bioRxiv preprint doi: https://doi.org/10.1101/429175; this version posted September 27, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

A high resolution view of adaptive events

Han Mei[1], Barbara Arbeithuber[2], Marzia A. Cremona[3], Michael DeGiorgio[2] and Anton Nekrutenko[1]

[1]Dept. of Biochemistry and Molecular Biology, The Pennsylvania State University
[2]Dept. of Biology, The Pennsylvania State University
[3]Dept. of Statistics, The Pennsylvania State University

bioRxiv preprint doi: https://doi.org/10.1101/429175; this version posted September 27, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Abstract

Coadaptation between bacterial hosts and plasmids often involves a small number of highly reproducible mutations. Yet little is known about the underlying complex dynamics that leads to such a single "correct" solution. Observing mutations in fine detail along the adaptation trajectory is necessary for understanding this phenomenon. We studied coadaptation between *E. coli* and a common artificial plasmid, pBR322, in a continuous turbidostat culture. To obtain a high resolution picture of early adaptive events, we used a highly sensitive duplex sequencing strategy to directly observe and track mutations with frequencies undetectable with conventional methods. The observed highly reproducible trajectories are governed by clonal interference and show rapid increases in the frequencies of beneficial mutations controlling plasmid replication followed by a profound diversity crash corresponding to the emergence of chromosomal variants. To the best of our knowledge our study represents the first comprehensive assessment of adaptive processes at a very fine level of resolution. Our work highlights the hidden complexity of coadaptation, and provides an experimental and theoretical foundation for future studies.

bioRxiv preprint doi: https://doi.org/10.1101/429175; this version posted September 27, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Introduction

Mutations continuously feed prokaryotic genetic diversity, which, when exploited by selection, enables microbes to colonize most environments, including our own bodies, and to defy antibiotic agents and host defences. Our understanding of these adaptive processes has been greatly advanced by experimental studies in bacteria [1–4], bacteriophages [5–7], yeast [8–10], and other systems (for review see [11–13]). The fate of beneficial mutations in clonal asexual populations depends on their supply (a product of population size and the mutation rate) and the distribution of fitness effects. If the supply of beneficial mutations is low relative to the selection strength, then a new variant quickly spreads and fixes. On the other hand, if mutational supply is high, then multiple coexisting variants interfere with each other—a phenomenon initially termed *periodic selection* [14] and known today as clonal interference [15–17]. Because population sizes of clonal asexual organisms are typically very large, there is a sufficient supply of beneficial mutations so that their frequencies are governed by clonal interference. Thus understanding the fine dynamics of clonal interference is necessary to predict adaptive events in such organisms [18,19].

To gain this understanding, one would need to try tracking all mutations within a population as soon as they arise. This may be technically challenging. The emergence of modern sequencing technologies has created an opportunity to explore these underlying dynamics [2,4,7]. However, these techniques only allow the observation of evolutionary trajectories of mutations that have risen to frequencies of ~1%—a threshold at which the noise of Illumina sequencing (the most accurate of the currently available sequencing technologies) obscures the signal [20,21]. Such a threshold is too high to obtain reliable insight into the underlying genetic variability within a large population. For example, Good et al.[2] used one milliliter of the overnight *E. coli* culture for sequencing time points from the long term evolution experiment. This volume conservatively contains ~10^8 cells. As a result, at sequencing resolution of 1%, one will miss all sequence variants present in fewer than ~10^6 cells. A profound gain in resolution has been made by the application of molecular barcodes within an experimental evolution setting [9,10,22]. These studies have shown that at the beginning of an experiment, a large pool of lineages carrying beneficial mutations is present within the frequency band between 10^{-8} and 10^{-4}, and a vast majority of adaptive lineages have frequencies under 0.01% (well below what is detectable with conventional Illumina sequencing). These data provided us with invaluable insight into early adaptive events. However, they do not convey the nature of genetic variants, as the barcode counts are only proxies for the number of cells carrying a particular, yet unknown, beneficial mutation. Venkataram et al. [22] recognized this limitation and used barcoding to select several hundred lineages carrying beneficial

bioRxiv preprint doi: https://doi.org/10.1101/429175; this version posted September 27, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

mutations for subsequent whole genome sequencing. Still, this is only the tip of the iceberg, with the majority of variants remaining below the surface.

A number of techniques have been recently developed to allow sequencing-based detection of very rare genetic changes that may be suitable to directly observe early adaptive events. Of these, duplex sequencing is the most sensitive, with a theoretical resolution threshold of $<10^{-7}$ [21,23]. It is based on using unique sequence tags to label individual molecules of the input DNA prior to preparation of Illumina sequencing libraries. During the amplification steps of library preparation, each of these molecules gives rise to multiple descendants. After sequencing, the descendants of each original DNA fragment are identified and grouped together using tags—i.e., one simply sorts tags in sequencing reads lexicographically, and all reads containing the same tag are bundled into *families*. These families (usually with at least three members) form single strand consensus sequences (SSCS) for the forward or the reverse strand, respectively. Complementary SSCS are then grouped to produce duplex consensus sequences (DCS). A true sequence variant is present in all reads within a family forming a duplex. In contrast, sequencing and amplification errors will manifest themselves as "polymorphisms" within a family, and so they can be identified and reliably removed. The resolution of duplex sequencing combined with the power of experimental evolution approaches provides a unique opportunity to directly observe adaptive events as they appear.

Observing emerging mutations and tracking their evolutionary trajectories requires an experimental system that satisfies two criteria. First, it must have a sufficiently small genome to allow for ultra-deep duplex sequencing. This is because the great dynamic range of duplex sequencing requires high depth (i.e., identifying variants with frequency of 10^{-6} requires sequencing depth of $>10^6$). Second, such a system needs to evolve on a timescale that fits into a short term experimental evolution setup, such as the one provided by chemostat or turbidostat devices. To satisfy these requirements, we selected one of the logistically simplest and exceptionally well understood systems—plasmid vector pBR322. It has a small genome of 4,361 bp, which is amenable to deep sequencing. As a synthetic plasmid it has no evolutionary history of being associated with *E. coli*. Transforming it into naïve *E. coli* cells triggers a series of adaptive events in the plasmid itself as well as in the host genome [24–28]. These changes take place within as few as several hundred generations, which is an easily achievable timeframe for an experimental evolution experiment. Because acquisition of the plasmid incurs a metabolic burden on the host cells, it reduces growth rate relative to their plasmid-free isogenic counterparts [27] and results in the rapid decline of transformed cells in antibiotic-free chemostat cultures [29]. pBR322 contains two antibiotic resistance genes encoding a β-lactamase and a TetC efflux pump, conferring resistance to

bioRxiv preprint doi: https://doi.org/10.1101/429175; this version posted September 27, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

ampicillin and tetracycline, respectively. Incubating transformants in the presence of these antibiotics eliminates plasmid loss as these genes become essential for host survival. Furthermore, the plasmid-encoded *tetC* gene plays a critical role in the host-plasmid co-adaptation. Propagation of naïve cells transformed with *tetC*-containing plasmid changes the host phenotype to a state where the plasmid carriage is no longer detrimental, and this change occurs in only 500 generations [24]. The evolved host has higher fitness in the presence of the plasmid, and this effect, caused by mutation(s) on the bacterial chromosome, requires a plasmid carrying *tetC* such as pBR322. Biochemically, this may be explained by *tetC* assuming other roles beyond tetracycline efflux (e.g., potassium uptake [28]).

The coevolution between a host and a plasmid provides a well defined experimental setting for direct observation of clonal interference. When a population (large enough to allow for clonal interference to drive adaptation) of naïve cells transformed with the plasmid is propagated within an antibiotic-containing medium, a number of competing beneficial mutations arise and follow the dynamics that was observed by Levy et al. [9] with multiple adaptive lineages co-existing at relatively low frequencies. Subsequently large effect mutation(s) sweep to fixation and lead to the crash of genetic variability within the population. Lenski et al. established that this large fitness effect change arises within ~500 generation[24]. Thus, by taking regular samples across the duration of the adaptation experiment, we were hoping to reconstruct the trajectory of adaptive events.

In this study we directly monitored the fate of beneficial mutations arising during coadaptation of a bacterial host and a plasmid in a continuous culture. Our objective was to test repeatability of early adaptation, to develop a simple deterministic model that would capture the observed behavior, and to generate predictions for future experiments.

Results and Discussion

Experimental evolution setup

We chose to use a turbidostat, a device that maintains constant cell density with no restrictions on the supply of nutrients, for this experimental evolution experiment, as it puts additional selective pressure on host cells to increase their maximum growth rate [30,31]. We selected *E. coli* DH5α F- *recA1* as the host, in which recombination is inhibited and conjugation is disabled. The cells were transformed with pBR322—a low copy number plasmid (15–20 per cell) [32,33] that minimizes the selection of traits favoring competition among the plasmid population [34] and that carries two antibiotic resistance loci. The turbidostat was set up in the following way. At time point zero, the turbidostat was inoculated with an

bioRxiv preprint doi: https://doi.org/10.1101/429175; this version posted September 27, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

aliquot of overnight DH5α culture transformed with pBR322. To establish the time course of the adaptation, approximately ⅔ of the volume (~8 mL) was taken from the incubation vessel every 12 hours, constituting time points 1, 2, 3, and so on. Plasmid DNA was isolated from that volume and subjected to duplex sequencing without any additional manipulation. Sampling such a relatively large volume allowed us to perform thorough population level sampling. Two experiments including two replicates each were performed: a short term (60 hours; replicates R1 and R2) and a long term (318 hours; replicates R6 and R7). Fig. 1 shows a graphical representation of our sampling strategy with the optical density at 600 nm (OD_{600}) dips corresponding to sampling points. Duplex sequencing was performed at 0, 12, 24, 36, 48, and 60 hour points in the short term experiments and at 72, 84, 108, 156, 240, and 318 hour points in the long term experiments. In each case, both replicates were sequenced for a total of 24 duplex sequencing datasets. In addition, we performed conventional high depth sequencing of the host genome before (naïve) and after (evolved) the long term experiments.

Emergence of polymorphisms followed by variation crash

Frequencies and positions of single nucleotide variants (SNVs) and insertions/deletions (indels) called from duplex sequencing are shown in Table S1 and Fig. 2. The initial, monomorphic, samples (time point zero, short term experiments, replicates R1 and R2) contained five SNVs supported by a single duplex family each (implying that our sample contained only a single molecule carrying each SNV). These sites did not propagate through the experiment and disappear in subsequent time points. We call such positions *flickering* sites, which likely represent neutral variants lost due to genetic drift. Flickering sites are present in most (but not all) time points, and the vast majority of them are supported by only one duplex family. The other group of variants is represented by likely adaptive SNVs that persist through the time points. None of them appear in the initial sample, and their frequencies increase initially (short term experiment) and start to drop off as the experiment progresses (long term experiment; Fig. 2) so that, at the terminal time point of one of the replicates of the long term experiment, pBR322 reverses completely to the wild-type. These SNVs cluster exclusively in two regions within the replication origin of pBR322: positions 3,027–3,035 and 3,118. In total, all experiments yielded 145 SNVs. The mutational spectrum (Fig. S1) showed a transition-to-transversion ratio of 1:1.9, which matches that of the long term evolution experiment of Wielgoss et al.[35]. One insertion and two deletions were identified in total (Table S1). The insertion was located at the replication origin, and was observed only in replicate R7 of the long experiment. One deletion that was located outside the origin was found in multiple turbidostat runs. The other deletion was within the origin, and was only observed in the R1 replicate of the short

bioRxiv preprint doi: https://doi.org/10.1101/429175; this version posted September 27, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

experiment. Indels did not exhibit any frequency fluctuation, which could be due to the rate of indel mutations being generally lower than that of base-pair substitutions [36–38].

Because the distance between SNVs located within the replication origin is shorter than the length of duplex reads, we examined our data for evidence of linkage. No linkage between any SNVs was observed. This is not unexpected because given the *E. coli* mutation rate is ~10^{-10}–10^{-9} per nucleotide per generation [35–38] and the highest frequency attained by these SNVs is ~0.9% it is highly unlikely to have two mutation events within a single molecule.

Variable sites control plasmid copy number

The mechanism controlling pBR322 replication has been well studied [39]. Two complementary RNA molecules are encoded within the *ori* region: RNAI and RNAII. RNAII is first transcribed and hybridizes with plasmid DNA to form a primer complex for DNA replication. The other molecule—RNAI—is transcribed from the complementary strand. RNAI inhibits the formation of RNAII-DNA complexes, a replication primer, by interacting with RNAII via three loop regions. Substitutions within region 3,027-3,035 (Fig. S2A) were reported to increase pBR322 copy number [40]. This interval represents loop II' of RNAI. Meanwhile, base 3,118 corresponds to the -35 promoter region of RNAII (Fig. S2B). Changes at this base were reported to increase the copy number as well [41]. The absolute majority of all variants found in this study are confined to these regions.

Evolved cells have higher fitness conferred by chromosomal mutations

At each collection point we sampled ⅔ of volume from the incubation vessel, which was immediately topped-off with sterile medium diluting the remaining cells and causing optical density dips in Fig. 1. In each case, OD_{600} values quickly returned to the 0.8 threshold. The speed of this recovery can be used as a relative fitness estimate: the steeper the slope, the fitter the cells. This approach demonstrated consistent fitness increases over the course of turbidostat runs (Figs. S3 and S4, left panels). Contrasting the fitness of terminal clones from replicates R6 and R7 against the clones from the initial time point zero stocks confirmed this pattern (Figs. S3 and S4, right panels). The fact that pBR322 has completely reverted to wild-type in R6 (Table S1) and was on the way to purge all variation in R7 suggests that the increase of fitness can be attributed to changes within the bacterial chromosome. Whole genome sequencing of initial and terminal clones from the long term experiments revealed a high frequency nonsynonymous T-to-A transversion in *treB* (Val112Glu) in both replicates (Table S2). In addition, R6

bioRxiv preprint doi: https://doi.org/10.1101/429175; this version posted September 27, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

possessed another nonsynonymous substitution in *recA* (Asp161Gly) that effectively re-establishes the RecA1 genotype, back to RecA.

The product of the *treB* gene transports trehalose into the cell as trehalose 6-phosphate, which is further converted to glucose 6-phosphate and glucose [42]. In addition to its role as a carbon source, intracellular trehalose can act as an osmoprotectant at high osmolarity [43–45]. The T-to-A transversion in *treB* observed in our work has been suggested to be a consequence of adaptation of *E. coli* to LB medium and has been identified in three parallel populations in two recent mutation accumulation experiments [46,47].

The *recA1* → *recA* reversal in replicate R6 is noteworthy, as it is close to fixation (frequency of 0.91; Table S1) and reinstates recombination [48–50]. Coincidentally, the same replicate R6 of the long experiment purges plasmid variation completely at the end of sampling, while the other replicate, R7, which did not experience *recA1* → *recA* reversal continues to harbor lingering variation within the plasmid, and has a much lower frequency of the *treB* variant. Recombination has been shown to accelerate adaptation in yeast populations by combining beneficial mutations from different backgrounds and by separating deleterious mutations apart [51]. However, none of the variants we observed are linked, thus it is unclear if RecA has any direct effect on the pattern observed here.

Modelling the dynamics of plasmid mutations

The observations we described thus far suggest the following scenario. A naïve host is transformed with a plasmid conferring antibiotic resistance. Incubation of the transformed cells in an antibiotic containing environment guarantees retention of an otherwise disposable plasmid. The high density of the turbidostat environment creates a strong selection pressure for shortening generation time. Initially, selection drives the increase in the frequencies of adaptive changes controlling plasmid copy number, as it translates into the elevated production of the efflux pump, thus helping to offset the inhibitory effect of tetracycline on the protein synthesis. This increases the fitness of cells carrying the mutated pBR322 relative to the wild-type plasmid. However, higher copy number has an associated metabolic cost—once it reaches a certain frequency threshold, it proves too expensive and, in fact, eventually decreases fitness. Simultaneously, a mutation with a large fitness effect arises on the bacterial chromosome. Given the frequencies observed in the experiment, it is highly unlikely that such a chromosomal mutation will arise on the background of an existing plasmid mutations. Its spread to fixation obliterates plasmid variation, as we observed in Fig. 2.

bioRxiv preprint doi: https://doi.org/10.1101/429175; this version posted September 27, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

To solidify this reasoning into a quantitative framework, we developed a model drawing on the previous work of De Gelder et al. [52]. Assuming that there is no back mutation on both the plasmid and the chromosome, we denote the number of cells with pBR322 bearing mutations at a certain generation n by m_n, and those with wild-type pBR322 by w_n. At any generation, m_n increases due to (1) doubling of mutant cells from w_{n-1} occurring at constant rate μ (number of individual cells per generation) and (2) doubling of m_{n-1} with a selection coefficient $\sigma_1^{(n-1)}$. Putting these two components together the number of mutation-carrying cells at generation n is given by

$$(1) \qquad m_n = 2\mu w_{n-1} + 2^{1+\sigma_1^{(n-1)}} m_{n-1} .$$

The selection coefficient $\sigma_1^{(n-1)}$ changes from generation to generation due to two factors. First, adaptive plasmid variants increase pBR322 copy number leading to better tetracycline removal ability. This factor is denoted by the constant positive coefficient b. Second, increased plasmid copy number imposes a higher burden, inversely proportional to plasmid frequency—characterized as a negative frequency-dependent manner. This effect is denoted by the positive constant a. Therefore $\sigma_1^{(n-1)}$ can be expressed as $\sigma_1^{(n-1)} = b - am_{n-1}$. At the same time, the number of cells containing wild-type plasmid w_n increases from previous generation of w_{n-1} as given by

$$(2) \qquad w_n = 2^{1+\sigma_2^{(n-1)}}(1-\mu)w_{n-1} ,$$

where the selection coefficient $\sigma_2^{(n-1)}$ reflects beneficial mutations accumulating on the chromosome and is given by $\sigma_2^{(n-1)} = c(n-1) + d$, where c and d are positive constants.

At each sampled time point we can directly observe the fraction of mutants in the population (denoted as $\beta^{(n)}$) as the sum of alternative allele frequencies at sites 3,027–3,035 and 3,118. This corresponds to $\frac{m_n}{m_n+w_n}$, where m_n and w_n are given by equations (1) and (2), and thus $\beta^{(n)}$ is a function of the two selection coefficients $\sigma_1^{(n-1)}$ and $\sigma_2^{(n-1)}$, and the mutation rate μ:

$$\beta^{(n)}(\sigma_1^{(n-1)}, \sigma_2^{(n-1)}, \mu) = \frac{m_n}{m_n+w_n} .$$

Fitting this model to the empirical data as described in *Material and Methods* helped us explain the observed dynamics of our system. Briefly, we empirically estimated $\beta^{(n)}$ from the observed allele frequencies at each sampling point. We then set biologically plausible intervals of possible values for the a, b, c, d, and μ parameters, and employed numerical optimization to infer a set of parameters

bioRxiv preprint doi: https://doi.org/10.1101/429175; this version posted September 27, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

minimizing the differences between simulated and empirical values of $\beta^{(n)}$. Results of this optimization are given in Table 1. Allele frequencies reconstructed using the described model with optimal parameters closely follow the observed dynamics (Fig. 3) and allow us to make and experimentally test predictions about the behaviour of this system. In particular, it allows for predicting the effects of the increases in tetracycline concentration (Fig. S6) and mutation rate (Fig. S7) on variant frequencies and their trajectories. To model the increase in the tetracycline concentration we decreased the *c* component of the $\sigma_2^{(n-1)}$ selection coefficient. This effectively increases the beneficial effect of plasmid mutations expressed by $\sigma_1^{(n-1)}$ selection coefficient. This increased the maximum frequency of plasmid mutations and prolonged their persistency before ultimately yielding to chromosomal changes. Likewise the effect of increase of the mutation rate allowed for even higher frequencies of plasmid variants. While this dynamics is inherent to our model it provides an insight into the magnitude of selection coefficients and the temporal scale at which plasmid variants remain in the population before being eliminated. This information leads to a series of future experiments our group will be undertaking.

Plasmid heteroplasmy likely prevents alterations in copy number

The dynamics described here initially favors beneficial changes within the plasmid DNA that are quickly overridden by chromosomal mutations that attain high frequency in the evolving population. As a result of this co-adaptation process, the plasmid remains unchanged. A number of experimental evolution studies report similar findings where co-adaptation is driven exclusively by chromosomal mutations and leaves plasmids unaltered [24,53–56]. However, it is not entirely clear why this is the case. At least for pBR322 it is not unreasonable to assume that some combination of mutations affecting plasmid replication could successfully balance tetracycline efflux with copy number burden and reach high frequency at least in a short term. The fact that this does not happen could be a consequence of pBR322 segregation dynamics—the plasmid is partitioned between daughter cells randomly and likely unequally. As pBR322 copy number is only ~20 copies per cell, this sampling error may be significant. Because mutations occur during replication of just one of multiple plasmids copies present within a single cell, this cell is rendered *heteroplasmic*—it contains a mixture of mutated and wild-type pBR322 molecules. The stochasticity of partitioning of plasmid molecules during cell division means that this ratio will be different in each of the daughter cells (also see [57]). In others words they will be heteroplasmic to a different degree. Recent data from another ColE1 plasmid (pB1000 in *Haemophilus influenzae*) shows heteroplasmic cells to have higher overall number of plasmid copies compared to cells that are monomorphic for the mutated plasmid (here, like in our case, the mutation leads to copy number

bioRxiv preprint doi: https://doi.org/10.1101/429175; this version posted September 27, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

increase)[58]. Thus the differences in the extent of heteroplasmy (a mutant-to-wild-type ratio within a cell) will lead to different plasmid copy numbers within the daughter cells. This would explain the behavior we observe in our system. Our experiment is performed at a constant antibiotic concentration. This means that the maximum beneficial effects of the increase in the number of efflux pumps is likely associated with a single, "optimal" number of plasmid copies per cell, which balances incurred metabolic cost. It is likely that such optimum is only achieved in heteroplasmic cells because monomorphic mutated cells may produce too many copies of the plasmid to remain competitive in the population. In addition, RNAI, which contains most mutations in our study, acts in *trans* by binding to RNAII at the origin of replication. As a result even a single mutated molecule may influence replication of wild-type molecules within a cell. Thus there must be a narrow optimal value for the degree of heteroplasmy that translates into the optimal pBR322 copy number conferring selective advantage in a given antibiotic concentration. However, the very nature of pBR322 segregation makes it very difficult to reliably establish cells with this "optimal" number in the population because every cell division effectively randomizes the relative numbers of mutated and wild-type plasmids in the next generation. As a result the positive effect of the increase in the copy number is only achieved along a precise optimal value ridge of the adaptive landscape. Every time a cell divides its descendants fall from the ridge because they produce either too few or too many plasmid copies. Likewise ascending the ridge is only possible by chance when a cell division results in just the right degree of heteroplasmy. Because the bacterial populations we are sampling are large, we are able to observe these events until a large fitness effect mutation obliterates the variation by changing the topography of the adaptive landscape (the variation is likely still there but cannot be observed with our method). Addressing these issues will require development of technique allowing tracing individual plasmids in bacterial population (e.g., a finer resolution of the approach recently reported by Rodriguez-Beltran et al. [59]).

In conclusion we have applied the highly sensitive duplex sequencing approach to trace the evolution of pBR322 plasmid during its adaptation to *E. coli* DH5α host. We uncovered previously hidden fine scale dynamics where multiple mutations leading to the increase in the copy number compete ultimately losing to a single chromosomal substitution. Our results suggest that there is a strong incentive not to alter copy number even if it can provide a degree of selective advantage. This incentive is likely rooted in the complex interplay between mutated and wild-type plasmids constrained within a single cell and underscores the importance of understanding of intracellular plasmid variability.

bioRxiv preprint doi: https://doi.org/10.1101/429175; this version posted September 27, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Material and Methods

Strains and plasmids

E. coli strain DH5α was obtained from Invitrogen (cat. 18265017). The cells were transformed with plasmid pBR322 (NEB cat. N3033S) according to the manufacturer's protocol. pBR322 carries the replication origin, *tetC*, encoding a tetracycline efflux pump, *bla*, conferring resistance to ampicillin, and *rop*, controlling replication [60]. All experiments were performed in Luria-Bertani (LB) broth (EMD cat. 110285), including those in turbidostat and agar plates, that was supplemented with 30 μg/mL tetracycline (Sigma-Aldrich cat. T3258) and 0.05% antifoam B (Sigma-Aldrich cat. A5757). Transformants were spread over a LB agar plate. A series of single colonies were picked, grown in LB, and then stored in 20% glycerol stock in a -80°C freezer. One glycerol stock was used as the founder to inoculate all turbidostat runs.

Turbidostat set-up and experimental evolution

The turbidostat set-up was described by Takahashi et al. [61]. It consists of three connected parts (Fig. S5): a carboy containing LB medium, a bacteria growing chamber that is revolved magnetically, and a waste tank. The carboy and the chamber were pressurized by an aquarium air pump. The volume in the chamber was kept constant at 13 mL. OD_{600} was measured in 30 second intervals and LB was added as necessary to maintain OD_{600} constant at 0.8.

Fifty μL of the glycerol stock was inoculated into 13 mL of LB medium and grown for 14 hours at 250 rpm, 37°C. Four mL cell culture was inoculated into turbidostat. One mL culture was frozen as glycerol stock in -80°C as the initial sample. The remaining culture was labelled as time point 0 and collected by centrifugation at 4,000 g for 5 minutes. The pellet was then transferred to -20°C. Turbidostat runs were carried out in an incubator to maintain 37°C. Eight mL culture was taken from the turbidostat at regular 12 hour intervals, which constitute time points for the subsequent analysis. The turbidostat was immediately refilled by fresh broth to the 13 mL mark. When the turbidostat run was completed, the terminal sample was also collected and archived as glycerol stock in -80°C.

bioRxiv preprint doi: https://doi.org/10.1101/429175; this version posted September 27, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Population fitness determination

OD_{600} values were constantly monitored, and growth curves were inferred from them. We found that exponential growth between OD_{600} values of 0.6 to 0.8 corresponded to the maximum growth rate, and therefore we used it as a proxy for overall population fitness. Growth rate was defined as

$$\ln(0.8/0.6)/t$$

where t is the duration in minutes in which OD_{600} values increased from 0.6 to 0.8. The generation time was calculated as

$$\ln(2)/\text{growth rate}$$

We found that the generation time varied over the course of the two long term replicates. Detailed information regarding growth rate and generation time are displayed in Table S3. For simplicity, the generation time was assumed as 60 min when we modeled the dynamics of variants (Fig. 3).

Fitness re-measurement of samples representing the initial and terminal time points

Fifty µL glycerol stock from each of the initial and terminal samples from the two long term replicates R6 and R7 was inoculated into turbidostat in triplicate. OD_{600} was monitored and growth curves were inferred. Exponential growth between OD_{600} values of 0.6 to 0.8 was used to estimate fitness.

Duplex sequencing and calling plasmid variants

Plasmid at selected time points was extracted by mini-prep (Qiagen cat. 27104). Duplex sequencing library preparation and variant calling were performed as previously described [20,62]. Briefly, 100 ng of plasmid was sheared to 550 bp on a M220 platform (Covaris) according to the manufacturer's instructions at the Pennsylvania State University Genomics Core Facility. The sheared DNA was subjected to end-repair (NEB cat. E6050S), size selection (Beckman Coulter cat. A63881), 3'-end dA-tailing (NEB cat. M0212L), duplex adaptor ligation (NEB cat. M0202T), and PCR amplification (Kapa Biosystems cat. kk2612). PCR amplicons were quantified by qPCR (Kapa Biosystems cat. kk4873) and sequenced on an Illumina MiSeq platform using 301-nt paired-end reads. DCS were generated from raw reads and mapped to the reference to call variants. The variant calling workflow is available on Galaxy [63] at https://usegalaxy.org/u/hanmei/w/du-var-calling.

bioRxiv preprint doi: https://doi.org/10.1101/429175; this version posted September 27, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Whole-genome sequencing and calling chromosomal variants

Genomic DNA was extracted (Qiagen cat. 51304) from samples representing the initial and terminal time points of the two long term replicates. Whole-genome sequencing was performed as described in our previous article [64]. Briefly, sequencing libraries were prepared as described for duplex libraries, except that 2 μg of DNA was used as the starting material for shearing and adaptors from the Illumina TruSeq Kit were used. Without PCR amplification, ligated libraries were directly subjected to MiSeq sequencing. Variants were called against the DH5α reference genome (GenBank accession number CP017100) using the haploid variant calling workflow on Galaxy at https://usegalaxy.org/u/hanmei/w/haploid-var-calling.

Data modeling

We first defined a *data_generation* function to simulate data based on the mathematical model described by equations (1) and (2) in *Modelling the dynamics of plasmid mutations*. Given the values of the five parameters (a, b, c, d, μ), this function returned numbers of mutants m_n, number of wild-type cells w_n, and mutant frequency $\frac{m_n}{m_n+w_n}$ over 318 generations. We also empirically estimated the mutant frequency at the generations corresponding to the time points selected for plasmid duplex sequencing. In particular, at each of these empirically determined generations, we computed the empirical mutant frequency as the sum of frequencies of mutations in bases 3,027-3,035 and 3,118 determined in turbidostat experiments. We then defined the *loss* function as the sum of squared differences between simulated and empirical mutation frequencies.

This *loss* function was optimized over the five parameters (a, b, c, d, μ) using the R function *optim*, setting the bounds on the parameters as shown in Table S4 (choice of these bound is discussed in the next paragraph). Note that *optim* was run with a maximum of 10^6 iterations in order to facilitate convergence of the optimization process. Since the simulated mutant frequency shows a complex dependence on the parameters (a, b, c, d, μ), the *loss* function is likely to have multiple local minima. As a consequence, the set of optimal parameters (a, b, c, d, μ) returned by of the *optim* function strongly depend on their initial values. To effectively explore the parameter space, we randomly sampled initial values for the five parameters (uniformly on a $\log_{10}$ scale, according to the bounds in Table S2) and repeated the optimization process until 1,000 converged optimization processes were obtained. Each repetition returned a set of parameters and the corresponding *loss* value. The five values (a, b, c, d, μ) generating the smallest *loss* among the 1,000 obtained were chosen as optimal parameters and presented in Table 1. The

bioRxiv preprint doi: https://doi.org/10.1101/429175; this version posted September 27, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

steps implementing parameters optimization are described in the supplementary jupyter notebook "Figure_3_parameter_optimization.ipynb" at https://github.com/hanmei5191/pBR322-variant-dynamics.

We defined the five parameter intervals in logarithmic scale to ensure that they were wide enough to randomly sample the parameters. Therefore, a large parameter space was explored. The mutation rate μ in De Gelder et al. [52] was determined with the order of magnitude of 10^{-5}. We set the interval for μ from 10^{-10} to 10^{-2}. The other four parameters (*a*, *b*, *c*, *d*) were involved in calculation of $\sigma_1^{(n-1)}$ and $\sigma_2^{(n-1)}$. As coefficients, $\sigma_1^{(n-1)}$ and $\sigma_2^{(n-1)}$ were generally expected to be smaller than 1. We set the intervals for (*b*, *c*, *d*) the same as μ. The interval for parameter *a* was set on smaller values (from 10^{-14} to 10^{-6}) because it served as the coefficient of m_{n-1} and we tried to maintain the product am_{n-1} reasonably small.

bioRxiv preprint doi: https://doi.org/10.1101/429175; this version posted September 27, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Data deposition

Raw sequencing reads in this study have been deposited at the NCBI SRA database as BioProject PRJNA485503.

Supplementary files

Supplementary files include Table S1 (xlsx), Table S2 (xlsx), Table S3 (xlsx), and Table S4 (xlsx), jupyter notebooks and data to perform the parameter optimization shown in Fig. 3 and to generate predictions shown in Figs. S6 and S7. The notebooks can be found at https://github.com/hanmei5191/pBR322-variant-dynamics.

Acknowledgements

Authors are grateful to Jim Bull for his suggestions that have greatly improved our manuscript. Nick Stoler has aided in the tuning of the duplex data analysis pipeline. This study has been funded by the funds provided by the Eberly College of Science at the Pennsylvania State University and NIH Grants U41 HG006620 and R01 AI134384-01 as well as NSF ABI Grant 1661497.

bioRxiv preprint doi: https://doi.org/10.1101/429175; this version posted September 27, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Reference

1. Tenaillon, O. *et al.* Tempo and mode of genome evolution in a 50,000-generation experiment. *Nature Publishing Group* **536,** (2016).
2. Good, B. H., McDonald, M. J., Barrick, J. E., Lenski, R. E. & Desai, M. M. The dynamics of molecular evolution over 60,000 generations. *Nature* **551,** 45–50 (2017).
3. Tenaillon, O. *et al.* The Molecular Diversity of Adaptive Convergence. *Science* **335,** 457–461 (2012).
4. Barrick, J. E. & Lenski, R. E. Genome-wide Mutational Diversity in an Evolving Population of Escherichia coli. *Cold Spring Harb. Symp. Quant. Biol.* **74,** 119–129 (2009).
5. Bull, J. J. *et al.* Exceptional convergent evolution in a virus. *Genetics* **147,** 1497–1507 (1997).
6. Wichman, H. A., Badgett, M. R., Scott, L. A., Boulianne, C. M. & Bull, J. J. Different trajectories of parallel evolution during viral adaptation. *Science* **285,** 422–424 (1999).
7. Dickins, B. & Nekrutenko, A. High-resolution mapping of evolutionary trajectories in a phage. *Genome Biol. Evol.* **1,** 294–307 (2009).
8. Kao, K. C. & Sherlock, G. Molecular characterization of clonal interference during adaptive evolution in asexual populations of Saccharomyces cerevisiae. *Nat. Genet.* **40,** 1499–1504 (2008).
9. Levy, S. F. *et al.* Quantitative evolutionary dynamics using high-resolution lineage tracking. *Nature* **519,** 181–186 (2015).
10. Blundell, J. R. *et al.* The dynamics of adaptive genetic diversity during the early stages of clonal evolution. (2017). doi:10.1101/170589
11. Long, A., Liti, G., Luptak, A. & Tenaillon, O. Elucidating the molecular architecture of adaptation via evolve and resequence experiments. *Nat. Rev. Genet.* (2015). doi:10.1038/nrg3937
12. Bruger, E. L. & Marx, C. J. A decade of genome sequencing has revolutionized studies of experimental evolution. *Curr. Opin. Microbiol.* **45,** 149–155 (2018).
13. Cvijović, I., Nguyen Ba, A. N. & Desai, M. M. Experimental Studies of Evolutionary Dynamics in

bioRxiv preprint doi: https://doi.org/10.1101/429175; this version posted September 27, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Microbes. *Trends Genet.* (2018). doi:10.1016/j.tig.2018.06.004

14. Atwood, K. C., Schneider, L. K. & Ryan, F. J. Periodic Selection in Escherichia Coli. *Proceedings of the National Academy of Sciences* **37,** 146–155 (1951).
15. Gerrish, P. J. & Lenski, R. E. The fate of competing beneficial mutations in an asexual population. in *Contemporary Issues in Genetics and Evolution* 127–144 (1998).
16. Miller, C. R., Joyce, P. & Wichman, H. A. Mutational effects and population dynamics during viral adaptation challenge current models. *Genetics* **187,** 185–202 (2011).
17. Desai, M. M. & Fisher, D. S. Beneficial mutation selection balance and the effect of linkage on positive selection. *Genetics* **176,** 1759–1798 (2007).
18. Strelkowa, N. & Lässig, M. Clonal interference in the evolution of influenza. *Genetics* **192,** 671–682 (2012).
19. Lässig, M., Mustonen, V. & Walczak, A. M. Predicting evolution. *Nat Ecol Evol* **1,** 77 (2017).
20. Rebolledo-Jaramillo, B. *et al.* Maternal age effect and severe germ-line bottleneck in the inheritance of human mitochondrial DNA. *Proc. Natl. Acad. Sci. U. S. A.* **111,** 15474–15479 (2014).
21. Salk, J. J., Schmitt, M. W. & Loeb, L. A. Enhancing the accuracy of next-generation sequencing for detecting rare and subclonal mutations. *Nat. Rev. Genet.* **19,** 269–285 (2018).
22. Venkataram, S. *et al.* Development of a Comprehensive Genotype-to-Fitness Map of Adaptation-Driving Mutations in Yeast. *Cell* **166,** 1585–1596.e22 (2016).
23. Schmitt, M. W. *et al.* Detection of ultra-rare mutations by next-generation sequencing. *Proc. Natl. Acad. Sci. U. S. A.* **109,** 14508–14513 (2012).
24. Lenski, R. E., Simpson, S. C. & Nguyen, T. T. Genetic analysis of a plasmid-encoded, host genotype-specific enhancement of bacterial fitness. *J. Bacteriol.* **176,** 3140–3147 (1994).
25. Lenski, R. E. & Bouma, J. E. Effects of segregation and selection on instability of plasmid pACYC184 in Escherichia coli B. *J. Bacteriol.* **169,** 5314–5316 (1987).

bioRxiv preprint doi: https://doi.org/10.1101/429175; this version posted September 27, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

26. Modi, R. I. & Adams, J. Coevolution in bacterial-plasmid populations. *Evolution* **45,** 656–667 (1991).

27. McDermott, P. J., Gowland, P. & Gowland, P. C. Adaptation of Escherichia coli growth rates to the presence of pBR322. *Lett. Appl. Microbiol.* **17,** 139–143 (1993).

28. Hellweger, F. L. Escherichia coli adapts to tetracycline resistance plasmid (pBR322) by mutating endogenous potassium transport: in silico hypothesis testing. *FEMS Microbiol. Ecol.* **83,** 622–631 (2013).

29. Jones, I. M., Primrose, S. B., Robinson, A. & Ellwood, D. C. Maintenance of some ColE1-type plasmids in chemostat culture. *Mol. Gen. Genet.* **180,** 579–584 (1980).

30. Bull, J. J., Millstein, J., Orcutt, J. & Wichman, H. A. Evolutionary feedback mediated through population density, illustrated with viruses in chemostats. *Am. Nat.* **167,** E39–51 (2006).

31. Gresham, D. & Hong, J. The functional basis of adaptive evolution in chemostats. *FEMS Microbiol. Rev.* (2014).

32. Twigg, A. J. & Sherratt, D. Trans-complementable copy-number mutants of plasmid ColE1. *Nature* **283,** 216–218 (1980).

33. Plotka, M., Wozniak, M. & Kaczorowski, T. Quantification of Plasmid Copy Number with Single Colour Droplet Digital PCR. *PLoS One* **12,** e0169846 (2017).

34. Turner, P. E. & Chao, L. Sex and the evolution of intrahost competition in RNA virus phi6. *Genetics* **150,** 523–532 (1998).

35. Wielgoss, S. *et al.* Mutation Rate Inferred From Synonymous Substitutions in a Long-Term Evolution Experiment With Escherichia coli. *G3* **1,** 183–186 (2011).

36. Lee, H., Popodi, E., Tang, H. & Foster, P. L. Rate and molecular spectrum of spontaneous mutations in the bacterium Escherichia coli as determined by whole-genome sequencing. *Proc. Natl. Acad. Sci. U. S. A.* **109,** E2774–83 (2012).

37. Drake, J. W. A constant rate of spontaneous mutation in DNA-based microbes. *Proc. Natl. Acad. Sci.*

bioRxiv preprint doi: https://doi.org/10.1101/429175; this version posted September 27, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

U. S. A. **88,** 7160–7164 (1991).

38. Jee, J. *et al.* Rates and mechanisms of bacterial mutagenesis from maximum-depth sequencing. *Nature* **534,** 693–696 (2016).
39. Eguchi, Y., Itoh, T. & Tomizawa, J. Antisense RNA. *Annu. Rev. Biochem.* **60,** 631–652 (1991).
40. Davison, J. Mechanism of control of DNA replication and incompatibility in ColE1-type plasmids--a review. *Gene* **28,** 1–15 (1984).
41. Castagnoli, L., Lacatena, R. M. & Cesareni, G. Analysis of dominant copy number mutants of the plasmid pMB1. *Nucleic Acids Res.* **13,** 5353–5367 (1985).
42. Klein, W., Horlacher, R. & Boos, W. Molecular analysis of treB encoding the Escherichia coli enzyme II specific for trehalose. *J. Bacteriol.* **177,** 4043–4052 (1995).
43. Boos, W. *et al.* Trehalose transport and metabolism in Escherichia coli. *J. Bacteriol.* **172,** 3450–3461 (1990).
44. Rimmele, M. & Boos, W. Trehalose-6-phosphate hydrolase of Escherichia coli. *J. Bacteriol.* **176,** 5654–5664 (1994).
45. Park, M., Mitchell, W. J. & Rafii, F. Effect of trehalose and trehalose transport on the tolerance of clostridium perfringens to environmental stress in a wild type strain and its fluoroquinolone-resistant mutant.
46. Blaby, I. K. *et al.* Experimental evolution of a facultative thermophile from a mesophilic ancestor. *Appl. Environ. Microbiol.* **78,** 144–155 (2012).
47. Behringer, M. G. *et al.* Escherichia colicultures maintain stable subpopulation structure during long-term evolution. *Proceedings of the National Academy of Sciences* **115,** E4642–E4650 (2018).
48. Kawashima, H., Horii, T., Ogawa, T. & Ogawa, H. Functional domains of Escherichia coli recA protein deduced from the mutational sites in the gene. *Mol. Gen. Genet.* **193,** 288–292 (1984).
49. Bryant, F. R. Construction of a recombinase-deficient mutant recA protein that retains

bioRxiv preprint doi: https://doi.org/10.1101/429175; this version posted September 27, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

single-stranded DNA-dependent ATPase activity. *J. Biol. Chem.* **263,** 8716–8723 (1988).

50. Sancar, A., Stachelek, C., Konigsberg, W. & Rupp, W. D. Sequences of the recA gene and protein. *Proc. Natl. Acad. Sci. U. S. A.* **77,** 2611–2615 (1980).
51. McDonald, M. J., Rice, D. P. & Desai, M. M. Sex speeds adaptation by altering the dynamics of molecular evolution. *Nature* **531,** 233–236 (2016).
52. De Gelder, L. *et al.* Combining mathematical models and statistical methods to understand and predict the dynamics of antibiotic-sensitive mutants in a population of resistant bacteria during experimental evolution. *Genetics* **168,** 1131–1144 (2004).
53. San Millan, A. *et al.* Positive selection and compensatory adaptation interact to stabilize non-transmissible plasmids. *Nat. Commun.* **5,** 5208 (2014).
54. San Millan, A. *et al.* Small-plasmid-mediated antibiotic resistance is enhanced by increases in plasmid copy number and bacterial fitness. *Antimicrob. Agents Chemother.* **59,** 3335–3341 (2015).
55. Loftie-Eaton, W. *et al.* Compensatory mutations improve general permissiveness to antibiotic resistance plasmids. *Nat Ecol Evol* **1,** 1354–1363 (2017).
56. Harrison, E., Guymer, D., Spiers, A. J., Paterson, S. & Brockhurst, M. A. Parallel compensatory evolution stabilizes plasmids across the parasitism-mutualism continuum. *Curr. Biol.* **25,** 2034–2039 (2015).
57. Paulsson, J. & Ehrenberg, M. Trade-off between segregational stability and metabolic burden: a mathematical model of plasmid ColE1 replication control. *J. Mol. Biol.* **279,** 73–88 (1998).
58. Santos-Lopez, A. *et al.* A Naturally Occurring Single Nucleotide Polymorphism in a Multicopy Plasmid Produces a Reversible Increase in Antibiotic Resistance. *Antimicrob. Agents Chemother.* **61,** (2017).
59. Rodriguez-Beltran, J. *et al.* Multicopy plasmids allow bacteria to escape from fitness trade-offs during evolutionary innovation. *Nature Ecology & Evolution* **2,** 873–881 (2018).

bioRxiv preprint doi: https://doi.org/10.1101/429175; this version posted September 27, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

60. Eguchi, Y. & Tomizawa, J. Complex formed by complementary RNA stem-loops and its stabilization by a protein: function of CoIE1 Rom protein. *Cell* **60,** 199–209 (1990).

61. Takahashi, C. N., Miller, A. W., Ekness, F., Dunham, M. J. & Klavins, E. A low cost, customizable turbidostat for use in synthetic circuit characterization. *ACS Synth. Biol.* **4,** 32–38 (2015).

62. Stoler, N., Arbeithuber, B., Guiblet, W., Makova, K. D. & Nekrutenko, A. Streamlined analysis of duplex sequencing data with Du Novo. *Genome Biol.* **17,** 180 (2016).

63. Goecks, J., Nekrutenko, A., Taylor, J. & Galaxy Team. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.* **11,** R86 (2010).

64. Lariviere, D., Mei, H., Freeberg, M., Taylor, J. & Nekrutenko, A. Understanding trivial challenges of microbial genomics: An assembly example. (2018). doi:10.1101/347625

bioRxiv preprint doi: https://doi.org/10.1101/429175; this version posted September 27, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Figure Legends

Figure 1. The four replicates of the short term (R1, R2) and long term (R6, R7) turbidostat runs. OD_{600} values were constantly monitored and maintained at 0.8. Samples were taken every 12 h. Dips in OD values represent sampling points. At each sampling point, ⅔ of the turbidostat volume was extracted for duplex sequencing.

Figure 2. Locations and frequencies of variants detected at bases 3,027–3,035 and 3,118 on pBR322 with duplex sequencing. R1 and R2 denote the two replicates on short turbidostat run. R6 and R7 denote the two replicates of long turbidostat run. These variants contributed to change in plasmid copy number. Positions are projected to the y-axis, and colored individually. Sizes of the closed circles are proportional to the allele frequencies of corresponding variants.

Figure 3. A combination of empirical (R1, R2, R6, and R7) and predicted (simulation) allele frequencies. The simulation was performed with parameters obtained by numerical optimization (see *Material and Methods*). Different turbidostat runs are colored respectively. Y-axis shows the sum of frequencies at bases 3,027–3,035 and 3,118. X-axis represents generations. The x-axis here using generation as the unit corresponds to the x-axis using time in hour in Fig. 1 and Fig. 2. This conversion was accomplished by assuming that the generation time was constantly 60 min.

Supplementary Figure 1. Mutational spectrum across 145 SNVs relative to the reference sequence of pBR322 in our experiment. X-axis shows different mutation types. Y-axis shows the number of SNVs falling into respective mutation types.

Supplementary Figure 2. Secondary structures of RNAI and RNAII molecules. **A**. Positions 3,027 and 3,035 are boundaries of the region that contains the majority of beneficial mutations identified in this experiment. RNAII was reported 555 nt in length. Only RNAII loop regions interacting with RNAI are shown, and denoted as "RNAII_555". **B**. Mutations at base 3,118 (uppercase, black) increase transcription of RNAII.

Supplementary Figure 3. Growth rate of individual isolates from replicate R6 of the long turbidostat run. In the left panel, growth rate were estimated from slopes in Fig. 1. Growth rate the initial isolate is shown in a red triangle, and those of the remaining isolated are shown in black circles. Growth rate of the terminal isolate in replicate R6 cannot be obtained from Fig. 1, because the terminal isolate was stored in the glycerol stock and did not show a dip in OD values. However, it can be re-measured by

bioRxiv preprint doi: https://doi.org/10.1101/429175; this version posted September 27, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

inoculating the glycerol stock into turbidostat; this is shown in the right panel. In the right panel, initial and terminal isolated, both in red triangles, were re-measured in triplicates.

Supplementary Figure 4. Same as Fig. S3, growth rate of individual isolates from replicate R7 of the long turbidostat run.

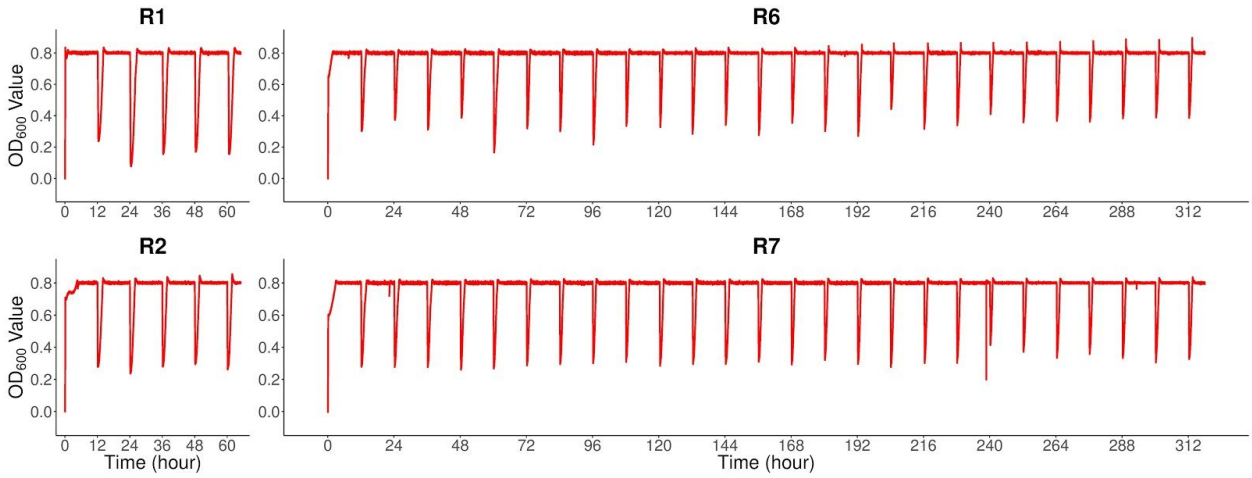
Supplementary Figure 5. Turbidostat setup.

Supplementary Figure 6. Prediction of frequencies of lineages carrying plasmid mutations at bases 3,027–3,035 and 3,118. The a, b, d, μ parameters were given the same value as those obtained by numerical optimization and used in Fig. 3. The c parameter was multiplied by 0.1, 0.01, and 0.001. Different c values generated different dynamics, which were shown in different colors.

Supplementary Figure 7. Same as Fig. S6. The a, b, c, d parameters were given the same value as those obtained by numerical optimization. The μ parameter was multiplied by 10, 50,and 100, generating different dynamics.

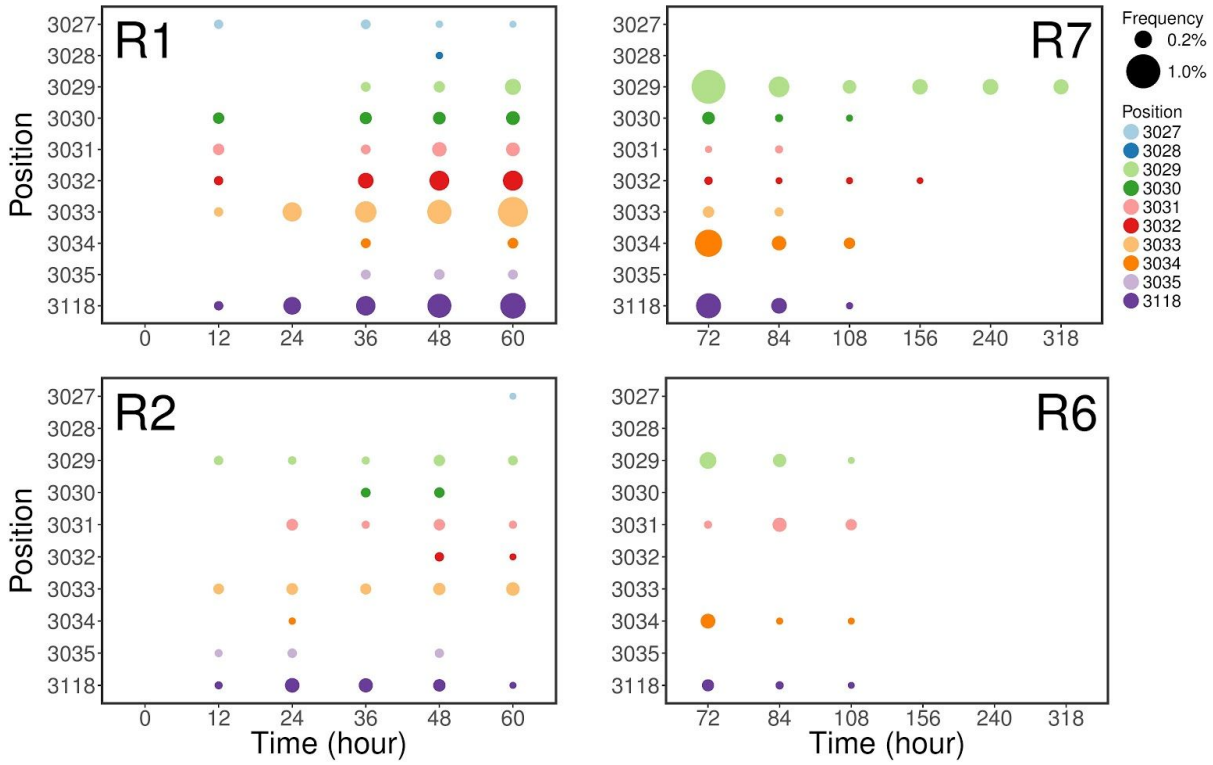
bioRxiv preprint doi: https://doi.org/10.1101/429175; this version posted September 27, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Figure 1. The four replicates of the short term (R1, R2) and long term (R6, R7) turbidostat runs. OD_{600} values were constantly monitored and maintained at 0.8. Samples were taken every 12 h. Dips in OD values represent sampling points. At each sampling point, ⅔ of the turbidostat volume was extracted for duplex sequencing.



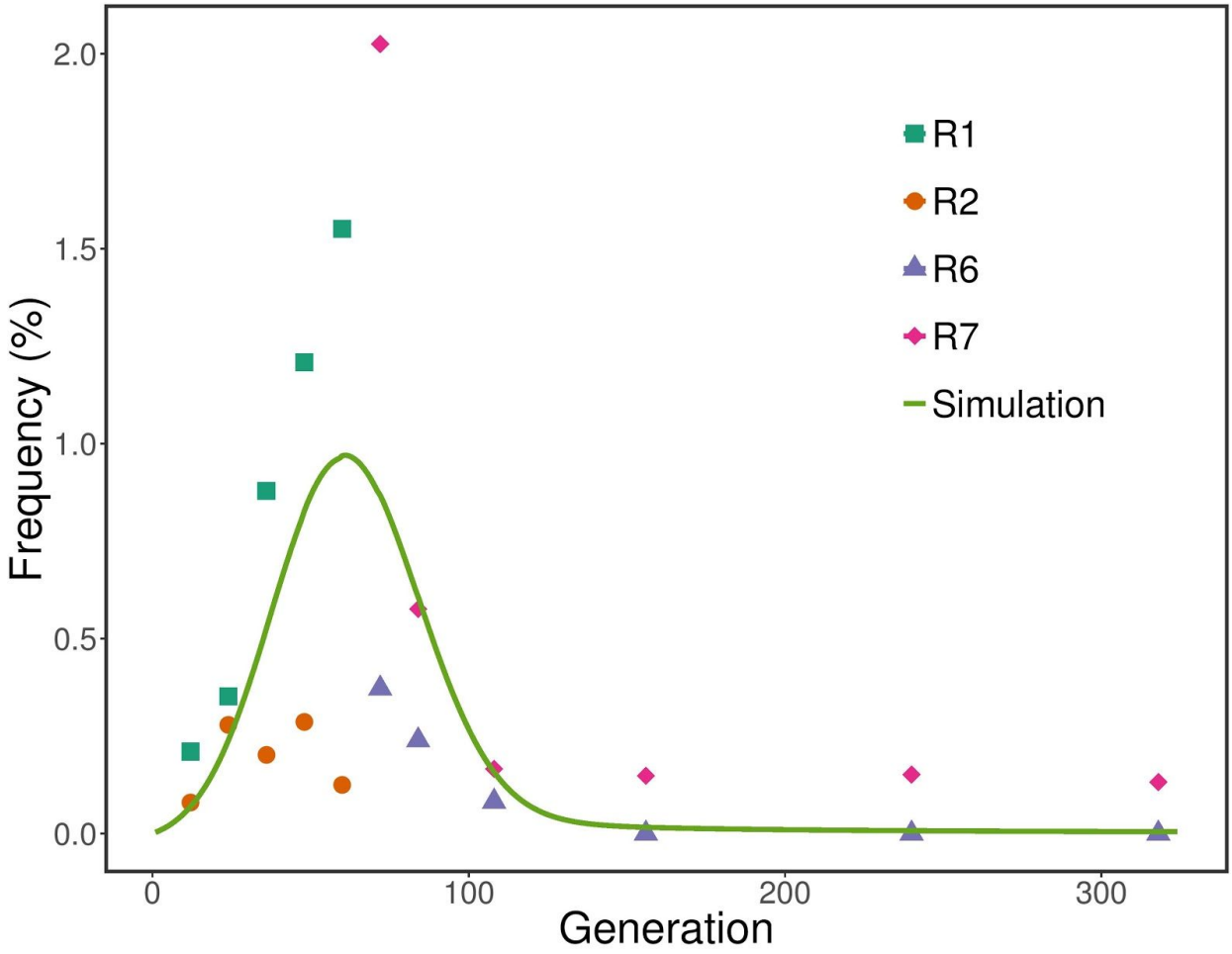
bioRxiv preprint doi: https://doi.org/10.1101/429175; this version posted September 27, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Figure 2. Locations and frequencies of variants detected at bases 3,027–3,035 and 3,118 on pBR322 with duplex sequencing. R1 and R2 denote the two replicates on short turbidostat run. R6 and R7 denote the two replicates of long turbidostat run. These variants contributed to change in plasmid copy number. Positions are projected to the y-axis, and colored individually. Sizes of the closed circles are proportional to the allele frequencies of corresponding variants.



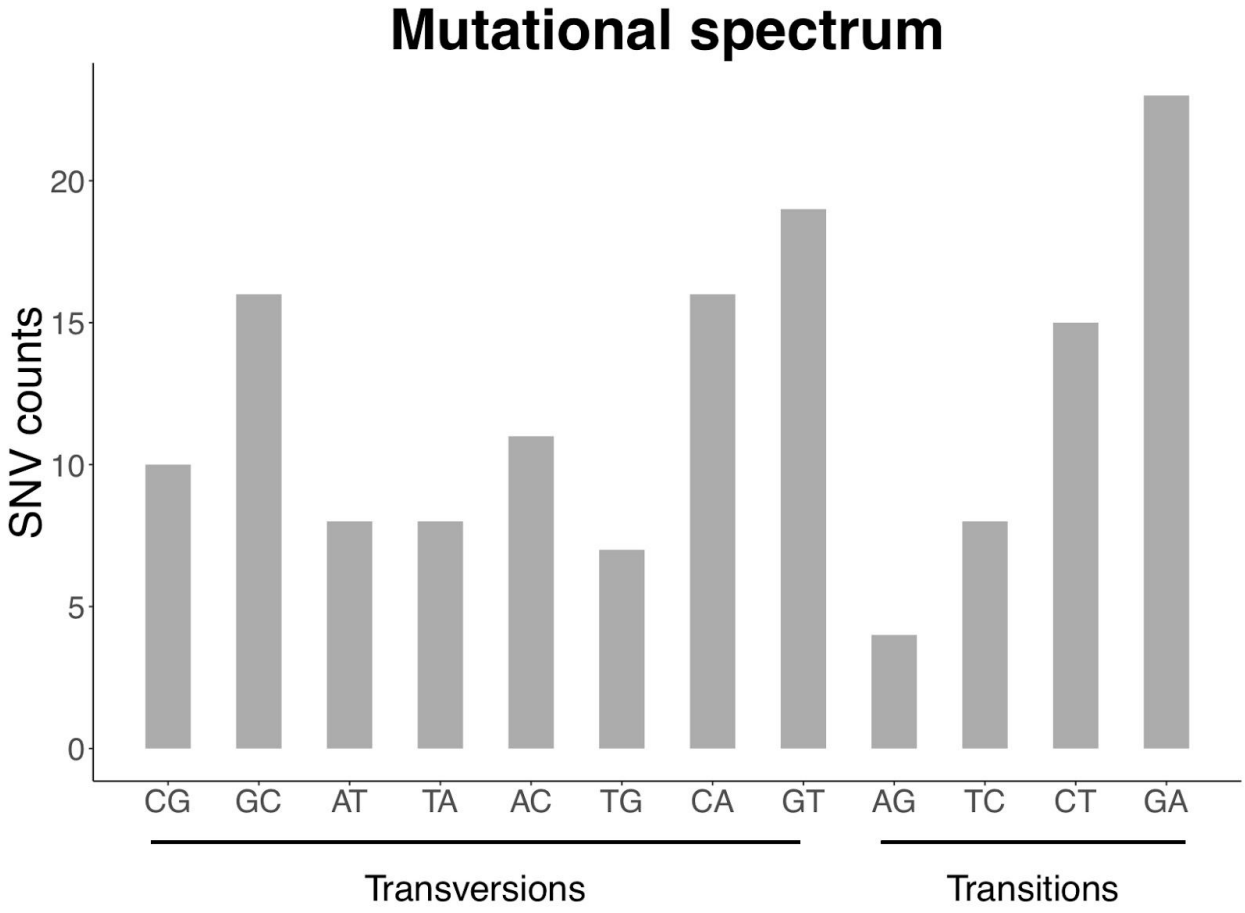
bioRxiv preprint doi: https://doi.org/10.1101/429175; this version posted September 27, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Figure 3. A combination of empirical (R1, R2, R6, and R7) and predicted (simulation) allele frequencies. The simulation was performed with parameters obtained by numerical optimization (see *Material and Methods*). Different turbidostat runs are colored respectively. Y-axis shows the sum of frequencies at bases 3,027–3,035 and 3,118. X-axis represents generations. The x-axis here using generation as the unit corresponds to the x-axis using time in hour in Fig. 1 and Fig. 2. This conversion was accomplished by assuming that the generation time was constantly 60 min.



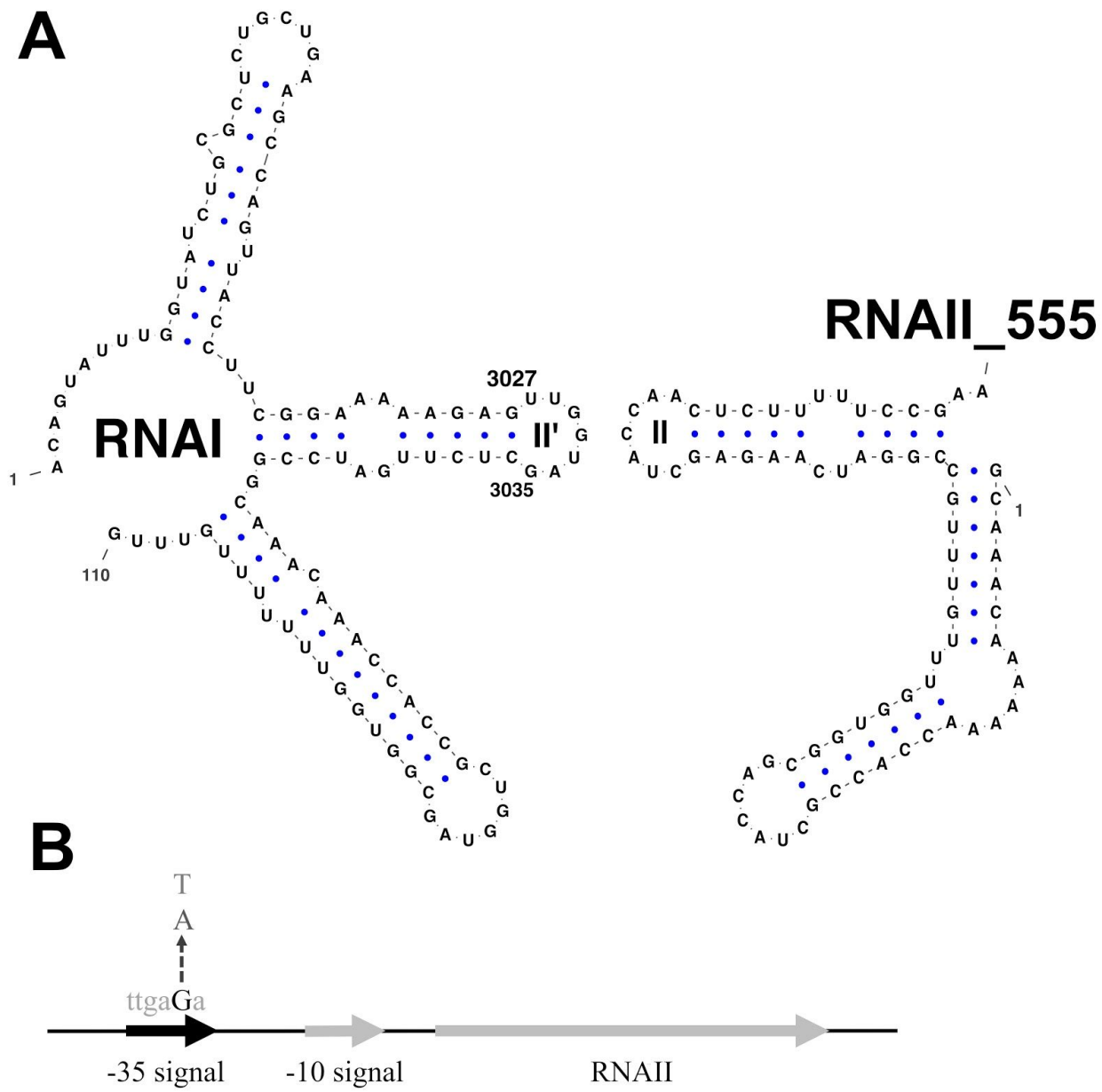
bioRxiv preprint doi: https://doi.org/10.1101/429175; this version posted September 27, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Supplementary Figure 1. Mutational spectrum across 145 SNVs relative to the reference sequence of pBR322 in our experiment. X-axis shows different mutation types. Y-axis shows the number of SNVs falling into respective mutation types.



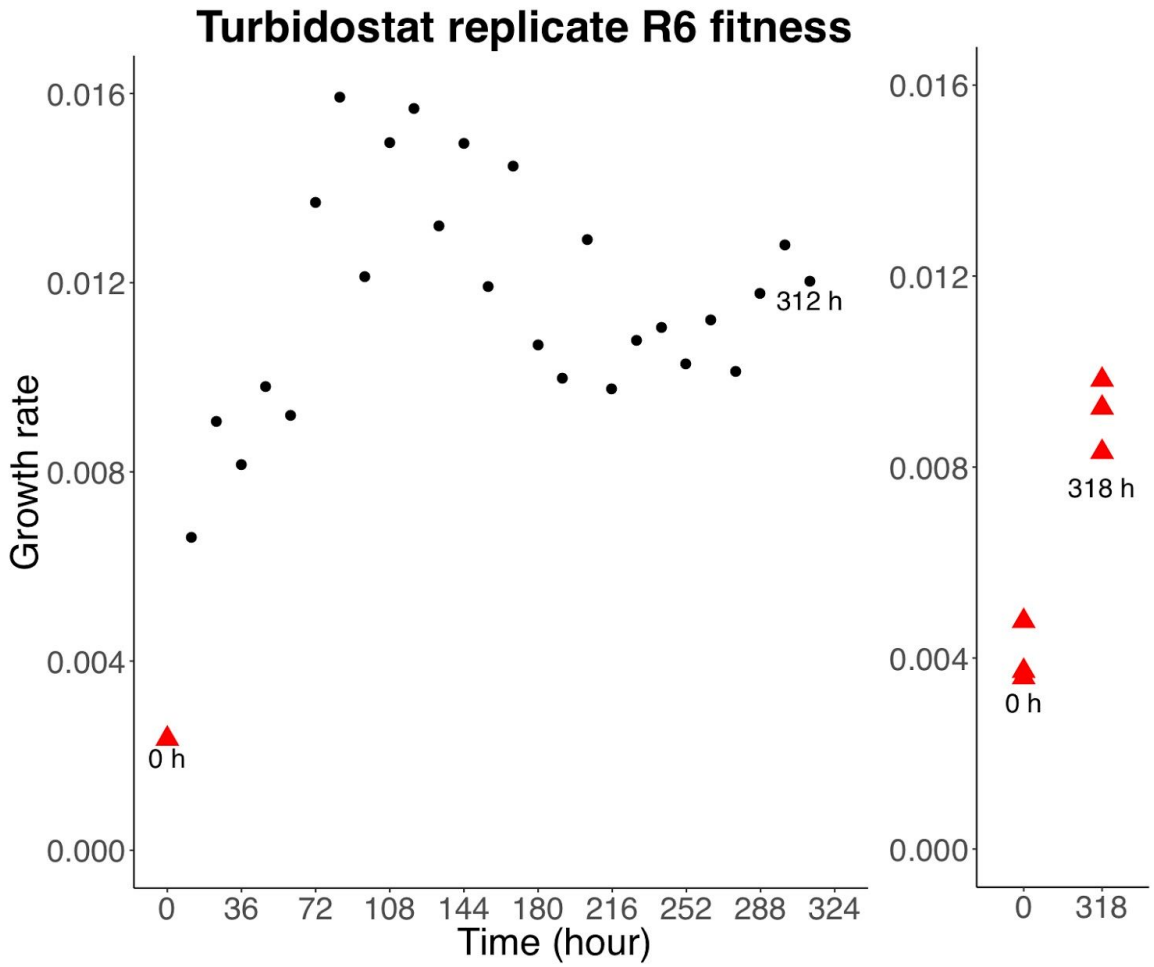
bioRxiv preprint doi: https://doi.org/10.1101/429175; this version posted September 27, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Supplementary Figure 2. Secondary structures of RNAI and RNAII molecules. **A**. Positions 3027 and 3035 are boundaries of the region that contains the majority of beneficial mutations identified in this experiment. RNAII was reported 555 nt in length. Only RNAII loop regions interacting with RNAI are shown, and denoted as "RNAII_555". **B**. Mutations at base 3,118 (uppercase, black) increase transcription of RNAII.



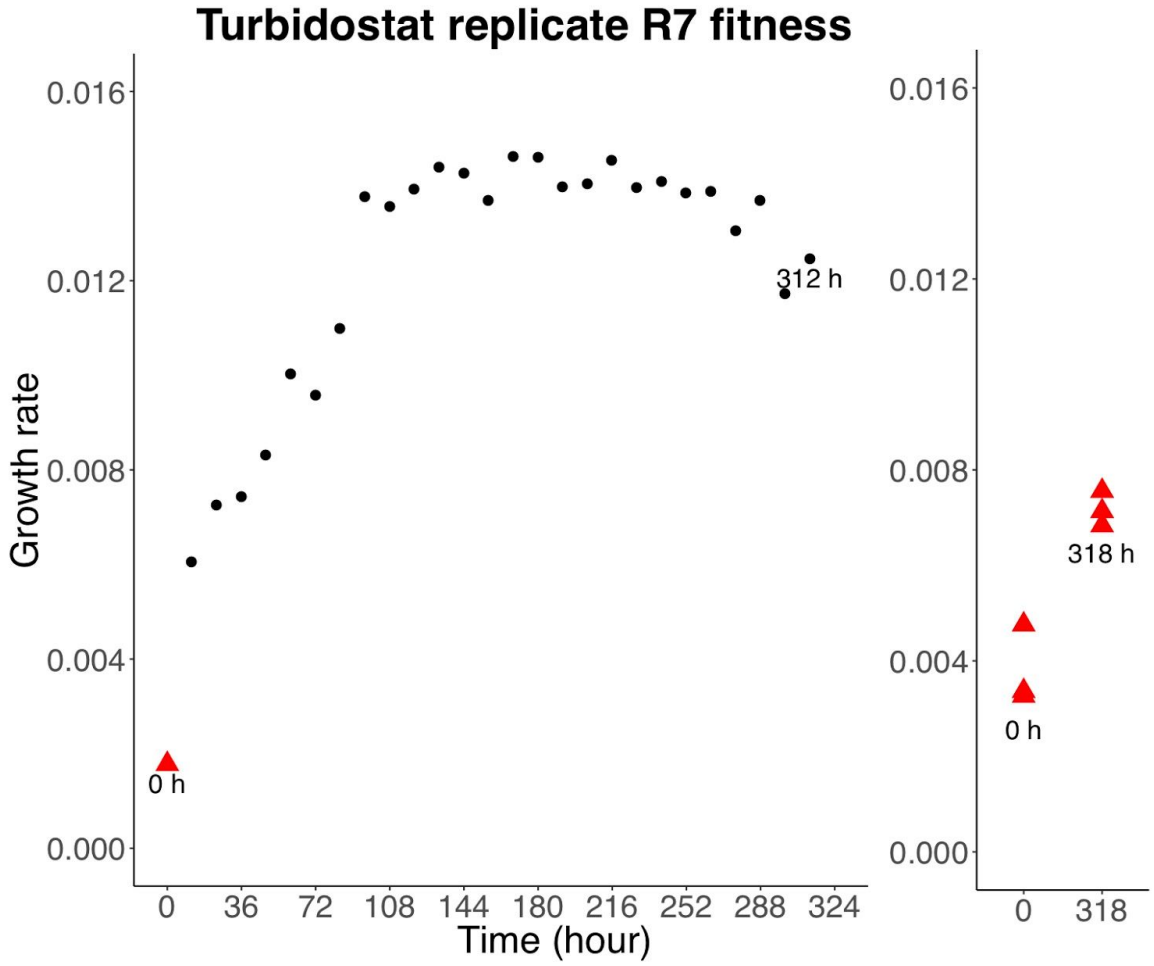
bioRxiv preprint doi: https://doi.org/10.1101/429175; this version posted September 27, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Supplementary Figure 3. Growth rate of individual isolates from replicate R6 of the long turbidostat run. In the left panel, growth rate were estimated from slopes in Fig. 1. Growth rate the initial isolate is shown in a red triangle, and those of the remaining isolated are shown in black circles. Growth rate of the terminal isolate in replicate R6 cannot be obtained from Fig. 1, because the terminal isolate was stored in the glycerol stock and did not show a dip in OD values. However, it can be re-measured by inoculating the glycerol stock into turbidostat; this is shown in the right panel. In the right panel, initial and terminal isolated, both in red triangles, were re-measured in triplicates.



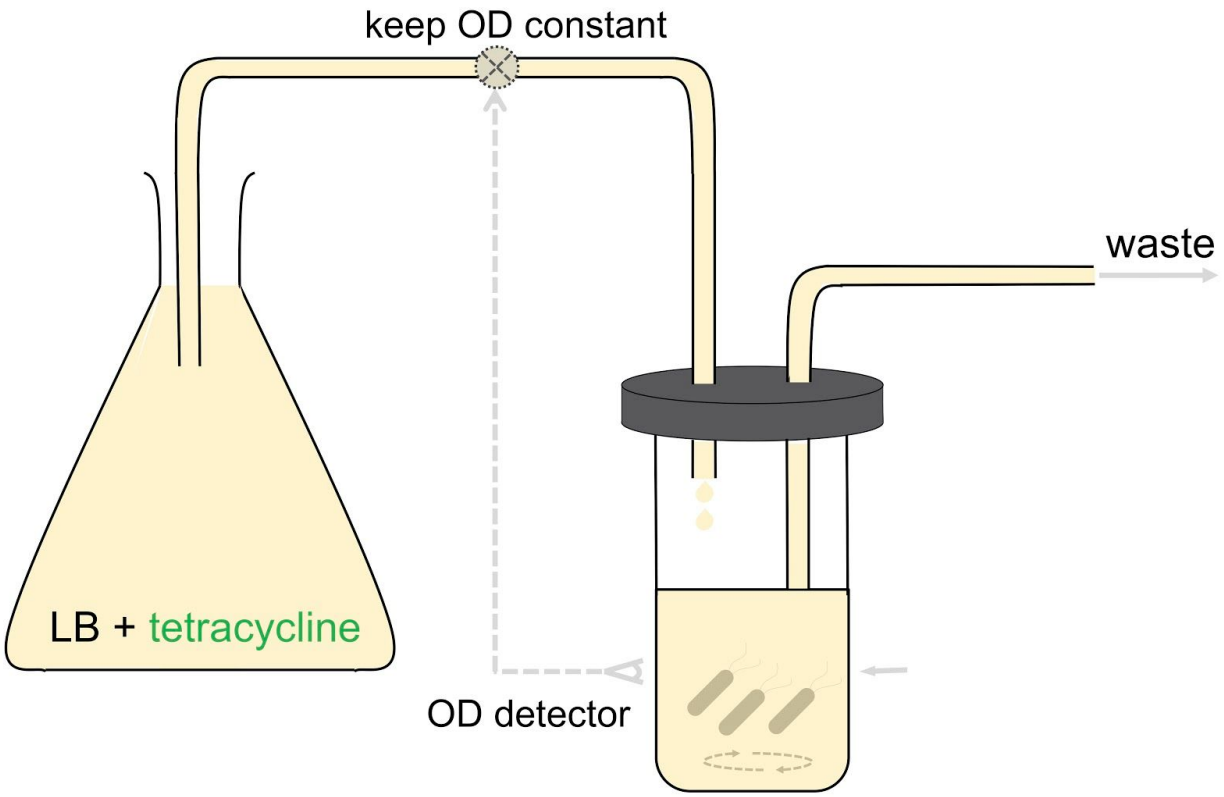
bioRxiv preprint doi: https://doi.org/10.1101/429175; this version posted September 27, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Supplementary Figure 4. Same as Fig. S3, growth rate of individual isolates from replicate R7 of the long turbidostat run.



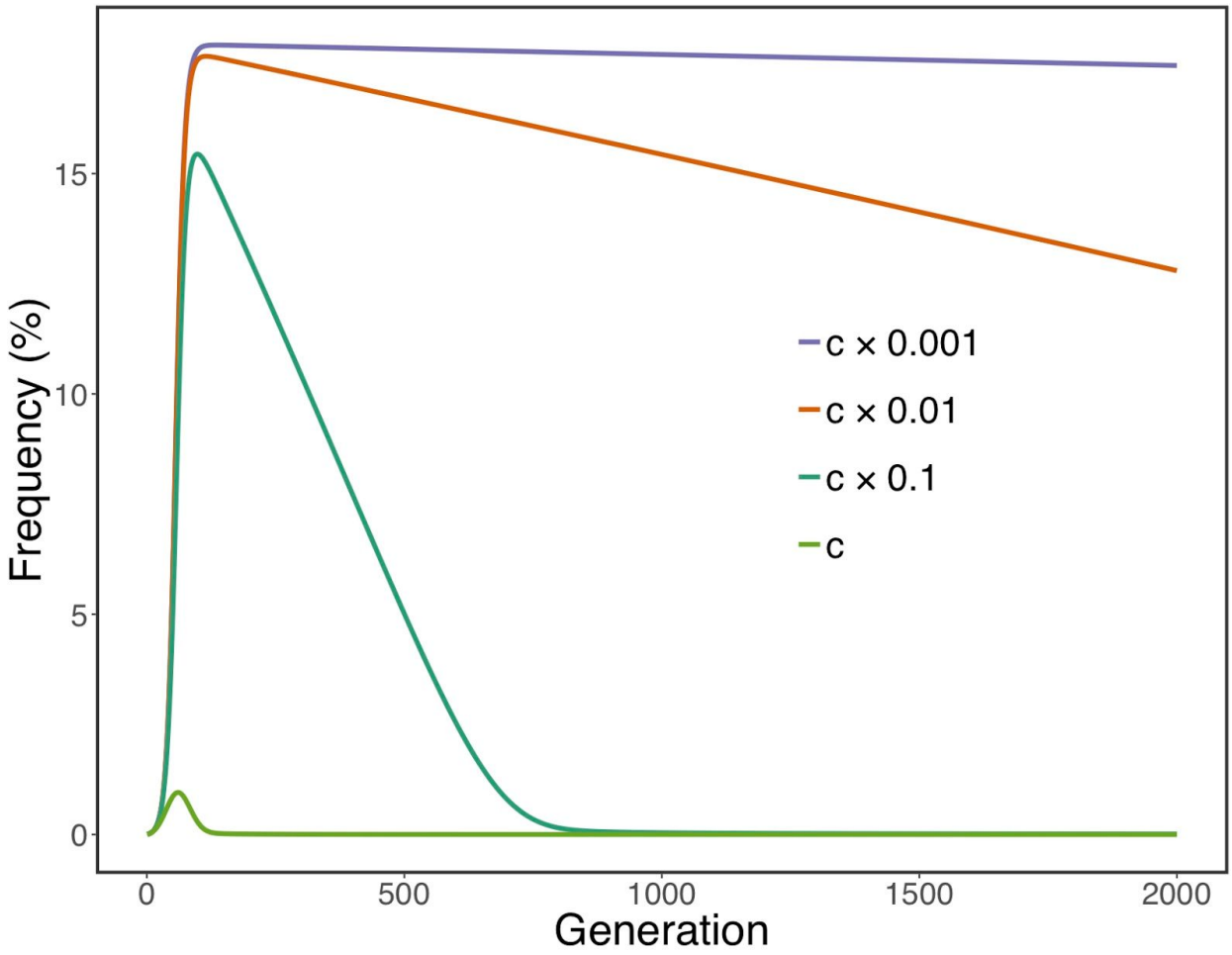
bioRxiv preprint doi: https://doi.org/10.1101/429175; this version posted September 27, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Supplementary Figure 5. Turbidostat setup.



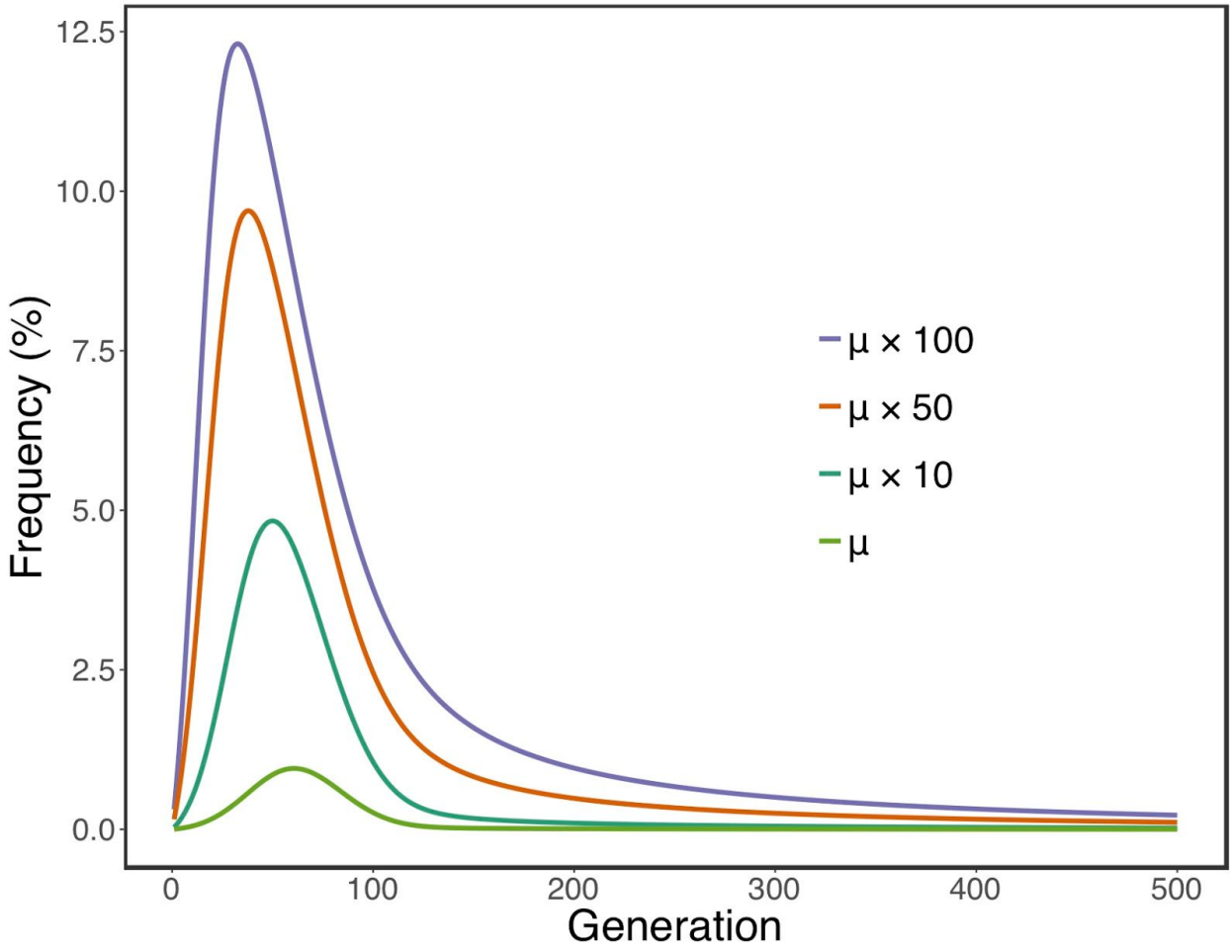
bioRxiv preprint doi: https://doi.org/10.1101/429175; this version posted September 27, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Supplementary Figure 6. Prediction of frequencies of lineages carrying plasmid mutations at bases 3,027–3,035 and 3,118. The *a*, *b*, *d*, μ parameters were given the same value as those obtained by numerical optimization and used in Fig. 3. The *c* parameter was multiplied by 0.1, 0.01, and 0.001. Different *c* values generated different dynamics, which were shown in different colors.



bioRxiv preprint doi: https://doi.org/10.1101/429175; this version posted September 27, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Supplementary Figure 7. Same as Fig. S6. The *a*, *b*, *c*, *d* parameters were given the same value as those obtained by numerical optimization. The μ parameter was multiplied by 10, 50,and 100, generating different dynamics.



bioRxiv preprint doi: https://doi.org/10.1101/429175; this version posted September 27, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Table 1. Optimized estimator for the model.

Parameter	a	b	c	d	μ
Estimator	1.36×10^{-12}	1.68×10^{-1}	2.69×10^{-3}	3.82×10^{-4}	3.13×10^{-5}

Table S1. Plasmid SNVs and INDELs identified in all four turbidostat runs.

Table S2. Chromosomal mutations identified in the two long term replicates—R6 and R7.

Table S3. Growth rate and generation time over the course of the two long term turbidostat replicates.

Table S4. Bounds, in $\log_{10}$ scale, for the parameters to be drawn uniformly at random to model the dynamics of plasmid.