ARTICLE

doi:10.1038/nature14279

Quantitative evolutionary dynamics using high-resolution lineage tracking

Sasha F. Levy[1,2,3]*, Jamie R. Blundell[4,5]*, Sandeep Venkataram[5], Dmitri A. Petrov[5], Daniel S. Fisher[4,5] & Gavin Sherlock[1]

Evolution of large asexual cell populations underlies ~30% of deaths worldwide, including those caused by bacteria, fungi, parasites, and cancer. However, the dynamics underlying these evolutionary processes remain poorly understood because they involve many competing beneficial lineages, most of which never rise above extremely low frequencies in the population. To observe these normally hidden evolutionary dynamics, we constructed a sequencing-based ultra high-resolution lineage tracking system in *Saccharomyces cerevisiae* that allowed us to monitor the relative frequencies of ~500,000 lineages simultaneously. In contrast to some expectations, we found that the spectrum of fitness effects of beneficial mutations is neither exponential nor monotonic. Early adaptation is a predictable consequence of this spectrum and is strikingly reproducible, but the initial small-effect mutations are soon outcompeted by rarer large-effect mutations that result in variability between replicates. These results suggest that early evolutionary dynamics may be deterministic for a period of time before stochastic effects become important.

A major focus of biomedical research has been to identify mutations responsible for increased pathogenicity, cancer progression, or drug resistance in large evolving asexual cell populations[1–12]. Yet, even characterizing all mutations underlying a disease is not sufficient to understand its progression. Rather, a quantitative understanding of the evolutionary dynamics is necessary to determine which adaptive mutations contribute significantly to driving the population fitness higher, and which are serendipitous or inconsequential. Mutations identified through genome sequencing are likely to constitute only the 'tip of the iceberg', with many beneficial mutations that impact the evolutionary dynamics never rising above extremely low frequencies[13,14].

A lineage trajectory, the size of a small subpopulation of cells over time, can be used to discover a beneficial mutation present at an extremely low frequency, and to measure its time of occurrence and selective advantage (Fig. 1a)[1,15–18]. A lineage increasing in size faster than can be explained by stochastic drift indicates that a beneficial mutation has occurred and risen to a high enough frequency to grow almost deterministically (that is, it has 'established'). Most beneficial mutations will drift to extinction before establishing (Supplementary Information section 4.1 and 4.4). For those that do establish, the exponential rate at which a lineage grows is a measure of the fitness effect (s) of the mutation. Extrapolating back the exponential growth, the establishment time (τ) can be inferred: this is a rough estimate of when the mutation occurred[19] (Supplementary Information section 4.1 and 4.2). A systematic characterization of the distributions of s and τ for beneficial mutations has been lacking, although these are fundamental to the evolutionary dynamics of large populations[20].

The major experimental challenge is developing a method to quantitatively measure the trajectories of large numbers of small lineages. Large lineages will accumulate multiple beneficial mutations contemporaneously, confounding measurements of s and τ (Fig. 1a, multiple mutations, Supplementary Information section 4.5). Small lineages are unlikely to acquire a beneficial mutation at all, so many trajectories must be observed to characterize the distributions of s and τ. DNA barcodes offer a powerful way to simultaneously track multiple lineages[21–23], yet technical barriers have limited the number of barcodes that can be inserted into cells[24]. Here we constructed a system capable of inserting ~500,000 random DNA barcodes into an initially clonal yeast population. Using this system in populations of $\sim 10^8$ cells growing in a defined glucose-limited minimal medium, we identified ~25,000 lineages that gained a beneficial mutation within ~168 generations, measured s and τ for each, and determined the spectrum of mutation rates to each fitness effect. This spectrum results in a deterministic increase in the mean population fitness early, with stochastic events governing its trajectory later.

Lineage tracking with random barcodes

We generated yeast lineages by inserting a random 20-nucleotide barcode at a single location in the genome (Fig. 1b, Supplementary Information section 1.3). To achieve a large number of integration events, we inserted a 'landing pad' into a neutral location in the yeast genome that allows for high-frequency, site-specific genomic integration of plasmids via the Cre–*loxP* recombination system[25,26]. A plasmid library containing ~500,000 random barcodes was inserted into the genome at the landing pad. Barcoding requires ~48 generations of growth from a common ancestor (Extended Data Fig. 1). Adaptive mutations begin to occur during this initial growth and can be carried forward into the evolution.

The same barcoded yeast library was evolved in replicate experiments (E1 and E2) for ~168 generations in serial batch culture, diluting 1:250 every ~8 generations, with a bottleneck population size of $\sim 7 \times 10^7$ (Extended Data Fig. 1, Supplementary Information section 4.4). To count the relative frequency of each lineage across time, we isolated genomic DNA from the pooled population, amplified lineage tags using a two-step PCR protocol, and sequenced amplicons (Fig. 1b, Supplementary Information section 1.5 and 5.2).

Plotting the relative frequency of each barcode over ~168 generations shows a reproducible pattern of population dynamics across replicates (Fig. 2a and Extended Data Fig. 2a). Most lineages declined in frequency (blue lines, neutral lineages), but a modest fraction (~5%,

[1]Department of Genetics, Stanford University, Stanford, California 94305-5120, USA. [2]Laufer Center for Physical and Quantitative Biology, Stony Brook University, Stony Brook, New York 11794–5252, USA. [3]Department of Biochemistry and Cellular Biology, Stony Brook University, Stony Brook, New York 11794–5215, USA. [4]Department of Applied Physics, Stanford University, Stanford, California 94305, USA. [5]Department of Biology, Stanford University, Stanford, California 94305, USA.
*These authors contributed equally to this work.

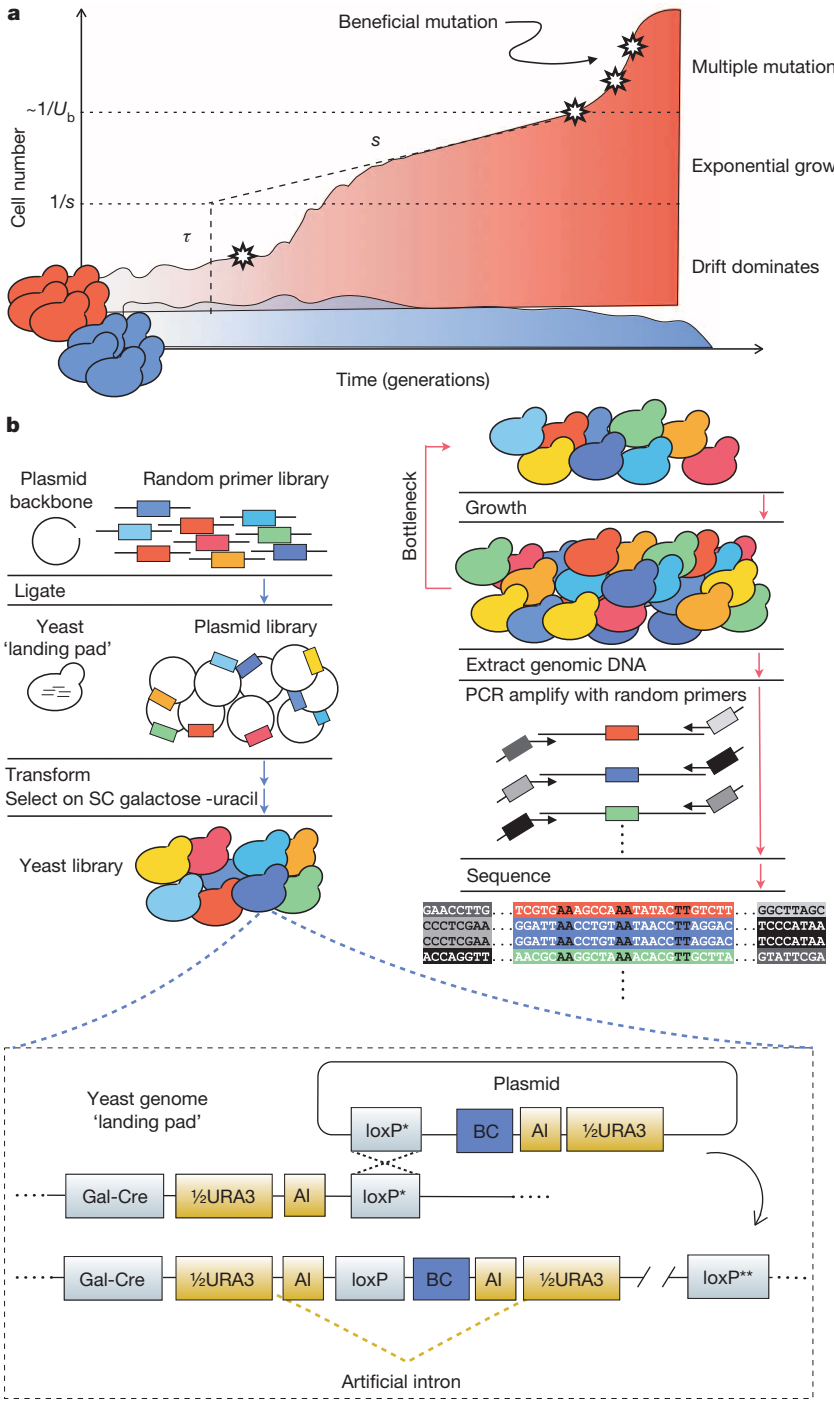


Figure 1 | Lineage tracking with random barcodes. **a**, Typical lineage trajectories. A small lineage that does not acquire a beneficial mutation (neutral, blue) will fluctuate in size due to drift before eventually being outcompeted. Rarely, a lineage will acquire a beneficial mutation (star) with a fitness effect of s (adaptive, red). In most cases, this beneficial mutation is lost to drift. If the beneficial mutants drift to a size $>\sim 1/s$ (lower dotted horizontal line), the lineage will begin to grow exponentially at a rate s. Extrapolating the exponential growth to the time at which the mutation is inferred to have reach a size $\sim 1/s$ yields the establishment time (τ, dashed vertical line) which roughly corresponds to the time when the mutation occurred with an uncertainty of $\sim 1/s$. At sizes $> \sim 1/U_b$ (upper dotted horizontal line), where U_b is the total beneficial mutation rate, the lineage will acquire additional beneficial mutations. **b**, Barcode insertion and sequencing. Left, sequences containing random 20 nucleotide barcodes (colours) are inserted first into a plasmid and then into a specific location in the genome. Bottom, recombination between two partially crippled *loxP* sites (*loxP**) integrates the plasmid into the genome and completes a *URA3* selectable marker, resulting in one functional and one crippled *loxP* site (*loxP***). The *URA3* marker is interrupted by an artificial intron containing the barcode. Right, to measure relative fitness, cells are passed through growth-bottleneck cycles of ~8 generations. Before each bottleneck, genomic DNA is extracted, lineage barcode tags are amplified using a two-step PCR protocol, and amplicons are sequenced. By inserting unique molecular identifiers[49] (also short random barcodes, grey bars) in early cycles of the PCR, PCR duplicates of the same template molecule (purple) are detected[49,50].

see below) had acquired a beneficial mutation that established (red lines, adaptive lineages). At later time points, the growth of adaptive lineages attenuates as the population mean fitness increases (clonal interference)[27].

To calculate the probability that a lineage contains an adaptive mutation, one must differentiate between a trajectory that increases due to an adaptive event from one that increases due to genetic drift and measurement errors. Because either scenario is rare, the right-hand tail of the distribution of read numbers is particularly important. Thus, we characterized the full distribution of noise that results from drift and sampling errors due to DNA isolation, amplification and sequencing, (black curve, Extended Data Fig. 2b, Supplementary Information section 5). The decline in frequency of neutral lineages is used to infer the increase in mean fitness of the population[4] (Fig. 2a and Extended Data Fig. 2a, b).

Using our estimates of noise and the mean fitness, we calculate the probability that a trajectory is explained by a mutation with fitness effect, s, having an establishment time, τ, over a broad range of s and τ (under a uniform prior in τ and an exponential prior in s, Extended Data 2c). If this exceeds the probability that no beneficial mutation occurs, we define the lineage as adaptive, with the peak of the probability our best estimate of s and τ (Supplementary Information section 7). Estimates of s and τ for each adaptive lineage are combined to calculate a second measurement of the increase in mean fitness (Fig. 2a and Extended Data Fig. 2a, insets). Our two methods to infer mean fitness agree, indicating that most lineages driving the mean fitness have been detected. Uncertainties in s and τ depend on the specific lineage trajectory; however, they are generally low ($\Delta s \pm 0.5\%$, $\Delta\tau \pm 10$ generations, Supplementary Information section 7.7).

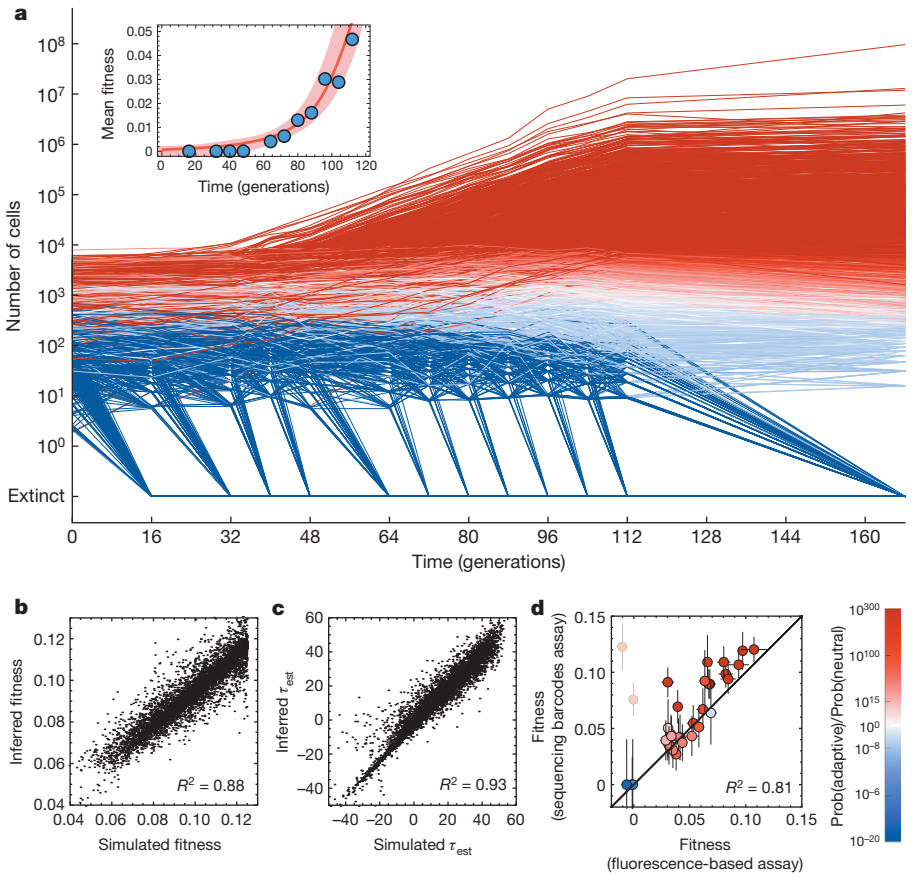


Figure 2 | Inferring the fitnesses and establishment times from lineage trajectories. **a**, Selected lineage trajectories from E1 coloured according to the probability that they contain an established beneficial mutation. The decline of adaptive lineages at later times is caused by the increase of the population mean fitness (inset). The population mean fitness is inferred from both the decline of neutral lineages (blue circles) and the growth of beneficial lineages (red line, Supplementary Information section 6.2). Shading indicates the error in mean fitness. **b**, **c**, The inferred fitnesses (**b**) and establishment times (**c**) from analysis of simulated trajectories correlate strongly with the known simulated values. **d**, Scatter plot of the fitness of 33 clones picked from E2 at generation 88 inferred by sequencing and pairwise competition (colouring as in (**a**), with outliers lightened in colour and excluded from correlation). Error bars represent one standard deviation.

To validate estimates of s and τ, we first analysed a simulated data set with comparable levels of noise to our experiment (Supplementary Information section 12). We find a strong correlation between the known and inferred values for both s ($R^2 = 0.88$ in Fig. 2b) and τ ($R^2 = 0.93$ in Fig. 2c). Second, we picked 33 clones from generation 88 that belong to different adaptive lineages and performed pairwise competitive fitness assays on each (Supplementary Information section 2). We find a strong correlation between these two methods ($R^2 = 0.81$, Fig. 2d). Outliers (lighter coloured data points) are likely due to a neutral cell being sampled from a lineage containing mostly adaptive cells. Other deviations could be due to interactions between adaptive lineages (that is, frequency dependent fitness) or to multiple mutations on the same genome (Supplementary Information section 8).

In total, ~25,000 beneficial mutations with a fitness effect of >2% established before generation 112 in E1 (Fig. 3a), a number that is roughly consistent with E2 (Extended Data Fig. 3a) and simulated data (Supplementary Figs 44 and 45 and Supplementary Information section 12). Adaptation occurs quickly: by generation 112 the population mean fitness is over 5% higher than the ancestor, with some lineages having a fitness advantage of >10%. E1 and E2 share 48 generations of common growth. During this time, ~6,000 lineages acquire a beneficial mutation that is sampled into, and establishes in, both replicates (Fig. 3a and Extended Data Fig. 3a, purple circles). We define these mutations as 'pre-existing': their presence is not an artefact of our experiment, but a general expectation for large populations grown from a single cell.

Beneficial mutation rates

To estimate the spectrum of beneficial mutation rates in the serial batch conditions, we consider only lineages that are identified as adaptive in one replicate but not the other (that is, are unlikely to contain mutations that occurred before barcoding, Supplementary Information section 9 and 10). Analysing the total number of cells with each s yields the best estimate of the mutation rate spectrum (Fig. 3a and Extended Data Fig. 3a, insets, and Supplementary Information section 11.1). These estimates are worse for fitness effects that have only occurred a few times. We find that beneficial mutations with $s > 5\%$ occur at a rate of $\sim 1 \times 10^{-6}$ per cell per generation (Supplementary Information section 11.2, Fig. 3a and Extended Data Fig. 3a, insets), a rate that is consistent across replicates. Using a fluctuation test[28,29], we find that the ancestor to our barcoded strains has a spontaneous mutation rate in non-repeat regions of $\sim 4 \times 10^{-10}$ per nucleotide per generation (Supplementary Information section 1.7)[30,31]. This implies that mutations in ~0.04% of the genome, ~5,000 bases, confer beneficial fitness effects of >5%. This target size is broadly consistent with previous reports[32,33], although it will certainly depend on the selective conditions. The beneficial mutation rate includes all events that have a heritable effect on fitness, and could include point mutations, indels, large genomic rearrangements or duplications, whole-genome duplications, and possibly even heritable epigenetic modifications. Reported beneficial mutation rates depend on the range of fitness effects that can establish and be detected. For example, if we include lower fitness effect mutations that are mostly pre-existing ($2\% < s < 5\%$), we find a beneficial mutation rate that is ~50× higher. However, as we discuss below, the total beneficial mutation rate is not necessary for a predictive understanding of the evolutionary dynamics. Instead, knowledge of the rate of mutation to the range of fitness effects that drive the dynamics is what is needed.

Mutation rate spectrum

Several authors have used extreme value theory to predict that the spectrum of beneficial mutation rates is exponential[34,35], with some experiments that sample small numbers of beneficial mutations supporting[17,36,37] or contradicting[38] this prediction. We do not find support for an exponential or even a monotonically decreasing distribution. Rather, most mutations we observe are confined to a narrow range of fitness effects ($2\% < s < 5\%$). At larger fitnesses, the distribution is relatively flat with two slight peaks in the fitness ranges 7–8% and 10–11%, a feature that is

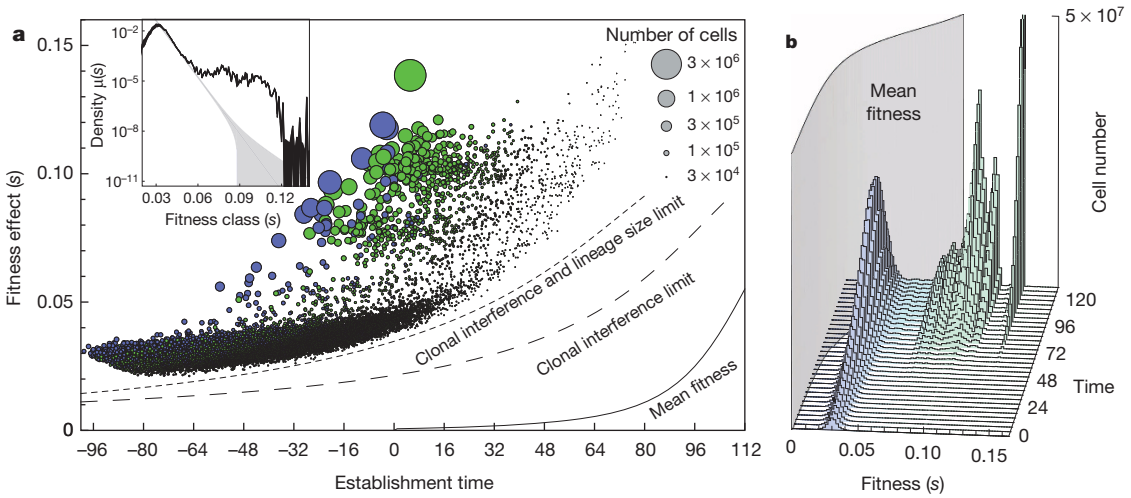


Figure 3 | Fitness effects, establishment times, and population dynamics. **a**, Scatter plot of τ and s of all ~25,000 beneficial mutations (circles) identified in E1. Circle area represents the size of the lineage at generation 88. Purple circles indicate lineages with mutations that occurred in the period of common growth ($t < 0$) that were sampled into, and established in, E1 and E2. Green circles indicate lineages that were identified as adaptive in only one replicate and likely contain mutations that arose after $t = 0$. Lines indicate the time limits before which mutations must occur in order to establish (large dash) or be observed (small dash). These limits trail the mean fitness (solid line) by ~$1/s$ generations. Inset, the spectrum of mutation rates, $\mu(s)$, as a function of fitness effect, s inferred from mutations that likely occurred after $t = 0$ (Supplementary Information section 10.2). The y axis is the mutation rate density, so the mutation rate to a range, Δs, is obtained by multiplying this density by Δs. The total beneficial mutation rate to $s > 5\%$ is inferred to be $\sim 1 \times 10^{-6}$ and is consistent across replicates. The observed spectrum is not exponential (grey line, with the error range shaded). **b**, the distribution of the number of adaptive cells binned by their fitness over time. As the mean fitness (grey curtain) surpasses the fitness of a subpopulation, cells with that fitness begin to decline in frequency.

consistent across replicates (Fig. 3a and Extended Data Fig. 3a, insets). Mutation rates to these two peaks are consistent with genomic target sizes of loss-of-function mutations for a single gene (~300 base pairs[31]); these have previously been shown to be adaptive in yeast grown in simple environments[1,3,4,39]. Weaker effect mutations ($s < 2\%$), which are hard to detect because they are rapidly outcompeted before establishing, do not occur at high enough rates to impact the population dynamics (Supplementary Information section 9.3).

Distribution of establishment times

For mutations that establish, τ roughly corresponds to the time at which a beneficial mutation occurred, with an uncertainty of a few times $1/s$ due to variability in initial stochastic drift (Supplementary Information section 4.1). Establishment times are broadly distributed ($-90 < \tau < 48$). Lineages containing beneficial mutations with very negative τ ($-90 < \tau < -40$) are usually identified as adaptive in both replicates (Fig. 3a and Extended Data Fig. 3a, purple). Establishment times as negative as -90 generations are expected[39] because of beneficial mutations that occur during the period of common growth ($t < 0$, Supplementary Information section 10.1). Indeed the number of pre-existing beneficial mutations is broadly consistent with the beneficial mutation rate we infer (Supplementary Information section 10). We observe very few mutations with $\tau > 48$ for the reasons that follow.

A beneficial mutation with a fitness effect s, that occurs in generation t will typically take another ~ $1/s$ generations to reach a size large enough to grow exponentially[19]. If before this time the mean fitness has increased by more than s, the mutation will decline in frequency and never grow exponentially. Thus, there is a time limit after which a beneficial mutation that occurs is unlikely to establish (Fig. 3a and Extended Data Fig. 3a, larger dashed lines). This time limit is shorter for smaller s for two reasons: (1) small s mutations must drift to higher numbers in order to establish, and (2) the mean fitness of the population surpasses its fitness advantage in a shorter time. A mutation with $s < 2\%$ is therefore extremely unlikely to establish because this limit is reached quickly. Thus, a fundamental lower limit on which fitness-effects can establish emerges from the population dynamics.

In addition to establishing, beneficial mutations in our assay must also grow to a large enough number to be detectable above the number of neutral (ancestral) cells remaining in its lineage. This shortens the time window in which a beneficial mutation must occur to be observed (Fig. 3a and Extended Data Fig. 3a, smaller dashed lines and Supplementary Information section 9). Beneficial mutations we are unable to detect (those occurring close to, or after, the time limit) never reach sizes much above their establishment number ($1/s$), are rapidly outcompeted, and typically go extinct. Such mutations are unlikely to have a significant impact on the population dynamics. Deleterious mutations are largely irrelevant here: given the mean fitness increases by a few percent in ~80 generations, a deleterious mutation will not rise to high frequency unless it occurs contemporaneously in a cell with a large beneficial mutation, and even then is unlikely to reach high frequencies[40].

Overall population dynamics

Plotting the fitness distribution of all adaptive cells over time reveals that massive clonal interference underlies the population dynamics (Fig. 3b and Extended Data Fig. 3b). Many beneficial mutations (~20,000 observed in E1, ~11,000 observed in E2) of small s ($2\% < s < 5\%$, the 'low fitness class') drive the mean fitness early ($t < 72$), but begin to be outcompeted by cells with larger s (~10%) that stem from fewer beneficial mutations (~5,000 in E1 and ~3,000 in E2). For the first ~80 generations the mean fitness trajectory in both replicates is strikingly similar (grey curtain, Fig. 3b and Extended Data Fig 3b and Supplementary Information section 6.5). However, by ~112 generations, the mean fitness is being driven by ~100 of the most beneficial mutations ($s >$ 10%). Because mutations to these higher fitness effects are rare, they display stochastic establishment times that lead to differences in the mean fitness between the two replicates at late times (Supplementary Information section 6.5). In E2, these higher fitness mutations happen to establish earlier, contributing to a quicker decline in the low fitness class, and fewer observed adaptive lineages overall. By generation ~132, we observe that the low fitness class has shrunk to a small fraction of the population. This, however, does not mean that cells in this class are inconsequential: they prevent mutations with even smaller s from establishing. Because they are so numerous early in the evolution, some cells in this class are likely to accumulate additional beneficial mutations whose expansion could enable them to eventually outcompete cells that initially acquired higher s mutations.

Fitness effects that drive the early evolutionary dynamics in this large population are a predictable consequence of the population size and spectrum of mutation rates. The range of s at the highest frequency at time t (those that are dominating the increase in mean fitness) are those that maximize $st + \log(\mu(s))$, with $\mu(s)$ being the mutation rate to s (Supplementary Information section, 11.1). That is, the most important fitness effects at a given time are determined by a balance between being sufficiently probable to have established multiple times and sufficiently fit to have grown to large cell numbers.

Adaptive lineages that accumulate an additional beneficial mutation in a cell with an existing beneficial mutation (double mutants), can impact the dynamics. However, double mutants are rare before ~168 generations because most single mutants are not yet present at high enough cell numbers to acquire a second mutation that establishes. We estimate that fewer than ~50 of the inferred values of s and τ are impacted by double mutants (~0.2% of all adaptive lineages, Supplementary Information section 4.5). Ecological changes in the environment caused by mutants can result in frequency-dependent selection and impact the evolutionary dynamics. But, over the time range used to infer fitnesses (up to ~100 generations) our observations are consistent with the simplifying assumption that beneficial mutations have frequency-independent fitness effects and thus subpopulations only interact via competition against the mean fitness (Fig. 2a and Extended Data Fig. 2a, insets).

Discussion

Tracking a large number of small lineages provides a granular view of evolutionary dynamics that is not possible by other methods[1–3]. By focusing on sequencing just 0.002% of the genome, we gain almost five orders of magnitude in frequency resolution over genome sequencing approaches. This enables us to identify tens of thousands of independent beneficial mutations, some of which never reach frequencies above $\sim 10^{-5}$. By contrast, our previous population sequencing approach[1], which detected mutations at frequencies above ~1%, would have identified only ~15 adaptive lineages in this study (Fig. 4, Supplementary Information section 9.4). Furthermore, barcode tracking yields estimates of the fitness effects and occurrence times for all changes that convey substantial fitness advantage, whether or not they are amenable to being identified via genome sequencing.

Our results show that in an asexually evolving population of $\sim 10^8$ cells, a large number of independent beneficial mutations drive adaptation. While individually each mutation is rare and occurs stochastically, collectively they have a predictable impact on the population dynamics. In large populations therefore, the early evolutionary dynamics is almost deterministic: it only becomes stochastic when mutations so rare that they have occurred only a handful of times, or multiple mutations on the same genome, expand to an appreciable fraction of the population. Mutations with certain fitness effects play a far more important role in driving the dynamics than others, resulting in a subtle interplay between deterministic and stochastic effects.

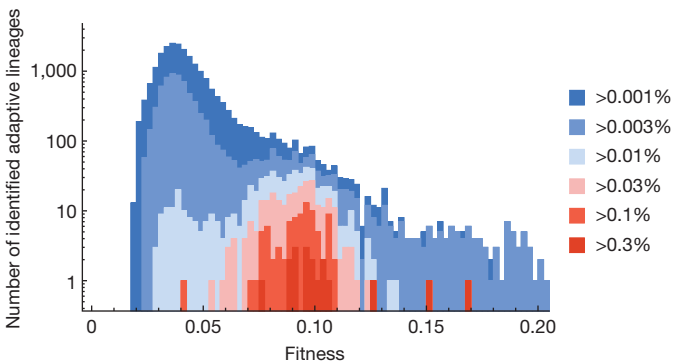


Figure 4 | The need for high frequency resolution. The fitness spectrum of adaptive lineages in replicate E1 that could be identified within the first 100 generations at different frequency resolution thresholds.

High-resolution lineage tracking is a powerful tool to study many questions important to evolution. Using this system across many environmental regimes, perhaps for longer periods of time than in this work, the relationships between adaptation rate, environment, and ecology could be quantitatively studied. A potential limitation of lineage tracking is that barcode diversity will always diminish over time. However, the possibility of adding barcodes at different times over the course of an evolution could provide a means to overcome this.

Cancer and microbial infections can have population sizes up to $\sim 10^{12}$ cells in a single individual, suggesting that massive clonal interference and complex population dynamics are likely to characterize disease progression and drug resistance[41–44]. Although mutations that rise to high frequencies are often emphasized, much larger numbers of low frequency mutations could be at least as important for disease progression or drug resistance. To study these low-frequency mutations, barcode tracking could be implemented in pathogenic microbes, cancer cell lines, or even animal tumour models[45–48]. Indeed, lineage tracking has the potential to identify the treatment regimes that most effectively slow the rate of adaptation. By randomly picking clones and sequencing their barcodes, one can cheaply identify many clones belonging to independent adaptive lineages. By sequencing the genomes of these clones, the mutational determinants for a broad range of beneficial fitness effects can be discovered. In combination with whole genome sequencing, lineage tracking therefore offers a powerful method by which to characterize the mutational spectrum underlying evolution, disease progression and drug resistance.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 11 September 2014; accepted 2 February 2015.
Published online 25 February; corrected online 11 March 2015.

1. Kvitek, D. J. & Sherlock, G. Whole genome, whole population sequencing reveals that loss of signaling networks is the major adaptive strategy in a constant environment. *PLoS Genet.* **9,** e1003972 (2013).
2. Herron, M. D. & Doebeli, M. Parallel evolutionary dynamics of adaptive diversification in *Escherichia coli*. *PLoS Biol.* **11,** e1001490 (2013).
3. Lang, G. I. *et al.* Pervasive genetic hitchhiking and clonal interference in forty evolving yeast populations. *Nature* **500,** 571–574 (2013).
4. Lang, G. I., Botstein, D. & Desai, M. M. Genetic variation and the fate of beneficial mutations in asexual populations. *Genetics* **188,** 647–661 (2011).
5. Ding, L. *et al.* Genome remodelling in a basal-like breast cancer metastasis and xenograft. *Nature* **464,** 999–1005 (2010).
6. Shah, S. P. *et al.* Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. *Nature* **461,** 809–813 (2009).
7. Mardis, E. R. *et al.* Recurring mutations found by sequencing an acute myeloid leukemia genome. *N. Engl. J. Med.* **361,** 1058–1066 (2009).
8. International Cancer Genome Consortium *et al.* International network of cancer genome projects. *Nature* **464,** 993–998 (2010).
9. Pleasance, E. D. *et al.* A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* **463,** 191–196 (2010).
10. Weinreich, D. M., Delaney, N. F., DePristo, M. A. & Hartl, D. L. Darwinian evolution can follow only very few mutational paths to fitter proteins. *Science* **312,** 111–114 (2006).
11. Young, B. C. *et al.* Evolutionary dynamics of *Staphylococcus aureus* during progression from carriage to disease. *Proc. Natl Acad. Sci. USA* **109,** 4550–4555 (2012).
12. Holden, M. T. G. *et al.* A genomic portrait of the emergence, evolution, and global spread of a methicillin-resistant *Staphylococcus aureus* pandemic. *Genome Res.* **23,** 653–664 (2013).
13. Desai, M. M., Walczak, A. M. & Fisher, D. S. Genetic diversity and the structure of genealogies in rapidly adapting populations. *Genetics* **193,** 565–585 (2013).
14. Neher, R. A. & Hallatschek, O. Genealogies of rapidly adapting populations. *Proc. Natl Acad. Sci. USA* **110,** 437–442 (2013).
15. Hegreness, M., Shoresh, N., Hartl, D. & Kishony, R. An equivalence principle for the incorporation of favorable mutations in asexual populations. *Science* **311,** 1615–1617 (2006).
16. Kao, K. C. & Sherlock, G. Molecular characterization of clonal interference during adaptive evolution in asexual populations of *Saccharomyces cerevisiae*. *Nature Genet.* **40,** 1499–1504 (2008).
17. Imhof, M. & Schlötterer, C. Fitness effects of advantageous mutations in evolving *Escherichia coli* populations. *Proc. Natl Acad. Sci. USA* **98,** 1113–1117 (2001).
18. Gerrits, A. *et al.* Cellular barcoding tool for clonal analysis in the hematopoietic system. *Blood* **115,** 2610–2618 (2010).
19. Desai, M. M. & Fisher, D. S. Beneficial mutation selection balance and the effect of linkage on positive selection. *Genetics* **176,** 1759–1798 (2007).

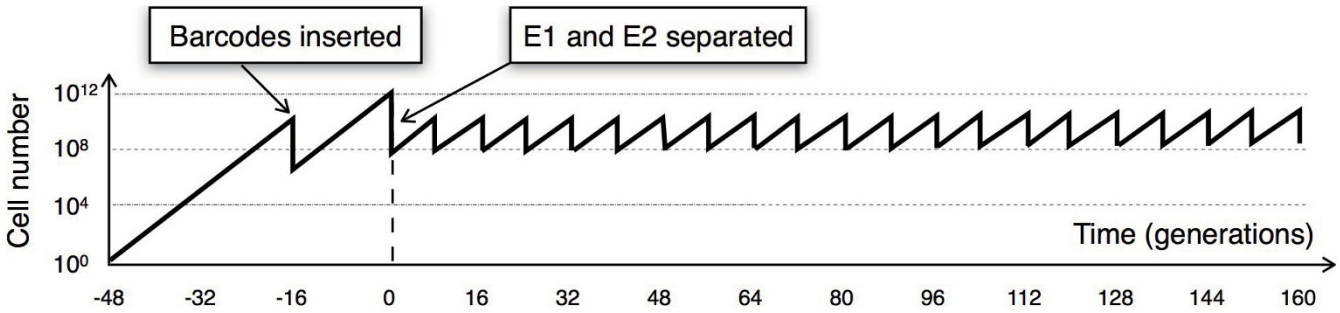
20. Charlesworth, B. The good fairy godmother of evolutionary genetics. *Curr. Biol.* **6,** 220 (1996).
21. Berns, K. *et al.* A large-scale RNAi screen in human cells identifies new components of the p53 pathway. *Nature* **428,** 431–437 (2004).
22. Smith, A. M. *et al.* Quantitative phenotyping via deep barcode sequencing. *Genome Res.* **19,** 1836–1842 (2009).
23. Lu, R., Neff, N. F., Quake, S. R. & Weissman, I. L. Tracking single hematopoietic stem cells *in vivo* using high-throughput sequencing in conjunction with viral genetic barcoding. *Nature Biotechnol.* **29,** 928–933 (2011).
24. Blundell, J. R. & Levy, S. F. Beyond genome sequencing: lineage tracking with barcodes to study the dynamics of evolution, infection, and cancer. *Genomics* **104,** 417–430 (2014).
25. Sternberg, N. & Hamilton, D. Bacteriophage P1 site-specific recombination. *J. Mol. Biol.* **150,** 467–486 (1981).
26. Austin, S., Ziese, M. & Sternberg, N. A novel role for site-specific recombination in maintenance of bacterial replicons. *Cell* **25,** 729–736 (1981).
27. Gerrish, P. J. & Lenski, R. E. The fate of competing beneficial mutations in an asexual population. *Genetica* **102–103,** 127–144 (1998).
28. Luria, S. E. & Delbrück, M. Mutations of bacteria from virus sensitivity to virus resistance. *Genetics* **28,** 491–511 (1943).
29. Lang, G. I. & Murray, A. W. Estimating the per-base-pair mutation rate in the yeast *Saccharomyces cerevisiae*. *Genetics* **178,** 67–82 (2008).
30. Lynch, M. *et al.* A genome-wide view of the spectrum of spontaneous mutations in yeast. *Proc. Natl Acad. Sci. USA* **105,** 9272–9277 (2008).
31. Zhu, Y. O., Siegal, M. L., Hall, D. W. & Petrov, D. A. Precise estimates of mutation rate and spectrum in yeast. *Proc. Natl Acad. Sci. USA* **111,** E2310–E2318 (2014).
32. Joseph, S. B. & Hall, D. W. Spontaneous mutations in diploid *Saccharomyces cerevisiae*: more beneficial than expected. *Genetics* **168,** 1817–1825 (2004).
33. Desai, M. M., Fisher, D. S. & Murray, A. W. The speed of evolution and maintenance of variation in asexual populations. *Curr. Biol.* **17,** 385–394 (2007).
34. Gillespie, J. H. Molecular evolution over the mutational landscape. *Evolution* **38,** 1116–1129 (1984).
35. Orr, H. A. The distribution of fitness effects among beneficial mutations. *Genetics* **163,** 1519–1526 (2003).
36. Kassen, R. & Bataillon, T. Distribution of fitness effects among beneficial mutations before selection in experimental populations of bacteria. *Nature Genet.* **38,** 484–488 (2006).
37. Rokyta, D. R., Joyce, P., Caudle, S. B. & Wichman, H. A. An empirical test of the mutational landscape model of adaptation using a single-stranded DNA virus. *Nature Genet.* **37,** 441–444 (2005).
38. Rokyta, D. R. *et al.* Beneficial fitness effects are not exponential for two viruses. *J. Mol. Evol.* **67,** 368–376 (2008).
39. Gresham, D. *et al.* The repertoire and dynamics of evolutionary adaptations to controlled nutrient-limited environments in yeast. *PLoS Genet.* **4,** e1000303 (2008).
40. Good, B. H., Rouzine, I. M., Balick, D. J., Hallatschek, O. & Desai, M. M. Distribution of fixed beneficial mutations and the rate of adaptation in asexual populations. *Proc. Natl Acad. Sci. USA* **109,** 4950–4955 (2012).
41. Salmon, S. E. & Smith, B. A. Immunoglobulin synthesis and total body tumor cell number in IgG multiple myeloma. *J. Clin. Invest.* **49,** 1114–1121 (1970).
42. Michaelson, J. S. *et al.* Predicting the survival of patients with breast carcinoma using tumor size. *Cancer* **95,** 713–723 (2002).
43. König, C., Simmen, H. P. & Blaser, J. Bacterial concentrations in pus and infected peritoneal fluid–implications for bactericidal activity of antibiotics. *J. Antimicrob. Chemother.* **42,** 227–232 (1998).
44. Wilson, M. L. & Gaido, L. Laboratory diagnosis of urinary tract infections in adult patients. *Clin. Infect. Dis.* **38,** 1150–1158 (2004).
45. Thomas, C. E., Ehrhardt, A. & Kay, M. A. Progress and problems with the use of viral vectors for gene therapy. *Nature Rev. Genet.* **4,** 346–358 (2003).
46. Bushman, F. *et al.* Genome-wide analysis of retroviral DNA integration. *Nature Rev. Microbiol.* **3,** 848–858 (2005).
47. Ran, F. A. *et al.* Double nicking by RNA-guided CRISPR Cas9 for enhanced genome editing specificity. *Cell* **154,** 1380–1389 (2013).
48. Mali, P. *et al.* RNA-guided human genome engineering via Cas9. *Science* **339,** 823–826 (2013).
49. Kivioja, T. *et al.* Counting absolute numbers of molecules using unique molecular identifiers. *Nature Methods* **9,** 72–74 (2011).
50. Lundberg, D. S., Yourstone, S., Mieczkowski, P., Jones, C. D. & Dangl, J. L. Practical innovations for high-throughput amplicon sequencing. *Nature Methods* **10,** 999–1002 (2013).

Supplementary Information is available in the online version of the paper.

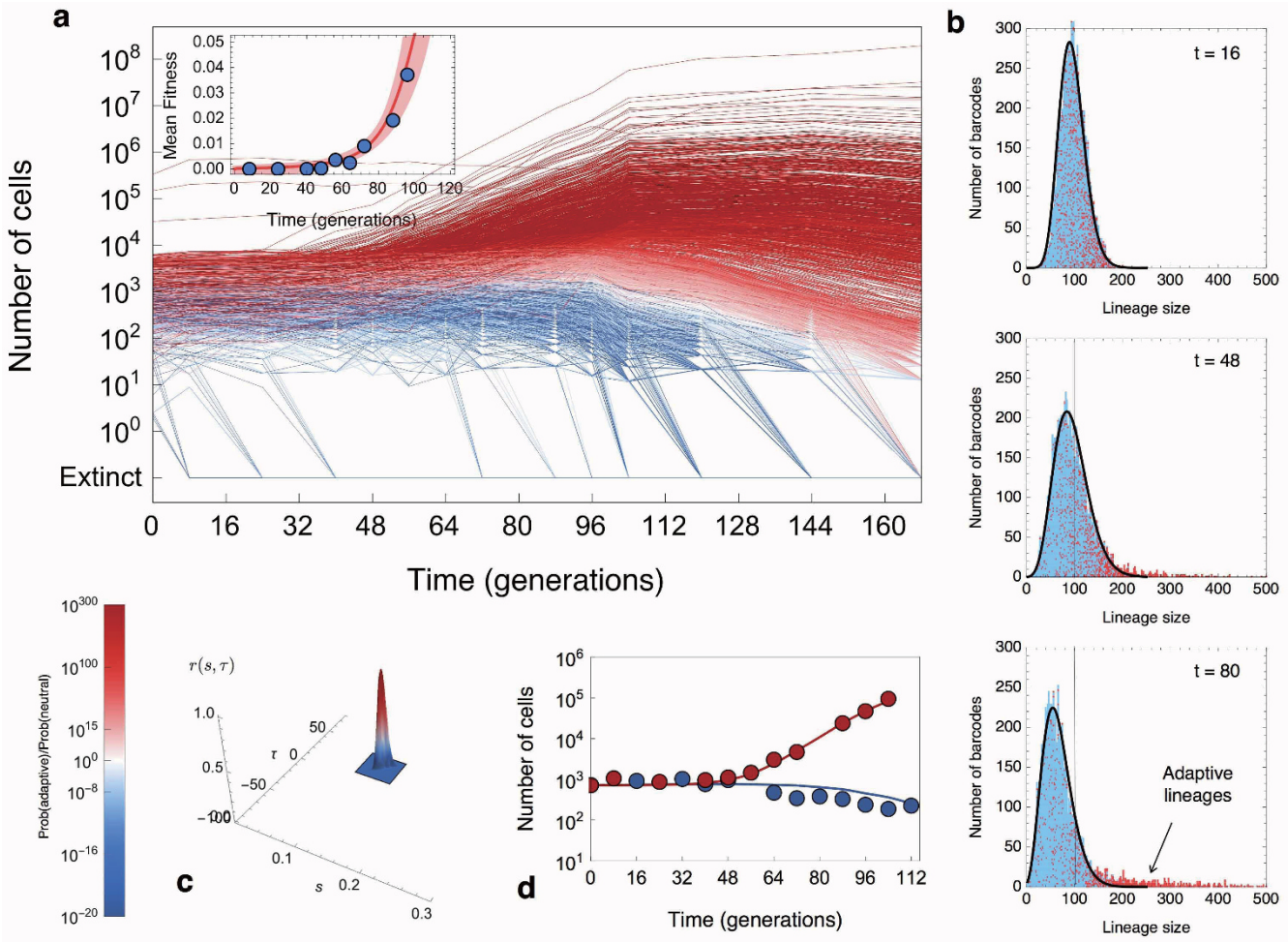
Acknowledgements The authors thank M. Siegal, K. Schwartz, B. Dunn, M. Jaffe, D. Kvitek, J. Thompson, D. Sellis, and Y. Zhu for discussions. FACS was performed at the Stanford Shared FACS Facility. We would like to acknowledge funding support from NIH grants R01 HG003328, 5-T32-HG-44-17 and R25 GM067110, NSF grants DMS-1120699 and PHY-1305433, Bio-X IIP6-63 grant from Stanford University, Gordon and Betty Moore Foundation grant no. 2919, and The Louis and Beatrice Laufer Center.

Author Contributions S.F.L. conceived of the barcoding system. S.F.L. and G.S. designed the barcoding system and evolution experiments. S.F.L., J.R.B., D.A.P., G.S. and D.S.F. developed the project vision. S.F.L. performed the barcoding and evolution experiments. S.V. and D.A.P. designed the pairwise competition assays. S.V. performed the pairwise competition assays. J.R.B. and D.S.F. developed theory and analysed the data. J.R.B. and S.F.L. wrote the paper. All authors edited the paper.

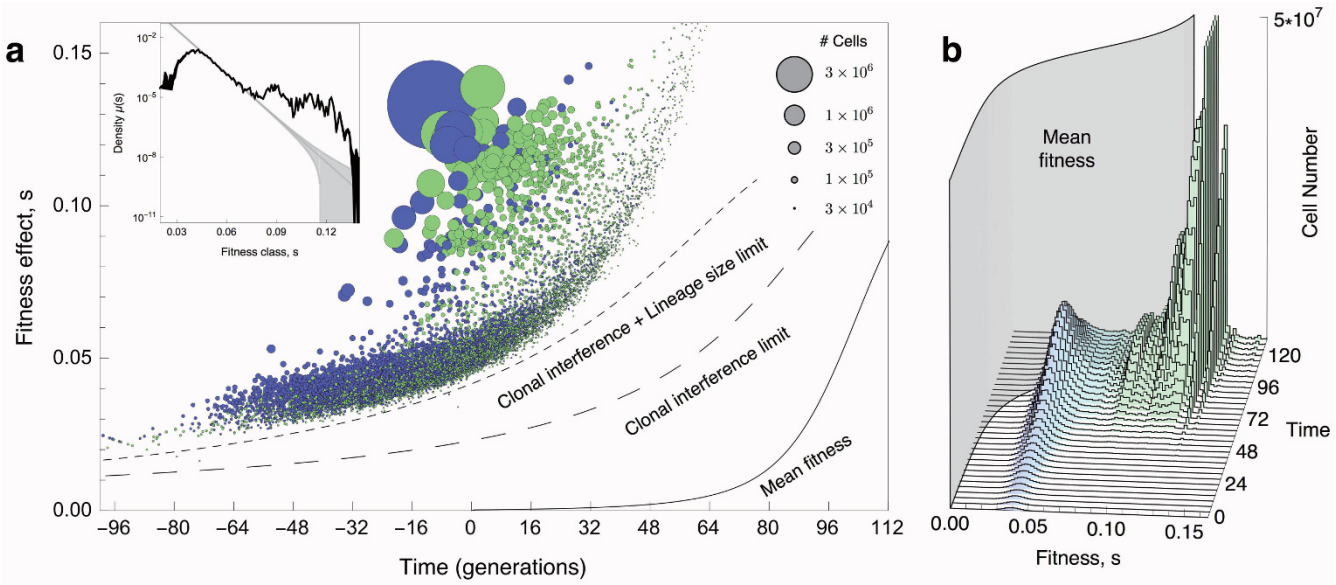
Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to D.S.F. (dsfisher@stanford.edu) or G.S. (gsherloc@stanford.edu).



Extended Data Figure 1 | Total population size over time. A single ancestral cell is grown for ~32 generations to ~10^{10} cells before barcodes are inserted. Cells that incorporate a barcode are grown for another 16 generations. The population is then divided into two replicates (E1 and E2) at $t = 0$. Beneficial mutations that occurred before barcoding can be sampled into both replicates.



Extended Data Figure 2 | Inferring the fitnesses and establishment times from lineage trajectories. **a**, Selected lineage trajectories and the mean fitness trajectory from replicate E2. **b**, The distribution of lineage sizes over time, for lineages that begin with ~100 ± 2 cells (vertical line). Adaptive lineages (red) begin to expand above the neutral expectation (black curve) and push neutral lineages to lower cell numbers (blue). **c**, The posterior probability distribution over s and τ for an adaptive lineage in E2. **d**, The measured trajectory of this lineage in E1 (unadaptive, blue circles) and E2 (adaptive, red circles) compared with the predicted trajectory with largest probability in E1 (blue line) and E2 (red line).



Extended Data Figure 3 | Fitness effects and establishment times for replicate E2. **a**, Scatter plot of τ and s of all ~14,000 beneficial mutations (circles) identified in E2. Circle area represents the size of the lineage at generation 88. Purple circles indicate lineages with mutations that occurred in the period of common growth ($t < 0$) that were sampled into, and established in, E1 and E2. Green circles indicate lineages that were identified as adaptive in only one replicate and likely contain mutations that arose after $t = 0$. Lines indicate the time limits before which mutations must occur in order to establish (large dash) or be observed (small dash). These limits trail the mean fitness (solid line) by ~$1/s$ generations. Inset, the spectrum of mutation rates, $\mu(s)$, as a function of fitness effect, **s** inferred from mutations that likely occurred after $t = 0$ (Supplementary Information section 10.2). The y axis is the mutation rate density, so the mutation rate to a range, Δs, is obtained by multiplying this by Δs. The total beneficial mutation rate to $s > 5\%$ is inferred to be $\sim 1 \times 10^{-6}$ and is consistent across replicates. The observed spectrum is not exponential (grey line, with the error range shaded). **b**, The distribution of the number of adaptive cells binned by their fitness over time. As the mean fitness (grey curtain) surpasses the fitness of a subpopulation, cells with that fitness begin to decline in frequency.