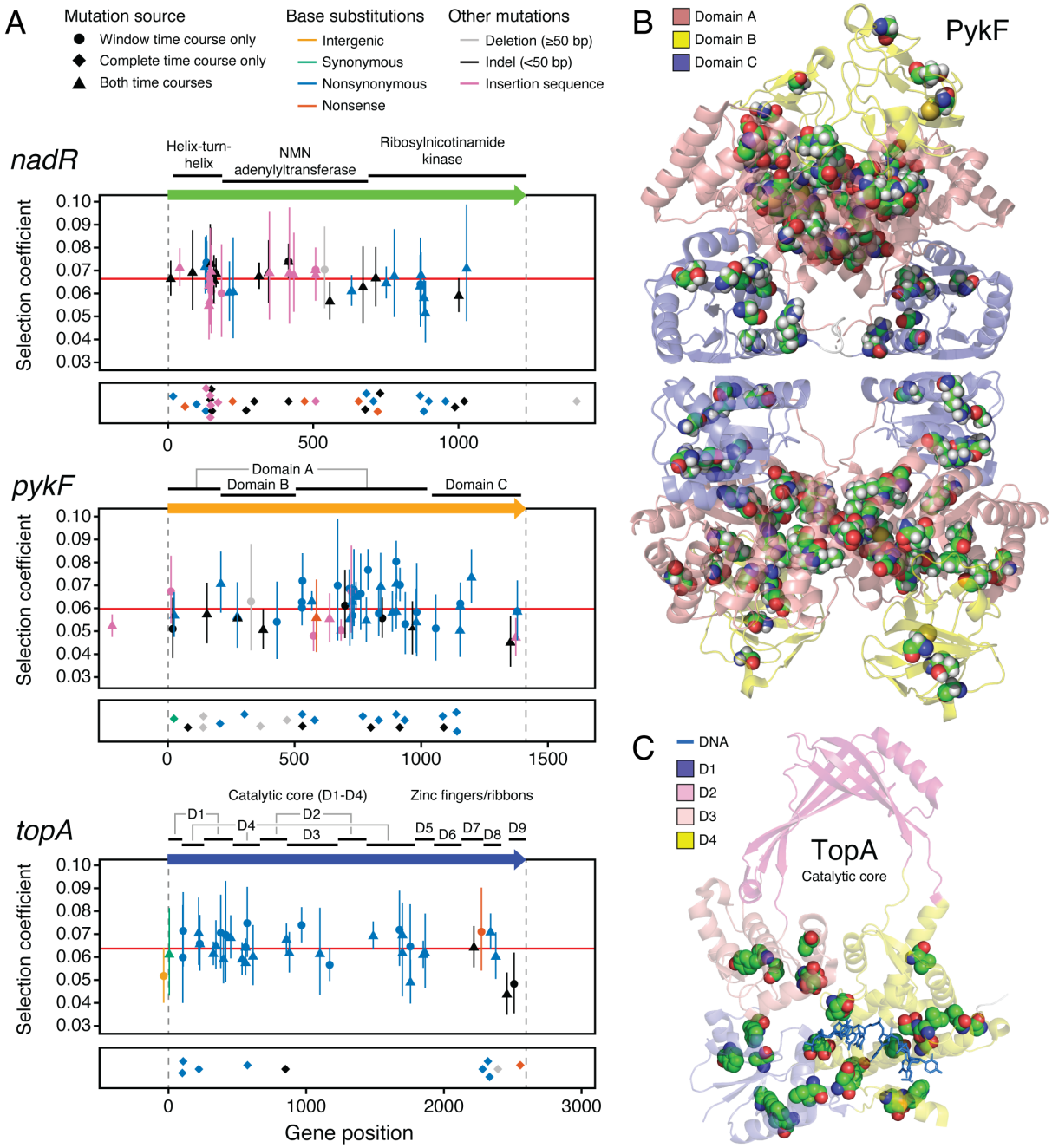
198 We were able to estimate a selection coefficient for each of the 161 mutations predicted in
199 the window time courses by fitting a binomial logistic model to the counts of reads supporting
200 the variant versus reference sequences over time. In all populations, there is a slight deceleration
201 in the rate at which the frequencies of the new mutations increase at generation 166 and later that
202 coincides with the onset of clonal interference (**Fig. 2E**, **Fig. 3**), At this point, genotypes with
203 beneficial mutations begin to make up an appreciable fraction of the population and compete
204 against one another rather than effectively only versus their ancestor. After correcting for this
205 increase in overall population fitness (see Methods), the mean selection coefficient that we
206 inferred for the tracked mutations in all six populations was 6.32% with a standard deviation of
207 0.74%. Although the distributions of selection coefficients estimated for mutations in *nadR*,
208 *pykF*, and *topA* overlap **(Fig. 4B)**, there was a significant stratification among these genes.
209 Mutations in *nadR* were 0.27% more beneficial than mutations in *topA*, on average, and this
210 difference was significant (p = 0.024, Kolmogorov-Smirnov test). In turn, mutations in *topA*
211 were 0.39% more beneficial than those in *pykF* (p = 0.00035, Kolmogorov-Smirnov test). The
212 six mutations in other genes (*spoT*, *yijC,* and *ybaL*) for which we were able to estimate selection
213 coefficients were roughly as beneficial as mutations in *nadR*, *pykF*, and *topA*.
214

215 **Beneficial mutations reveal different signatures of selection on gene function**

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

221 different types of mutations were also unevenly distributed throughout the sequences of these

222 three commonly hit genes and had noticeably different effects on bacterial fitness (**Fig. 5A**).



223

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

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

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

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

258 regulating glycolytic flux [37,38]. We observed an intermediate representation of disruptive
259 mutations in *pykF*, fewer than in *nadR* but more than in *topA* (**Fig. 4C**). Interestingly,
260 nonsynonymous base substitutions in *pykF* tend to have a larger selection coefficient than
261 disruptive mutations (p = 0.00390, Kolmogorov-Smirnov test) (**Fig. 5A**). This finding is in
262 agreement with a recent study of various *pykF* alleles that arose in the LTEE which found that
263 nearly all *pykF* point mutations were more beneficial than deletion of the *pykF* gene, both in the
264 ancestor and in evolved genetic backgrounds [39]. PykF forms a homotetramer in which each
265 polypeptide is folds into three structural domains [40,41]. The central domain C forms the active
266 site at the interface with domain B and the binding site for the allosteric effector fructose 1,6-
267 bisphosphate at the interface with domain A. The nonsynonymous mutations that we observed
268 are more concentrated than expected in domain C versus the other structural domains (p =
269 0.0050, one-tailed binomial test) (**Fig. 5B**). Overall, these results suggest that complete
270 inactivation of *pykF* is highly beneficial in the environment of our evolution experiment, but
271 mutations that alter its activity—likely in ways that reduce glycolytic flux—are even more so. It
272 has been suggested that reducing *pykF* activity is beneficial in the similar glucose-limited
273 conditions of the LTEE because this allows more PEP to be used for import of glucose into cells
274 by the phosphotransfer system [42].
275 DNA topoisomerase I (*topA*) relaxes negative supercoiling introduced into the chromosome
276 by replication and transcription [43]. The mutations we observed in *topA* are almost exclusively
277 single-base substitutions (**Fig. 4C**), suggesting that modulating the activity of this enzyme
278 provides a fitness benefit. Indeed, complete loss of *topA* function is lethal to *E. coli* without
279 compensatory mutations in DNA gyrase [44,45]. The structure of *E. coli* TopA consists of four
280 N-terminal domains (D1-D4) that make up the catalytic core and five C-terminal zinc finger and

281 ribbon domains (D5-D9) [46]. The few out-of-frame indels and the large deletion that we
282 observe truncate TopA within domains D7-D9, which interact with single-stranded DNA and
283 RNA polymerase but are not critical for catalysis. Considering only the catalytic core, we find
284 that nonsynonymous mutations are concentrated in domains D1 and D4 versus D2 and D3 (p =
285 0.0060, one-tailed binomial test) (**Fig. 5C**). D1 and D4 together form the ssDNA binding groove
286 leading to the active site, and D1 also forms part of the active site at its interface with D3 [47].
287 Several base substitutions in *topA* have been shown to increase positive supercoiling in evolved
288 LTEE strains [48,49]. The exact reason that this change in supercoiling is beneficial is unknown,
289 but it may be linked to increasing the expression of ribosomal RNAs [48], altering gene
290 regulation responses to starvation or stress [49], and/or increasing gene expression divergently
291 transcribed operons [50].
292

293 **Discussion**

294 We examined bacterial evolution during the initial stages of clonal competition when there is a
295 burst of beneficial genetic diversity as many new subpopulations with different mutations evolve
296 and begin to displace the ancestral genotype. We focused on eight genes known to accumulate
297 adaptive mutations in the >70,000 generation Lenski long-term evolution experiment (LTEE)
298 with *E. coli* that used nearly the same environment as our experiments. The only difference was
299 that we added four times as much of the limiting nutrient (glucose). By combining Illumina
300 sequencing of reads that incorporate molecular indexes for error correction, hybridization-based
301 capture of DNA encoding these genes, and dense temporal sampling, we were able to identify
302 more beneficial mutations and track them at much lower frequencies than is possible with
303 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

327 beneficial alleles that evolve in the LTEE [51]. Mutations in *spoT* and *fabR* were rarer than
328 expected from the LTEE. One possible explanation for these slight differences is the increased
329 concentration of glucose in our experiment compared to the LTEE. These minor deviations are
330 also reminiscent of how changing a different aspect of the environment (temperature) re-focuses
331 the mutations of largest benefit that succeed early onto different subsets of genes, nearly all of
332 which eventually accumulate beneficial mutations later in the LTEE environment, in related
333 evolution experiments [53,54]. Despite these subtle differences, we were still able to account for
334 majority of the genetic variation present in three of four of the four populations that we profiled
335 over the entire 500 generations by analyzing evolution in the eight candidate genes.
336 We also wanted to understand to what extent we gained early warning of driver mutations by
337 deeply profiling evolution in genes we expected to be under strong selection. In general, we were
338 able to begin tracking most mutations when they were above a frequency of 0.01%. This level of
339 profiling enabled us to first detect mutations an average of 75, 152, and 290 generations before
340 they surpassed frequencies of 0.1%, 1%, and 5%, respectively. Under the conditions of our
341 experiment these intervals take roughly 11, 23, and 44 days, respectively; so, even though we
342 made these predictions retrospectively, there would have been sufficient time to complete the
343 DNA isolation, library preparation, sequencing, and analysis steps quickly enough for this
344 approach to give early warning of specific genetic variants driving evolution of these
345 populations. The amount of lead time becomes disproportionately longer at higher frequencies
346 due to clonal interference between beneficial mutations. The chances and timescales of earlier
347 detection are expected to increase even more when there are ecological interactions or spatial
348 structure that further slow the takeover of new variants, as has been demonstrated and discussed
349 in other microbial evolution experiments [26,55,56].

350 A further prediction is that the genes in which we observe early, but unsuccessful beneficial
351 mutations will sustain mutations again and again until they are successful in a population's
352 evolutionary future. This prediction is limited by the nature of epistatic interactions. In the LTEE
353 and other microbial evolution experiments, diminishing returns epistasis dominates between
354 beneficial mutations in different genes [57–61]. That is, mutations in one gene that improve the
355 fitness of the ancestor tend to still be beneficial to evolved genotypes containing beneficial
356 mutations in other genes, just less so than when those other mutations are not present.
357 Subpopulations with mutations in both *nadR* and *pykF* evolve by 20,000 generations in all 12
358 LTEE populations, and cells that also contain a mutation in *topA* are found in six of the LTEE
359 populations at this point [31]. By this time, mutations in *ybaL* and *spoT* are also found in nine
360 and six LTEE populations, respectively. So, for five of the six genes in which we detected
361 multiple mutations in the initial burst phase, it is likely that nearly all of them would have
362 eventually accumulated beneficial mutations if we continued our experiment.
363 The genes in which we did not detect multiple mutations (*fabR*, *iclR*, and *hslU*) likely
364 represent other scenarios. Mutations in *fabR* transiently appear within the first 2,000 generations
365 of the LTEE [52]. They interact unfavorably with beneficial mutations in *spoT* and other genes,
366 such that a *fabR* mutation essentially precludes further adaptation by mutating the other set of
367 genes and vice-versa [52,62]. So, *fabR* mutations are unlikely to re-emerge in the future of these
368 populations. On the other hand, mutations in *iclR* and *hslU* appear to either require the presence
369 of mutations in other genes to become highly beneficial or may not be able to experience any
370 mutations that are beneficial enough to make them competitive early on in the LTEE. Of the 12
371 LTEE populations, 11 have cells mutations in *iclR* and 11 have mutations in *hslU* by 20,000
372 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

440 DNA Mini kit (Life Technologies). For each sample, 1 μg of gDNA was randomly fragmented
441 on a Covaris S2 focused-ultrasonicator.
442 Illumina libraries were constructed using the Kappa Biosystems LTP Library Preparation Kit
443 with the following modifications. End-repaired, fragmented DNA was T-tailed (rather than A-
444 tailed) in a 50 μl reaction including 10 mM dTTP and 5 units of Klenow fragment, exo^{-} (New
445 England Biolabs). Illumina adapters containing 12-base molecular indexes were ligated to the T-
446 tailed fragments as previously described [32], except full-length adapter sequences containing
447 unique external sample barcodes were directly ligated to the T-tailed dsDNA inserts to reduce
448 the risk of cross-contamination between samples. The full list of DNA sequence adaptors used is
449 provided in **Table S1**.
450
451 **Probe design and target capture**
452 Oligonucleotide probes consisting of 60-base xGen Lockdown probes (Integrated DNA
453 Technologies) were designed to tile across each of the eight genes of interest including upstream
454 promoter elements. Probes for each gene were compared to the entire *E. coli* B strain REL606
455 reference genome (GenBank: NC_012967.1) [69] using BLASTN [70]. The starting positions of
456 all probes in a set were shifted by one base at a time until every probe had only a single
457 significant predicted binding location (match with E-value $< 2\times10^{-5}$). The sequences of the final
458 set of 242 probes are provided in **Table S2**.
459 Capture was performed using a SeqCap EZ Exome Enrichment kit v3.0 (NimbleGen) with
460 several modifications to the protocol. First, 18 to 20 population samples with unique barcodes
461 were pooled together in a single capture reaction that contained a total of 1 μg of library DNA
462 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).

486

487 **Variant calling**

488 We used the *breseq* pipeline [52,72,73] (version 0.26.0) to call single-nucleotide variants (SNVs)
489 and structural variants (SVs) from the CSRs. We divided the genome sequence of the ancestral
490 *E. coli* REL606 strain into two types of reference regions for mapping in this analysis. The eight
491 regions of the genome tiled with probes—extended with hundreds of bases of flanking sequence
492 on both sides—were input as "targeted" sequences, and the remainder of the genome with the
493 identical eight regions masked to degenerate N bases was supplied as a "junction-only" reference
494 (to which reads are mapped without variant calling). All 116 samples were analyzed using
495 *breseq* in polymorphism prediction mode with all bias, minimum allele frequency, and read-
496 count filters disabled. Evidence items in the Genome Diff (GD) files for all samples were
497 combined using the *gdtools* utility program to generate a single merged GD file with each piece
498 of evidence listed a single time, regardless of how many times it was detected in different
499 samples. We then re-ran *breseq* using the same parameters except that this GD file was supplied
500 as an input user-evidence file to force output of variant and reference information for these
501 putative variants in every sample. Then, we extracted the number of variant reads supporting
502 each putative variant allele and the total number of reads at that reference location from the GD
503 file output by *breseq* and performed subsequent statistical tests and fitting steps in R (version
504 3.2.2) [74]. Scripts and data files for this analysis are available in GitHub
505 (https://github.com/barricklab/adaptome-capture).

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

509 also had frequency trajectories that tracked together. To identify these candidates for merging,
510 we analyzed each of the six window (generation 133 to 213) and four complete (generation 0 to
511 500) time courses separately. We only considered mutations that exceeded a threshold frequency
512 of 0.03% at some time during each time course as candidates for merging. Read alignment (RA)
513 evidence items were merged when they were located within 6 base pairs of one another and
514 within a normalized Canberra distance of 0.1 between the vectors of their frequency observations
515 across all of the time points in a dataset. All RA evidence pairs of this kind were found to co-
516 occur in the same sequencing reads. For these cases, the read counts for the first linked mutation
517 were used to represent the entire event. For example, if a deletion of three base pairs was
518 predicted by missing bases at positions x, y, and z; then the frequency of missing the first base
519 (x) was assigned to the entire three-base deletion mutation. For new junction (JC) evidence we
520 performed the same merging procedure but allowed linked mutations to be within a larger
521 window of 20 base pairs and within a normalized Canberra distance of 0.5. JC pairs passing
522 these criteria were only merged if they were also consistent with an IS-element insertion in terms
523 of their relative orientation and spacing. In this case the variant and total read counts were added
524 together for the two different junctions, as the junctions on each side of the inserted IS element
525 provide independent information for estimating the frequency of this type of mutation.
526
527 **Time course filtering and selection coefficient estimation**
528 After merging evidence of genetic variants into lists of putative mutations, we further eliminated
529 putative evolved alleles from consideration using several filtering steps. For the complete time
530 courses, we first required that non-zero frequencies be observed for a mutation in samples from
531 at least two different time points. We next applied a filter to eliminate spurious variants that can

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

536 For the window time courses, we further required that the estimated frequency of a putative
537 mutation be $\geq 10^{-4}$ at each of the last three time points that were sequenced (generations 193,
538 206, and 213). Then, we fit a binomial logistic model with slope and *x*-intercept terms to the time
539 courses of counts of variant and reference (total minus variant) observations for each mutation.
540 We filtered out any mutations for which this fit had an AIC < 200, Bonferroni-corrected *p*-value
541 for the slope differing from zero of > 0.05, or an *x*-intercept < -15. The slope fit from the
542 frequency trajectory of each mutation is an estimate of the selection coefficient of each mutation,
543 assuming the trajectories reflect competition purely against the ancestral strain. However, in the
544 latter half of the window time courses we detected a significant deviation from linearity for all
545 mutation trajectories, indicating that the overall population fitness had increased to a degree that
546 it decreased the rate at which all newly evolved genotypes with beneficial mutations increased in
547 frequency. The figures show the best models for a stepwise increase in population fitness
548 between the sequenced time points that improved the fits for all mutations in each population
549 considered separately. Because there was significant uncertainty in these estimates and the
550 fitness trajectories are expected to be highly similar between different populations, we used a
551 consensus model with one step-wise increase in fitness over time that best improved the fits for
552 all mutations from all populations to correct the estimated selection coefficients for this effect.
553

554 **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

578 Conceptualization: DED JEB.

579 Data Curation: DED JEB.

580 Funding Acquisition: DED JEB.

581 Investigation: DED.

582 Methodology: DED JEB.

583 Software – DED JEB.

584 Supervision – JEB.

585 Visualization: DED JEB.

586 Writing – Original Draft Preparation: DED JEB.

587 Writing – Review & Editing: DED JEB.

588

589 **References**

590

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

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

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

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

601 5. Park S-C, Krug J. Clonal interference in large populations. Proc Natl Acad Sci U S A.
602 2007;104: 18135–18140.

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

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

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

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

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

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

617 11. Landau DA, Carter SL, Stojanov P, McKenna A, Stevenson K, Lawrence MS, et al.
618 Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. Cell.
619 2013;152: 714–26. doi:10.1016/j.cell.2013.01.019

620 12. Ding L, Ley TJ, Larson DE, Miller C a., Koboldt DC, Welch JS, et al. Clonal evolution in
621 relapsed acute myeloid leukaemia revealed by whole-genome sequencing. Nature.
622 2012;481: 506–10. doi:10.1038/nature10738

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

626 14. Winstanley C, O'Brien S, Brockhurst MA. *Pseudomonas aeruginosa* evolutionary
627 adaptation and diversification in cystic fibrosis chronic lung infections. Trends Microbiol.
628 2016;24: 327–337. doi:10.1016/j.tim.2016.01.008

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

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

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

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

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

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

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

647 22. Zelder O, Hauer B. Environmentally directed mutations and their impact on industrial
648 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–

688 1247.

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

692 38. Kochanowski K, Volkmer B, Gerosa L, Van Rijsewijk BRH, Schmidt A, Heinemann M.
693 Functioning of a metabolic flux sensor in *Escherichia coli*. Proc Natl Acad Sci U S A.
694 2013;110: 1130–1135. doi:10.1073/pnas.1202582110

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

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

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

706 42. Woods R, Schneider D, Winkworth CL, Riley MA, Lenski RE. Tests of parallel molecular
707 evolution in a long-term experiment with *Escherichia coli*. Proc Natl Acad Sci U S A.
708 2006;103: 9107–9712. doi:10.1073/pnas.0602917103

709 43. Massé E, Drolet M. Relaxation of transcription-induced negative supercoiling is an
710 essential function of *Escherichia coli* DNA topoisomerase I. J Biol Chem. 1999;274:
711 16654–16658. doi:10.1074/jbc.274.23.16654

712 44. Dinardo S, Voelkel KA, Sternglanz R, Reynolds AE, Wright A. *Escherichia coli* DNA
713 topoisomerase I mutants have compensatory mutations in DNA gyrase genes. Cell.
714 1982;31: 43–51. doi:10.1016/0092-8674(82)90403-2

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

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

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

723 48. Crozat E, Philippe N, Lenski RE, Geiselmann J, Schneider D. Long-term experimental
724 evolution in *Escherichia coli*. XII. DNA topology as a key target of selection. Genetics.
725 2005;169: 523–532.

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

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

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

732 51. Ostrowski EA, Woods RJ, Lenski RE. The genetic basis of parallel and divergent
733 phenotypic responses in evolving populations of *Escherichia coli*. Proc R Soc B.
734 2008;275: 277–284. doi:10.1098/rspb.2007.1244

735 52. Deatherage DE, Traverse CC, Wolf LN, Barrick JE. Detecting rare structural variation in
736 evolving microbial populations from new sequence junctions using *breseq*. Front Genet.
737 2015;5: 468. doi:10.3389/fgene.2014.00468

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

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

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

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

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

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

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

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

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

762 62. Woods RJ, Barrick JE, Cooper TF, Shrestha U, Kauth MR, Lenski RE. Second-order
763 selection for evolvability in a large *Escherichia coli* population. Science. 2011;331: 1433–
764 1436.

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

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

768 64. Fischer S, Greipel L, Klockgether J, Dorda M, Wiehlmann L, Cramer N, et al. Multilocus
769 amplicon sequencing of *Pseudomonas aeruginosa* cystic fibrosis airways isolates
770 collected prior to and after early antipseudomonal chemotherapy. J Cyst Fibros. 2017;16:
771 346–352. doi:10.1016/j.jcf.2016.10.013

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

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

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

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

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

783 70. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+:
784 architecture and applications. BMC Bioinformatics. 2009;10: 421. doi:10.1186/1471-
785 2105-10-421

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

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

791 73. Barrick JE, Colburn G, Deatherage DE, Traverse CC, Strand MD, Borges JJ, et al.
792 Identifying structural variation in haploid microbial genomes from short-read
793 resequencing data using *breseq*. BMC Genomics. 2014;15: 1039. doi:10.1186/1471-2164-
794 15-1039

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

797 75. Wickham H. ggplot2: Elegant Graphics for Data Analysis. New York: Springer-Verlag;
798 2016.