KEY RESOURCES TABLE

Please see attached Microsoft Word Document for the Key Resources Table.

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information regarding the manuscript may be directed to the lead contact Dmitri Petrov (dpetrov@stanford.edu). Requests for reagents may be directed to, and will be fulfilled by Gavin Sherlock (gsherloc@standford.edu).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

The yeast strains used in this study can be grown and maintained using standard methods (e.g. YPD media in test tubes, glycerol stocks for long term storage at $-80°$C), but should be propagated in the selection environment (glucose limited minimal media) for optimal phenotypic and fitness measurements.

METHOD DETAILS

Sampling Clones

Evolved yeast clones were isolated by plating for single colonies from frozen samples of the generation-88 time point of a previously reported serial batch transfer evolution experiment seeded by a population of individually genome-barcoded yeast cells (500,000 barcodes total); each batch cycle consisted of 8 generations of growth with glucose as the known limiting nutrient at an initial concentration of 1.5% (Levy et al., 2015). We selected 3,840 colonies from replicate experiment E1 and 960 colonies from replicate E2 for a total of 4,800 individual evolved clones. A portion of each colony was resuspended in 20% glycerol in 96 well plates and immediately frozen at -80° C; the remaining portion was used to identify the barcode residing in that clones genome. To identify the barcode, we amplified the genomic region carrying the barcode with PCR using the following primers:

PS1 - CCCGCAGAGTACTGCAATTT

PS2 - TGCACGAAAAGCAAACAAAC

The PCR products were purified using ExoSap-It (Affymetrix # 78200) and sequenced by Sanger technology, using PS2 as the sequencing primer, and then identified from among the set of 500,000 barcodes described in Levy et al. (2015).

Pairwise fluorescence competition assay measurements

Fluorescence-based fitness assays were conducted as in Levy et al. (2015). Briefly, the individual clone to be assayed was grown in liquid culture and then mixed with a YFP-tagged ancestral clone in a 1:9 ratio. This

mixture was sampled over 32 generations (four 8-generation batch cycles) in conditions identical to the initial evolution experiments of Levy et al. (2015). The relative frequencies of the sample and the ancestor were estimated at each time-point using flow cytometry at the Stanford Shared FACS facility. An exponential model was then fit to these data to estimate fitness.

Pooled-clone fitness measurement assay

Overview All 4,800 isolated clones were pooled into a single culture, then mixed with a clone with ancestral fitness in a 1:9 ratio and competed by culturing the mixture under conditions identical to the initial evolution for 32 generations (four 8-generation cycles), with samples being stored at every transfer. Barcode frequencies were tracked using Illumina HiSeq technology, and fitness was estimated using these frequency trajectories. We performed the fitness measurement assay a total of 11 times, in four independent batches, each time with two or three replicate flasks (see below for complete details).

Design of the Fitness measurement assay The goal of the fitness measurement assay is to cheaply, easily and accurately measure the fitness of many barcoded clones in parallel; any desired set of clones (as long as each clone contains a unique barcode at the same genomic position) can be pooled and used in this assay. In this protocol, we competed a pool of 4,800 clones sampled from generation 88 of the evolution experiments of Levy et al. (2015) against a clone with ancestral fitness, for a period of 32 generations of batch culture competition. We then estimated the frequencies of the clones at five timepoints by Illumina amplicon sequencing of their DNA barcodes. The frequencies from four of these five timepoints (described in section 3.4) were then used to estimate fitness (s). We used the same growth conditions as the evolution experiments of Levy et al. (2015), where the population goes through a bottleneck of $\approx 5 \cdot 10^7$ cells every 48 hours, i.e., at the time it is transferred to fresh media (≈ 8 generations between transfers).

To accurately measure the fitness of many clones, each adaptive lineage must be at a large enough frequency such that stochastic effects at the bottleneck process are mitigated. In addition, most of the initial population must consist of the ancestral genotype, so that fitness is measured in a condition dominated by the ancestor. To fulfill all of these criteria we first pooled 4,800 sampled clones, then seeded this pool at an initial frequency of 10% in the population and competed against an ancestral strain that made up $\approx$ 90% of the initial population. Each sampled clone thus had a population size in the bottleneck of $\approx 0.1 \cdot 5 \cdot 10^7/5000 = 1,000$ cells. However, there are biological fluctuations from stochasticity in the lag time before new growth after dilution, in birth/death fluctuations near the bottleneck, and in the sampling induced by the dilution process itself. These give rise to fluctuations in the bottleneck population from one cycle to the next of $\pm\sqrt{\beta n}$. In our experiments, we estimated $\beta \approx 10$ (see: Quantification and Statistical Analysis) so that these effects are relatively small. Furthermore, for beneficial mutations with $s > 1\%$, the systematic increases in population due to selection are larger than the stochastic fluctuations. Thus the stochastic effects are relatively small for most adaptive lineages in our fitness assay.

The large population size and the short time for the assay (32 total generations) also ensured that any new adaptive mutations that arose during the course of the assay had no significant impact on the fitness estimates of any single clone. The fraction of lineages that will get taken over by new mutations with a particular range of s can be approximately bounded by$\mu/s \cdot e^{sT}$, for a representative s. In our fitness assays, the total adaptive mutation rate per generation μ for mutations with $s > 5\%$ is $\approx 10^6$ (Levy et al., 2015), highly adaptive mutations have $s = 10\%$, and the assay is conducted for $T = 32$ generations. This gives us an upper bound of 10^4 for the expected fraction of lineages dominated by a mutant that came up during the

evolution. The effect of new adaptive mutations is thus negligible in our fitness assays.

Ideally, the fraction of the population that consists of mutant clones would remain a small fraction of the total population throughout the 32 generations of growth in these assays. The dynamics of the limiting resource depend on the physiology of the dominant type(s) in the population, and if a non-ancestral type dominated it could change those dynamics in a way which affected different mutants differently. Additionally, if the ancestral clone gets to a low frequency, it becomes challenging to estimate fitness of a mutant relative to the ancestral type. As we describe in Figure S1, the mutant clones started at a higher frequency within the barcoded class than planned and reached substantial frequency in the total pool at late timepoints. This might be one of the sources of systematic variations of measured fitnesses between experiments that we observe. Future experiments would be well served to minimize the effect of large populations of mutants changing the environment.

Construction of a strain with ancestral fitness

We realized that if a barcoded ancestor was used in the competition experiments, a large number of reads (up to 90% of reads for the initial timepoint) would be spent sequencing the ancestral clone, leading to a waste of sequencing capacity when estimating the frequency of the 4,800 evolved clones in the pool. We attempted to use a barcode-less clone for the ancestor, but found that the PCR reactions failed when the barcode sequence was present in such a small proportion of the population. Therefore, we developed a barcoded ancestral strain with a restriction site at the barcode locus to serve as the reference strain, allowing us to remove the amplicons derived from the reference strain by restriction enzyme digestion after the PCR step, saving us a significant amount of sequencing cost.

We used the following primers in constructing the modified ancestor:

RE-SbfI-F	atcg cctgcagg aaacgaagataaatcatgtc
RE-ApaLI-2R	atcg gtgcac ctgtcaacactgttccaact
RE-ApaLI-2F	atcg gtgcac ataacttcgtataatgtatg
RE-XhoI-R	atcg ctcgag tcatgtaattagttatgtca

We used the plasmid pBAR3 (Levy et al., 2015) as a template to generate two separate PCR products from the above primers, which were then digested and ligated together to form the final construct for transformation. Primers RE-SbfI-F and RE-ApaLI-2R were used to generate the one PCR product (amplicon A), while primers RE-ApaLI-2F and RE-XhoI-R generated amplicon B. Amplicon A was then digested with SbfI and ApaLI, amplicon B was digested with ApaLI and XhoI and the pBAR3 plasmid was digested with SbfI and XhoI. These three digestion products were mixed and ligated simultaneously to generate complete plasmids containing the ApaLI site in the barcode region. The ligation product was transformed directly into SHA185 (Levy et al., 2015) to generate the modified ancestral clone. The presence of the ApaLI site in the barcode locus was verified through amplification of the barcode locus of these transformants, digestion of the resulting product by ApaLI and gel electrophoresis. A number of validated modified ancestral clones were screened for ancestral fitness using the fluorescent pairwise-competition fitness assays (described in Levy et al. (2015)), and the clone with the fitness closest to ancestral was selected for use in the sequencing based fitness measurement assays.

Optimizing sequencing costs

We determined that we needed $\approx$ 1 million reads per timepoint for Illumina amplicon sequencing of the 4,800-pool barcodes in order to accurately estimate the frequency of the clones. For the initial timepoint, we expect $\approx$ 200 reads per clone at this read depth. The frequencies of clones with large fitness effects (either positive or negative) are expected to change the most between samples (taken every 8 generations). A $\approx$ 20% fitness effect (the largest fitness effect observed in Levy et al. (2015)) would result in a four-fold change in frequency over 8 generations, resulting in the subsequent sample having $\approx$ 800 reads for the clone if it is adaptive and $\approx$ 50 reads for the clone if it is deleterious. These read depths are sufficient to have a small amount of sampling noise during the amplicon sequencing process (Levy et al. (2015) Supplementary Methods section 5), and thus $\approx$ 1 million reads per sample are adequate for our purposes. After allowing for variation in read depth when multiplexing samples (one sample per timepoint), the presence of a small amount of reads from the ancestral reference strain (due to low levels of undigested PCR products) and the use of a 25% Phi-X library spike-in to properly calibrate the Illumina machines, we ended up pooling 9 assays worth of samples ($\approx$ 40 samples) per lane of Illumina HiSeq 2000. As each lane results in $\approx$ 200 million reads, this gave us $\approx$ 4 million raw reads per sample. After removing reads from the reference strain and PCR duplicates, we had 1-3 million reads per sample for our estimation.

As we sequenced 5 timepoints per fitness assay replicate, this protocol costs $\approx$ \$0.06 USD to measure the fitness of a single clone per replicate. It takes a single person about one month to conduct both the fitness measurement assays and library preparation for amplicon sequencing, showing that this is truly a fast, accurate and cost-effective way to estimate the fitness of thousands of clones in parallel.

Pooling the 4,800 clones sampled at generation 88

The 4,800 sampled clones from generation 88 of the two evolution replicates of Levy et al. (2015) were stored in glycerol stocks in 50 96-well plates. As it was impractical to pool all of these clones together at once, we constructed the pool in batches of 192 clones (2 plates). Each clone was grown from freezer stock in 800μL of M3 medium (the medium used in the evolution experiments) in 96 well plate format at 30°C for 2 days so that all lineages reached saturation. 400μL 40% glycerol were added to each well and mixed, after which 400μL mixture from each well were pooled into a single vessel. Thus, for every two plates (192 total clones) we had an 80mL pool stored in two 50mL tubes at -80°C. This procedure was repeated for pairs of plates over the course of a few weeks until we had 25 frozen pools, each of which represented two plates worth of clones. As we used a multi-pronged pinner to take clones from frozen stock and pin them into 96-well plates, a small percent of clones were not successfully recovered from frozen stock and therefore not included in the pool. The 25 frozen pools were then thawed simultaneously at room temperature and mixed into a single vessel. This vessel thus contained cells from $\approx 4,800$ clones (excluding those that were not recovered from frozen stock). We dispensed 1 mL aliquots into 1.5mL eppendorf tubes, which were stored at -80°C.

We found that the clones have a wide range of frequencies in the pool, spanning nearly 3 orders of magnitude. To test whether this wide frequency range had a significant effect on fitness, we generated another pool of 500 of these clones where all clones were grown and pooled simultaneously, instead of in batches, and the fitness assay was begun without any freeze-thaw cycles to minimize the number of generations of pooled growth before the beginning of the fitness assay. Our fitness measurement results are highly consistent with the results of the 4,800 clone pool (Figure 2), suggesting that the wide range of initial frequencies, freeze-thaw effects nor the presence of additional generations of growth in the pool substantially change our fitness

estimates.

Conducting the fitness measurement assay

The fitness measurement assay was designed to assay the fitness of a large number of adapted clones in bulk against a reference clone. We conducted the fitness measurement assay on the pool of 4,800 clones in four batches with slightly different protocols.

To conduct the fitness assays, we first streaked the modified ancestral clone from frozen stock onto M3 agar plates. We selected a single colony and inoculated it into 3mL fresh M3 media and grew it for 2 days so that it reached saturation. 400μL of cell culture were then inoculated into 100mL M3 medium (the medium used in the evolution experiments of Levy et al. (2015)) in 500mL DeLong flasks (Bellco # 2510-00500). We also thawed out 1mL of the 4,800 clone pool, spun it down, removed supernatant, re-suspended the cells in M3 medium (to remove glycerol) and then inoculated the entire volume into a separate flask of 100mL M3 medium. After 2 days of growth at 30°C and 223 RPM in a shaking incubator, the cultures were saturated, and we mixed the ancestral culture with the pool in a 1:9 ratio accounting for variation in particle counts between the two cultures (Beckman Coulter) resulting in $\approx$ 100mL of mixture. 400μL of this mixture were then used to inoculate 3 replicate fitness assay cultures. The replicate fitness assay cultures were grown under conditions identical to the initial evolution conditions (Levy et al., 2015) for a total of 4 growth cycles or 32 generations with 1:250 dilutions for every transfer. The remainder of the 100mL culture after the initial mixture and each transfer was aliquoted in two 50mL conical tubes, spun down at 3000 rpm for 5 minutes, re-suspended in 6mL sorbitol solution (0.9M sorbitol, 0.1M Tris-HCL pH 7.5, 0.1M EDTA pH 8.0) and frozen at -20°C (-80°C is also acceptable). This procedure was done for three different batches of assays (batches # 1, # 3 and # 4). The 500 clone pool measurements followed a similar protocol except for not conducting the recovery growth from the freezer stock, as the 500 clone pool fitness assays were conducted without freezing the population (so one two-day growth cycle between the initial pooling and the mixing of the pool with the ancestor to begin the fitness assay).

For the batch # 2 containing two replicates (the third replicate did not generate sufficient sequencing data for analysis), after the initial 2 day growth of the separate ancestral and pool cultures in 100mL M3 media performed as for the first batch of experiments, we transferred $5 \cdot 10^7$ cells from each culture into 100mL of fresh M3 medium and grew them separately for 2 days before mixing and beginning the assay as before. This second 48-hour growth was done to accustom the cells to the medium and to minimize the freezer effects before beginning the assay. The number of cells/mL were determined using a Coulter particle counter to transfer $5 \cdot 10^7$ cells for each transfer, rather than the 400μL transfers done by Levy et al. (2015). This was done to ensure a more consistent dilution regime, and in practice worked out to nearly the same regime as the evolution experiments as we transferred $\approx$ 400μL per cycle under these conditions.

In effect, while all three batches were tracked for 32 generations of growth, we used the data from generations 8-32 for fitness estimation in batches 1 and 3, and data from generations 0-24 in batch 2.

DNA extractions from each sample

For each sample (representing one time-point in one replicate), we conducted DNA extractions as follows (starting from 50mL of cells spun down, then re-suspended and frozen in 5mL of sorbitol solution: 0.9M sorbitol, 0.1M Tris-HCL pH 7.5, 0.1M EDTA pH 8.0). We thaw the frozen samples at room temperature,

resuspend the cells by vortex and transfer 750μL of cells to a 2mL screw cap tube. The cells are then collected by high speed centrifugation, the supernatant is removed and the cells are washed in 500μL sterile H_2O. The water is again removed by centrifugation. We then add 200μL Triton SDS buffer (2% (v/v) Triton X-100, 1% (w/v) SDS, 100mM NaCl and 1mM Na_2 EDTA) to the cells, along with 200μL 25:24:1 phenol : chloroform : isoamyl alcohol and $\approx$ 200μL 0.1mm glass beads. This mixture is vortexed at high speed for 15 minutes. We then add 200μL pH8.0 TE buffer to the tubes in a fume hood, then spin the tubes for 2 minutes in a microcentrifuge at high speed to collect the cellular debris. The aqueous layer is transferred to a 2mL yellow phase lock tube (5 PRIME # 2302830), which is then spun for 5 minutes at high speed in a microcentrifuge. The supernatant from the phase lock tube is transferred to a clean 2mL eppindorf tube, along with 1mL cold 100% ethanol. This is mixed by inversion, which should visibly precipitate the DNA. The DNA is collected by centrifugation for 2 minutes at high speed, after which the supernatant is discarded. The DNA pellet is resuspended in 400μL TE buffer, to which we add 50μL 10mg/mL RNAse A and incubate for 15 minutes at 37°C. We add 10μL 4M ammonium acetate plus 1mL 100% ethanol to the mixture and mix by inversion. The DNA is collected again by centrifugation for 2 minutes at high speed, after which we remove the supernatant and let it air dry for 2 minutes before finally re-suspending the pellet in 150μL EB buffer (10mM Tris-Cl, pH 8.5). We dilute this re-suspended DNA to 75ng/μL in EB for use in the PCR reactions (lower yields are acceptable as long as the concentration is at least $\approx 40ng/\mu$L).

PCR amplification of the barcode locus

We used a two-step PCR protocol to amplify the barcodes from the DNA that is very similar to the protocol used in Levy et al. (2015).

We use barcoded primers for the first PCR cycle. Different combinations of forward and reverse primers are used for each sample so that we can multiplex many samples together in a single HiSeq lane. The "N" positions in these primers are random nucleotides used to uniquely index each amplicon product to remove PCR duplicates from downstream analysis. All of these primers are HPLC purified to ensure that they are the correct length.

Forward primers

FP1	ACACTCTTTCCCTACACGACGCTCTTCCGATCT	NNNNNNNN	CGATGTTT	AATATG-GACTAAAGGAGGCTTTT
FP2	ACACTCTTTCCCTACACGACGCTCTTCCGATCT	NNNNNNNN	ACAGTGTT	AATATG-GACTAAAGGAGGCTTTT
FP3	ACACTCTTTCCCTACACGACGCTCTTCCGATCT	NNNNNNNN	TGACCATT	AATATG-GACTAAAGGAGGCTTTT
FP4	ACACTCTTTCCCTACACGACGCTCTTCCGATCT	NNNNNNNN	GCCAATTT	AATATG-GACTAAAGGAGGCTTTT
FP5	ACACTCTTTCCCTACACGACGCTCTTCCGATCT	NNNNNNNN	ATCACGTT	AATATG-GACTAAAGGAGGCTTTT
FP6	ACACTCTTTCCCTACACGACGCTCTTCCGATCT	NNNNNNNN	CAGATCTT	AATATG-GACTAAAGGAGGCTTTT
FP7	ACACTCTTTCCCTACACGACGCTCTTCCGATCT	NNNNNNNN	GGCTACTT	AATATG-GACTAAAGGAGGCTTTT
FP8	ACACTCTTTCCCTACACGACGCTCTTCCGATCT	NNNNNNNN	TAGCTTTT	AATATG-GACTAAAGGAGGCTTTT

Reverse primers

RP1	CTCGGCATTCCTGCTGAACCGCTCTTCCGATCT	NNNNNNNN	TATATACGC	TC-GAATTCAAGCTTAGATCTGATA
RP2	CTCGGCATTCCTGCTGAACCGCTCTTCCGATCT	NNNNNNNN	CGCTCTATC	TC-GAATTCAAGCTTAGATCTGATA
RP3	CTCGGCATTCCTGCTGAACCGCTCTTCCGATCT	NNNNNNNN	GAGACGTCT	TC-GAATTCAAGCTTAGATCTGATA
RP4	CTCGGCATTCCTGCTGAACCGCTCTTCCGATCT	NNNNNNNN	ATACTGCGT	TC-GAATTCAAGCTTAGATCTGATA
RP5	CTCGGCATTCCTGCTGAACCGCTCTTCCGATCT	NNNNNNNN	ACTAGCAGA	TC-GAATTCAAGCTTAGATCTGATA
RP6	CTCGGCATTCCTGCTGAACCGCTCTTCCGATCT	NNNNNNNN	TGAGCTAGC	TC-GAATTCAAGCTTAGATCTGATA
RP7	CTCGGCATTCCTGCTGAACCGCTCTTCCGATCT	NNNNNNNN	CTGCTACTC	TC-GAATTCAAGCTTAGATCTGATA
RP8	CTCGGCATTCCTGCTGAACCGCTCTTCCGATCT	NNNNNNNN	GCGTACGCA	TC-GAATTCAAGCTTAGATCTGATA

For the first cycle, for each sample, we performed 12 PCR reactions

Master Mix:

- 325μL OneTaq 2x master mix (NEB # M0482L)
- 13μL 10uM FP
- 13μL 10uM RP
- 156μL sample DNA (diluted to 75ng/μL or the entire DNA sample if between 40 - 75ng/μL DNA)
- 143μL dH_2O
- 650μL total

50μL of master mix is aliquoted into 12 wells of a 96 well plate and the following PCR reaction is run on a thermocycler:

1. 94°C 10 minutes
2. 94°C 3 minutes
3. 55°C 1 minute
4. 68°C 1 minute
5. Repeat steps 2-4 for a total of 3 cycles
6. 68°C 1 minute
7. Hold at 4°C

We then add 250μL of P1 buffer from QIAquick PCR purification kits (Qiagen # 28106) to each PCR reaction and then perform PCR cleanups following the standard Qiagen protocol in two columns (6 PCR reactions pooled into each column). This results in 50μL eluate of purified PCR product in two tubes for each sample.

For the second step of PCR, we use the following HPLC purified primers (where x is a phosphothioate group)

PE2 - xAATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCxT (read1)

PE1 - xCAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCxT (read2)

Master mix:

- 175μL PrimestarMAX 2x master mix (Clontech # R045B)
- 7μL 10μM PE1
- 7μL 10μM PE2
- 90μL purified PCR product
- 71μL dH_2O
- 350μL total

50μL of this master mix is added to each of 6 wells in a 96 well PCR plate, and the following reaction is run:

1. 98°C 2 minutes
2. 98°C 10 seconds
3. 69°C 15 seconds
4. 72°C 15 seconds
5. Repeat steps 2-4 for a total of 24 cycles
6. 72°C 1 minute
7. Hold at 4°C

250μL of Buffer P1 from the Quiagen kit is again added to each of these PCR wells, and all 6 wells are used in a single PCR purification protocol to generate a single tube with 50μL elutent with purified PCR product from the sample.

Removal of the reference strain amplicons using restriction digestion and size selection

We conducted the ApaLI digest of the reference strain reads as follows. We added 60μL H_2O, 10μL 10XCutSmart buffer (NEB # B7204S), 5μL ApaLI enzyme (NEB # R0507S) and 25μL of the purified

PCR product to a single tube and digested for 2 hours at 37°C. After digestion, we did a standard PCR purification using the Qiagen QIAquick PCR purification kit (Qiagen # 28106) on the 100μL of digestion mixture and eluted in 30μL of buffer EB. After digestion, we conducted size selection using E-Gels (ThermoFisher # G661002) with 25μL of the purified digestion product and selected the band at $\approx$ 350 bp for sequencing.

Multiplexing and amplicon sequencing

We used Qubit HS kits (ThermoFisher # Q-33120) to quantify the concentration of our size-selected product for each sample and mixed them in equimolar ratios into a single sample for high-throughput Illumina sequencing. Our samples were submitted to the Stanford PAN facility (`pan.stanford.edu`) for Bioanalyzer analysis and then sequenced either with NGX Bio `www.ngxbio.com` or at the Stanford Center for Genomics and Personalized Medicine (`scgpm.stanford.edu`) with 2x101 paired end sequencing technology on Illumina HiSeq 2000 machines. Samples were sequenced with 25% phi-X genomic library spike-in (provided by the sequencing facility) to avoid calibration problems due to amplicon sequencing.

Initial processing of the amplicon sequencing data

Our initial processing of the sequencing data included de-multiplexing the sequencing data to separate reads from different samples, removing PCR duplicates, and determining the number of reads in each sample for each barcode. Complete source code can be found at `https://github.com/sunthedeep/BarcodeCounter`.

Briefly, the pipeline uses bowtie2 to identify the sample, pcr duplicate and lineage tag barcode sequences from each read in the FASTQ file. After removing PCR duplicates from the data and demultiplexing the data by sample, we identify all unique sequences in each sample and their number of occurrences using a simple lookup table. We then map all of these unique sequences to the database of 500,000 barcode sequences identified by Levy et al., (2015) using NCBI blastn with parameters ("-outfmt 6 -word_size 12 -evalue 0.0001") to count the number of reads mapping to each of the known 500,000 barcodes in each sample. We account for barcodes known to be in the database with nearly identical sequences by considering such barcode clusters as a single lineage, and provide scripts to identify previously undetected barcode clusters from the sample data. These barcode counts provide the input for our fitness estimation procedure described below.

Whole-genome sequencing

DNA extraction, Library Construction and whole-genome sequencing Clones selected for sequencing were streaked onto either M3 or YPD agar plates from freezer stocks for single colonies. One single colony for each clone was inoculated into either 1mL M3 or YPD (in a 96 deep-well plate) and grown overnight at 30° C without shaking. These cultures were used to perform DNA extractions using either the BioBasic 96 yeast genomic DNA extraction kit (BioBasic # BS8357) or the Zymo YeaStar Genomic DNA kit (Zymo # D2002). Libraries were constructed using Nextera technology with the protocol of Kryazhimskiy et al. (2014). We multiplexed up to 96 libraries per Illumina HiSeq 2000 lane; samples were sequenced at the Stanford Center for Genomics and Personalized Medicine with 2x101 paired end sequencing technology. Libraries that generated less than 5x average genome-wide coverage were removed from further analysis.

Some lineages (defined by unique barcode IDs) were sequenced multiple times, either due to low coverage in one library or due to sequencing multiple independent clones containing the same barcode ID. Variants called from all libraries with the same barcode ID, regardless of origin, were combined together. Importantly, please note that while the libraries were mapped to a non-reference genome which includes the barcode locus sequence, all variants reported in this manuscript both in the main text and the supplemental files have been lifted over to the coordinate system of the Saccharomyces Genome Database (SGD; `www.yeastgenome.org`) R64 Saccharomyces cerevisiae reference genome for convenience.

FASTQ processing, GATK-based variant calling and filtering For each sample, we received two fastq files, one for each read of the paired end sequencing ("forward.fastq" and "reverse.fastq"). We trimmed the first 15 bases and the last 3 bases of each read as well as any adapter sequences using TrimGalore (version 0.3.7 Available at: `http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/`).

```
perl trim_galore -a CTGTCTCTTATACACATCT -a2 CTGTCTCTTATACACATCT --length 50
    ↪  --clip_R1 15 --clip_R2 15 --three_prime_clip_R1 3 --three_prime_clip_R2 3 --
    ↪ paired -o OUTPUTDIR forward.fastq reverse.fastq
```

Reads were mapped using Novoalign (version 3.02.02, Novocraft Technologies) to a modified version of the sacCer3 S288C *S. cerevisiae* reference genome that includes the DNA barcode locus (Levy et al., 2015) in the sequence.

```
novoalign -d referenceGenome.fasta -f forward.trimmed.fastq reverse.trimmed.fastq -l 75 -H
    ↪ 22 -o SAM READGROUPINFO -r Random >library.novoalign.sam
```

The mapped reads were then sorted using PicardTools version 1.105(1632) (Broad Institute, `http://broadinstitute.github.io/picard`)

```
java -Xmx2g -jar SortSam.jar INPUT=library.novoalign.sam OUTPUT=library.novoalign.bam
    ↪ SORT_ORDER=coordinate
```

We used PicardTools again to remove PCR duplicates

```
java -Xmx2g -jar MarkDuplicates.jar ASSUME_SORTED=true REMOVE_DUPLICATES=true
    ↪ INPUT=library.novoalign.bam OUTPUT=library.novoalign.dedup.bam
```

After building an index for the bam file with PicardTools, we then got global coverage metrics with GATK version 3.2.2 (McKenna et. al. 2010) DepthOfCoverage function and per base pair coverage statistics using Bedtools v2.17.0 Quinlan et. al. 2010 genomecov. We plotted the per base pair coverage statistics to identify whole chromosome aneuploidy events.

We genotyped the libraries using GATKs Unified Genotyper.

```
java -jar -Xmx2g GenomeAnalysisTK -T UnifiedGenotyper -R referenceGenome -I library.
    ↪ novoalign.dedup.bam -ploidy 2 --genotype_likelihoods_model BOTH -stand_call_conf
    ↪ 30 -stand_emit_conf 10 -o library.gatk.vcf
```

We initially filtered variants from GATK as follows:

```
java -Xmx2g -jar GenomeAnalysisTK.jar -T VariantFiltration -R referenceGenome --variant
    ↪ library.gatk.vcf --out library.gatk.filtered.vcf --filterExpression ''QD < 10.0 || FS >
    ↪ 20.0 || MQ < 50.0 || MQRankSum < -12.5 || ReadPosRankSum < -8.0 || AN > 10 ||
    ↪ AF<0.25'' --filterName ''my_filter''
```

We also manually filtered all variants called by GATK by remapping them with the CLC genomics workbench and manually validating variants using the resulting read pileups. This procedure eliminated 11% of our GATK variant calls. We then tested the validity of 57 of these filtered variants using Sanger sequencing, and identified no false positive calls. We combined all of the filtered variants across all of the libraries into a single file ("allLibraries.gatk.filtered.vcf") using GATK after lifting over the variant coordinates to the standard UCSC sacCer3 reference genome. We then conducted additional variant filtering using custom scripts, where we removed any mitochondrial variants, variants not passing the GATK filter and variants annotated in reference genome as being in repetitive elements (telomeres, centromeres, replication origins, transposable elements containing "Ty", "delta", "sigma" or "tau" in their name) or low complexity regions defined by the Tandem Repeat Finder (Benson et al., 1999) with the recommended parameters (2 7 7 80 10 50 500 ngs). We also removed all variants with less than 3 reads of support for the derived allele. Heterozygous calls by GATK were validated by first testing whether they had at least 3 reads of support for both ancestral and derived alleles, and passed a binomial filter with $p > 5\%$ for deviation from an equal proportion of ancestral and derived reads. Variants that failed either of these filters were reclassified as homozygous. Heterozygous calls that did pass were then checked to see if they resided in homopolymer repeat regions or in sites with multiple derived alleles across the entire dataset. Such variants were removed from the dataset as likely mapping errors.

As we found that mutations in the nutrient sensing pathway were highly adaptive, we searched the raw variant calls of clones with $s > 5\%$ but no nutrient sensing pathway mutations for filtered variants in this pathway and added them back into our mutation list (the mutations reported in the main text include these variants). This was done for a total of 3 clones (one *IRA1*, one *IRA2* and one *CYR1*).

Copy number variant detection We tested for the presence of copy number variants using a number of software packages, including CNVnator and SVDetect, along with specific manual surveys of the coverage density around the HXT6/7 locus as amplifications of this locus have been shown to be adaptive in previous chemostat laboratory evolution experiments. However, we were unable to detect any copy number events either at this locus or genome-wide with high confidence.

Structural variant detection with CLC-Bio We systematically looked for the existence of structural variation in our sequenced clones, i.e., for the presence of insertions and/or deletions larger than the maximum of 5-10 bp typically detected by our GATK-based variant calling pipeline, as well as chromosomal inversions and translocations. We performed a workflow, described below, utilizing CLC Genomics Workbench version 8.5 (QIAGEN Aarhus A/S; `www.clcbio.com`; API version:850; Build number:20150904114350; Build date:1509041143; Build rev:131279. Platform:Mac OS X 10.10.5; Architecture:x86_64 (64 bit); Processor cores:24; Java version:1.8.0_60 (Oracle Corporation)). Note that we will call the program "CLC Workbench" for brevity. First we imported the Illumina paired end fastq.gz files for each clone into CLC Workbench, using the parameters "paired reads", "remove failed reads", "paired-end (forward-reverse)", minimum distance 25, maximum distance 1000, Illumina pipeline 1.8 and later quality scores.

We then mapped the reads to the unmodified *S. cerevisiae* (strain S288C) reference genome (downloaded from the Saccharomyces Genome Database (SGD; `www.yeastgenome.org`) R64-1-1 and then imported into CLC Workbench). We did not use any masking during the mapping and used the following mapping parameters: mismatch cost 2, lineage gap cost, insertion and deletion costs 3, length fraction 0.5, similarity fraction 0.5, auto-detect paired distances, map randomly for non-specific matches.

Reads were then trimmed by using the "Trim Sequences" function; trimming was done based on quality scores (limit 0.05); ambiguous nucleotides (maximum of 2) were also trimmed. Reads below 15 nucleotides in length were discarded. Any Nextera adapter sequences were trimmed from reads using the following sequence and parameters for trimming: sequence for adapter trimming CTGTCTCTTATACAC, strand "plus", remove adapter, mismatch cost 2, gap cost 3, allow internal matches with minimum score 4, allow end matches with minimum score at end 1.

We then ran the "InDels and Structural Variants" function, using these mapped and trimmed reads, with the parameters "p-value threshold 0.001" and "maximum number of mismatches 3", and saved the "breakpoints" output files in tab-delimited formats. These variants were filtered to remove structural variants with less than 3 reads of support, present in more than 3 strains or closer than 300bp from the ends of each chromosome. The variants were annotated with gene annotations (file SGD_features.tab) from the Saccharomyces Genome Database (`www.yeastgenome.org`).

After structural variant calling was completed, we filtered out structural variants that occurred in previously known repetitive elements annotated in the SGD database (telomeres, centromeres, replication origins, transposable elements containing "Ty", "delta", "sigma" or "tau" in their name) as before.

Determination of Mating Type and Ploidy

Mating type assays Mating type testing was conducted for 960 clones from replicate E2 and 192 clones from replicate E1. Standard Nat+ URA3- tester strains of both MATa and MAT mating types were grown as lawns on YPD agar plates, while the clones with unknown mating type (Nat-, URA3+) were arrayed and grown on independent YPD agar plates. Replica plating was used to transfer the clones with unknown mating type onto Nat+ Ura- SC-agar plates along with one of the tester strains. The presence of a colony on this plate was used to determine successful mating.

Propidium Iodide and Flow Cytometry Ploidy was initially tested using a simplified propidium iodide staining protocol designed for high throughput analysis, inspired by Cousin et al. (2009). Clones were grown to saturation in YPD liquid media in 96 well plate format. 200 mL of saturated culture was transferred to 96 well filter plates (Pall Life Sciences # 8039) and spun down to remove the spent media. These spun down cells were resuspended in 200 μL 70% ethanol in the filter plates and allowed to fix for at least 1 hour at room temp. Plates were then centrifuged again to remove the ethanol. Cells were resuspended in 50 μL RNAse A buffer (1mg/mL RNAse A in PBS) and incubated at 37° C for at least 6 hours (at most 18 hours). Treated cells were diluted 1:100 into 200 L of propidium iodide staining solution (50 μg/ml PI, 50 μM sodium citrate) and analyzed along with standards of known ploidy using the BD LSR II with an HTS attachment at the Stanford Shared FACS Facility (NIH grant # S10RR027431-01 for UV LSRII). We note that the filter plates can be re-used for ploidy analysis by thoroughly washing them with distilled water using a multichannel pipette.

High throughput benomyl assay A simpler high throughput ploidy test was developed using the drug benomyl. Clones were grown from frozen stock in 1mL liquid YPD in 96 well plates until saturation at 30° C without shaking. The saturated cultures were mixed by multichannel pipette, pinned onto YPD+20μg/mL benomyl (in DMSO) and YPD+DMSO (control) rectangular agar plates using a multi-pronged pinner, grown at 25° C for 48 hours, and then imaged. Under these conditions, diploid growth is strongly inhibited by benomyl but haploid growth is less affected.

Construction of gene deletions in the Ras-cAMP-PKA pathway

Gene deletions were constructed using standard yeast transformation methods to replace the gene of interest with a selectable marker cassette. *IRA1*, *IRA2*, *GPB1*, *GPB2*, *PDE1*, *PDE2* and the pseudogene control YFR059C were individually replaced with a selectable NatMX (nourseothricin) resistance marker in neutral barcoded yeast strains. For each target gene, the resistance marker was amplified from the pBAR1 plasmid (Levy et al., 2015) with primers flanked by 45 bp of sequence adjacent to each end of the appropriate yeast gene. Transformations were performed to delete the gene of interest using the lithium acetate based protocol of Gietz and Woods (2002). Each transformant was verified with gene-specific PCR reactions spanning both the 5 and 3 insertion breakpoints. We assayed the fitness of each deletion using pairwise competition assays described in Levy et al. (2015) and in the main text methods.

Primers for gene deletions:

IRA1 5' CTTCAGCATATAACATACAACAAGATTAAGGCTCTTTCTAAAATGTGGAGGCCCAGAATACCCTCC

IRA1 3' AAGGAAAAACGTATATAATCACTGCAATACTCTAATTTAAAATTATCGACACTGGATGGCGGC

IRA2 5' TATCAACTAAACTGTATACATTATCTTTCTTCAGGGAGAAGCATGTGGAGGCCCAGAATACCCTCC

IRA2 3' AGATAGATATTGATATTTCTTTCATTAGTTTATGTAACACCTCTATCGACACTGGATGGCGGC

GPB1 5' CGGCTACTTTAAGGCTTTCCGTACCAATTCTTCTACATAAGAATGTGGAGGCCCAGAATACCCTCC

GPB1 3' AATTTTCTCGTTTTCCTTTAGTCACTCTTGTCACATAAGGATTATTCGACACTGGATGGCGGC

GPB2 5' GATTCATTGGCAGGTCCATTGTCGCATTACTAAATCATAGGCATGTGGAGGCCCAGAATACCCTCC

GPB2 3' CTAAACAAAGTTTACAAAGTGAAAGCATTGAAAACTGCCTTTTTATCGACACTGGATGGCGGC

PDE1 5' GGTTCTTCTTCTTCATCCCCTTTTTTACCAATATTCCTTTTTATGTGGAGGCCCAGAATACCCTCC

PDE1 3' TAATGGAAAGAAGTTTCATTAGTTACTACTAGTATTTTGCTTGCTTCGACACTGGATGGCGGC

PDE2 5' GAGATCACTACTACTTAATTGAAGAAAACATAACCTATTGATATGTGGAGGCCCAGAAT-

ACCCTCC

PDE2 3' ATGTTTATACAATGAATGGTACAAGAAATTTTGATATTCTTGCTATCGACACTGGATGGCGGC

YHR095W 5' CCATCAAATGTCGCAGCAGCTCATGTTTACGTTTGCTGTCTTCTGTGGAGGCCCA-GAATACCCTCC

YHR095W 3' AATAAGCCCTAGAAACCTTACACCCTAATTTGCACAAGAAAACTATCGACACTGGATG-GCGGC

QUANTIFICATION AND STATISTICAL ANALYSIS

Statistical parameters including the exact values of n, precision measures (mean ± SEM) and statistical significance for various statistical tests are reported in the Main Text, Figures and the Figure Legends.

Introduction to fitness estimation methodology

The fitness of a barcoded lineage relative to the rest of the population determines how quickly it grows. If the number of cells in a lineage is large at the bottleneck, then during the $T = 8$ generations from cycle i to cycle $i + 1$ the bottleneck population, n, grows close to deterministically:

$$n_{i+1} \approx n_i e^{(s-\mu_i)T} \tag{1}$$

with μ_i the mean fitness of the population at time i. The time-dependent mean fitness cannot be measured directly, but the size of the total barcoded population that is neutral with respect to the ancestor, ρ_i, gives μ_i from ρ_{i+1}/ρ_i as s is the fitness of the barcoded lineage relative to these neutral lineages.

The sequencing measurements give estimates of the relative sizes of a barcoded lineage from the numbers of reads, r_i, of the barcode at successive time points as a fraction of the total reads, R_i. Comparing with the number of reads of the neutral barcodes, ρ_i, the fitness over cycle i is estimated by

$$\begin{aligned} \hat{s}_i &= \frac{1}{T}\left[\ln\left(r_{i+1}/R_{i+1}\right) - \ln\left(r_i/R_i\right)\right] + \mu_i \\ &= \frac{1}{T}\left[\ln\left(r_{i+1}/r_i\right) - \ln\left(\rho_{i+1}/\rho_i\right)\right] \end{aligned} \tag{2}$$

However, there are several sources of deviations of such estimates from the actual fitness. The experiments themselves contribute biological stochasticity in the growth and division of cells, sampling during the dilution at the end of each cycle, and subtle variability in conditions. The measurement process contributes counting noise from sequencing as well as potential variabilities and biases in DNA extraction and PCR amplification.

The biological noise, dilution sampling, and sequencing counting noise should all have variance proportional to the mean numbers of cells and/or reads. We find that for typically sized barcode lineages (~ 100 reads), deviations from deterministic trajectories scale as the square root of the number of reads, i.e.

$$\mathrm{Var}(r_i) \approx \kappa_i \langle r_i \rangle \tag{3}$$

where r_i is the number of reads at time i, $\langle r_i \rangle$ is the expected number of reads, and κ_i is a noise parameter inferred from the data which depends on the cycle, the replicate and the batch. Furthermore, we show that

for the collection of neutral lineages, the distributions of changes in read numbers from one cycle to the next are close to normal.

For large lineages ($> 10^3$ reads), however, the data exhibit larger than expected variations which do not decrease with numbers of reads. The sources of these variations are currently unknown. They set a limit of $\gtrsim 1\%$ per generation on the resolution of our fitness assay.

We use the data to crudely fit a multiplicative noise parameter α_i at each cycle in addition to the normal variance. For the fitness inferred over one cycle,

$$\hat{s_i} = \langle s_i | r_i,\, r_{i+1} \rangle \tag{4}$$

the variance is then roughly of the form:

$$\mathrm{Var}(s_i) = \frac{1}{T^2}\left(\frac{\kappa_i}{r_{i+1}} + \alpha_i^2\right) \tag{5}$$

To infer fitnesses, we use a model assuming Gaussian additive noise at low frequency and multiplicative noise at high frequency to combine the results from across the cycles, replicates and batches, weighted by the inverse variances.

In the next we further elucidate the fitness estimation process and break down the contributions to κ_i. We then carry out self-consistency checks and justify our noise model. Finally, we present the results of the fitness assay broken down by batch and replicate, and further discuss the hypothesis testing done in the main text.

Noise model

Read stochasticity From the dynamics of the numbers of cells in a lineage, we expect that the mean number of reads at time $i+1$ will be

$$\langle r_{i+1} \rangle = \frac{R_{i+1}}{R_i} r_i e^{(s-\mu_i)T} \tag{6}$$

and thus dependent on the total numbers of reads, R_i and R_{i+1}.

The stochasticity in the population dynamics and the counting variations from the sequencing both give additive noise so that we expect

$$\mathrm{Var}(r_{i+1}) \approx \kappa_i \langle r_{i+1} \rangle \tag{7}$$

where κ_i is a parameter fit from the data for many barcode families. For nearly neutral lineages, there is only a weak dependance of κ_i on s. Lineages with large s quickly reach a size where the additive model breaks down; we will use our multiplicative noise model to analyze their fluctuations.

The contributions to κ_i depend on the parameters of the particular measurement: the total number of barcoded cells at the bottleneck, $\mathcal{N}_i^B$, and the number of reads, R_i both of which vary considerably. The average number of reads per barcoded cell, the *coverage ratio*

$$C_i \equiv \frac{R_i}{\mathcal{N}_i^B} \tag{8}$$

strongly affects the noise magnitude: when the coverage ratio is low, read noise dominates; when the coverage ratio is high, the biological noise dominates. The contributions to the noise parameter are

$$\kappa_i = \underbrace{1}_{\text{Read noise at } i+1} + \underbrace{C_{i+1}/C_i}_{\text{Read noise at } i} + \underbrace{C_{i+1}(\beta_i+1)}_{\text{growth+dilution}} + \underbrace{\xi_i(1+C_{i+1}/C_i)}_{\text{Extraction/PCR noise}} \tag{9}$$

The first term comes from Poisson read noise at time $i+1$ and the second term from the Poisson read noise at time i, scaled from the coverage at i to the coverage at $i+1$. The third comes from the stochasticity in the growth of the cells. For a single cell at a bottleneck, the number of descendants at the end of the cycle averages $2^T e^{sT}$ with a variance of $\beta_i 2^{2T}$. Almost all this variability is likely to come from the earliest stages of the cycle as when the number of descendants becomes large, the fluctuations are averaged over. After dilution the biological stochasticity contributes β_i per cycle to the variance. In addition, there is a factor of 1 that comes from the Poisson dilution at the end of each cycle. The last term ξ_i accounts for the unknown additive parts of the effects of DNA extraction and PCR amplification.

We assume that the variations are Gaussian in nature. This assumption was inspired by the additive nature of the noise sources, and describe the data well. The assumption breaks down when r_i is low or when $C_{i+1}\beta_i$ is large, since the biological noise is likely to be non-Gaussian.

Number of mutants and coverage ratios In order to understand the balance between read noise and biological noise, we need to know the coverage ratio, C_i, at each timepoint. We know the total number of reads, R_i, at each timepoint; however, we do not have a direct measurement of the total number of barcoded cells, $\mathcal{N}_i^B$, at the bottleneck of each cycle. Since the total population saturates at a size that is roughly independent of the admixture of mutants and ancestral types, after dilution the total bottleneck population, $\mathcal{N}$, is roughly constant.

The barcoded portion can be inferred by noting that two portions of the total population, the unbarcoded ancestral cells with population $\mathcal{N}_i^U$, and the barcoded types that are neutral relative to the ancestor at time i, have the same fitness. Given f_i^n the fraction of the barcoded cells without adaptive mutations, the ratio of the neutral population sizes, $f_i^n \mathcal{N}_i^B/\mathcal{N}_i^U$ is thus constant.

If we know $f_i^n \mathcal{N}_i^B/\mathcal{N}_i^U$ at one timepoint, and f_i^n at all other timepoints, we can solve for $\mathcal{N}_i^B/\mathcal{N}_i^U$. We can then use $\mathcal{N}_i^B/\mathcal{N}_i^U$ and $\mathcal{N}_i^U = \mathcal{N} - \mathcal{N}_i^B$ to approximate $\mathcal{N}_i^B$ at each timepoint. Let $f_i^n \mathcal{N}_i^B/\mathcal{N}_i^U = q$. Then we have

$$\begin{aligned} \frac{f_i^n \mathcal{N}_i^B}{\mathcal{N} - \mathcal{N}_i^B} &= q \\ \mathcal{N}_i^B &= \frac{q}{q + f_i^n}\mathcal{N} \end{aligned} \tag{10}$$

Initially, the fraction barcoded is formulated to be $\mathcal{N}_0^B \cong 10\%\mathcal{N}$. This gives allows us to calculate q (≈ 0.03, similar across batches). Using the sequencing reads, we can obtain f_i^n at each timepoint and hence calculate an estimate for $\mathcal{N}_i^B$ at each timepoint.

Figure S1 shows this estimate of the barcoded fraction of the population. At late times, a considerable fraction of the population is barcoded, and a significant fraction of this barcoded population has adaptive mutations. The barcoded fraction increased rapidly as a consequence of its original diversity. Roughly 50% of the barcoded cells had fitness $> 6\%$ at the first timepoint, and 25% are diploid. The rest were nearly neutral haploids. By the end of the experiment, $> 90\%$ of barcoded cells were high fitness mutants. As barcoded

fraction of the pool increased, the read depth remained nearly constant. The coverage ratios decreased as a function of time. They started at around 0.3-0.5, but fell to about 0.04-0.07 by the end of the experiment. We will see that means by late times, the errors in fitness estimation are dominated by the read noise.

The beneficial mutants showed significant transient behavior in the first growth dilution cycle. In batches 1, 3, and 4, the barcoded fraction (and therefore the beneficial mutants) did not increase appreciably in the first cycle. Batch 2 was grown for one growth/dilution cycle before timepoint 1 and does not display transient behavior in its first cycle.

To remove the effect of transient behavior on the fitness assay, we used the sequencing data from timepoints 2-5 for batches 1, 3, and 4, and timepoints 1-4 for batch 2. The trajectories of the barcoded fractions are very similar across batches for the timepoints chosen, and avoids the latest timepoint in batch 2 where the barcoded types have nearly taken over the population.

While κ_i increases with increasing coverage ratio, the variance in the fitness estimate decreases with increasing read depth. For large coverage ratios (C_{i+1} & $C_i >> 1$), the variance reaches the minimal value

$$\mathrm{Var}(s_i) \approx \frac{1}{T^2}\frac{(1+\beta_i)}{n_i} \tag{11}$$

where n_i is the number of cells in the barcode family at the bottleneck at timepoint i. In this regime the noise is dominated by biological fluctuations. This sets a noise floor for the measurement. For our measurements, only the first cycle (which was not included in the analysis) was near this regime.

Inferences of s In addition to the additive sources, there appears to be a roughly frequency independent component of the noise. The source of this noise is unknown. For simplicity, as it does not affect much the results, we parametrize this by a multiplicative gaussian noise parameter α_i, fit within each batch for every pair of timepoints. We find that $\alpha_i \approx 0.1$/ cycle, largely independent of the cycle, replicate and batch. Then the assumed variance of our estimator is

$$\mathrm{Var}(s_i) = \frac{1}{T^2}\left(\frac{\kappa_i}{\langle r_{i+1}\rangle} + \alpha_i^2\right) \tag{12}$$

The fitness estimation algorithm proceeds in the following manner:

1. Identify lineages which are neutral relative to the ancestor for each replicate and batch individually.
2. Use the *collection* of these neutral lineages to estimate κ_i and μ_i.
3. Estimate α_i for each batch and timepoint from lineages with a large number of reads.

We then carry out the follow steps for **each barcode separately**:

1. Use formulae for $\hat{s}_i$ and $\mathrm{Var}(s_i)$ (Equations 2 and 12) to calculate fitness and error at each timepoint.
2. Average over timepoints, replicates, and batches, using inverse variance weighting by errors, to get an overall estimate of the fitness $\bar{s}$ of that barcode .

We give a more detailed account in the next two sections.

Checks on the noise model

We made a number of self-consistency checks to test the applicability of the simple additive noise model for lineages at low read depth. We analyzed the following quantities:

- Distributions of within-replicate variations
- Scaling with read numbers of between replicate variations
- Comparison of within replicate to between replicate variations

Our analysis suggests that there is good agreement between within-replicate variation and between-replicate variation for moderately sized ($\sim$ 100 reads) lineages. At late timepoints, the noise is dominated by the counting noise of sequencing. We show that this is due to the expansion of the barcoded lineages. We also discuss the frequency-independent deviations for large ($\sim$ 1000 reads) lineages, which limits the sensitivity of fitness assay to 10%/cycle (1.2%/generation).

Estimating κ within replicate By considering the dynamics of large groups of lineages with identical fitness together, we can test the noise model. The large set of lineages neutral relative to the ancestor ($\sim$ 1500) enables estimation of the noise parameter κ_i, with good enough statistics on the noise to test its normality. It also lets us infer the time-dependence of the mean fitness, μ_i, which is needed to obtain the fitness of the other lineages. We assume that the neutral lineages are virtually identical in both fitness and the magnitude of their biological fluctuations.

The model assumes that the deviations of the read numbers $r_{i+1} - \langle r_{i+1} \rangle$ are distributed as $\mathcal{N}(0, \kappa_i \langle r_{i+1} \rangle)$ for r_{i+1} large enough. We define the normalized differences Z_i as follows:

$$Z_i = \frac{r_{i+1} - \langle r_{i+1} \rangle}{\sqrt{\langle r_{i+1} \rangle}} \tag{13}$$

Given a collection of lineages with identical phenotype, the Z_i are identically distributed as $\mathcal{N}(0, \kappa_i)$. The total frequency of a phenotype f_i can be used to estimate μ_i. The distribution of scaled deviations can be used to find κ_i and test the noise model.

We carried out the following procedure using the $\sim$ 1500 lineages which were neutral in the experiments of Levy et al. (2015)

- Estimate mean fitness by $\mu_i = \log(f_{i+1}/f_i)/T$.
- Plot distribution of Z_i. Remove outliers (likely adaptive or outside regime of additive noise)
- Re-estimate μ_i. Re-plot Z_i.
- Set $\kappa_i = \mathrm{Var}(Z_i)$.

The first two rows of Figure S2 shows the distributions of the scaled deviations for each pair of adjacent timepoints in batch 1, replicate 1. The top row shows the histograms of Z_i from the neutral haploids as a function of time. The same analysis was carried out on the 1600 diploid lineages: the results are shown in the second row and are very consistent with the haploid inferences.

The normal distribution predicted from theory is plotted in red over the empirical distribution histogram. The counting noise limit is plotted in black. The noise starts off larger than the read counting noise limit, but is dominated by counting noise at the end. This is in concordance with our observation that the coverage ratio decreases at late times, and suggests that the extraction and amplification parts of the noise, ξ_i — which would be expected also to scale with the read depth — is a small fraction of κ_i.

The model fits quite well over the range of 1-2 standard deviations. Only 10-20 lineages were removed from each plot as outliers; most were a few standard deviations from the mean, and clearly adaptive. The normal fit is worst at the earliest timepoints, when the cells are first experiencing the evolution condition, and the latest timepoint, where the number of reads is smaller. Our analysis also showed that the noise parameter inferred from lineages of different sizes did not vary significantly when different sized lineages were used to infer it (from 30 reads up to 150). The diploids behaved similarly to the haploids in all these aspects, including the number of outliers.

The values of κ_i tend to start around 10, and drop down to around 2 at late times. The values vary between replicates and experiments as can be expected by the different coverage ratios. Two replicates have very low coverage at one timepoint (batch 2 replicate 2 and batch 3 replicate 3), which decreases the quality of fitness inferences for those datasets. The analysis for the diploids found similar κ_i values. The diploid κ_i tended to be slightly higher than the haploid κ_i: by 5-10% at early times, nearly identical at late times.

The fact that the fluctuations are dominated by counting noise at the end of the experiment suggest that ξ_i is small. If we set ξ_i to 0, and use the C_i estimated previously, we can calculate the biological noise parameters β_i using Equation 9. We get values β_i in the 12-17 range at early times and in the 5-10 range at late times.

Part of the difficulty estimating β_i comes from the fact that coverage ratios are low ($\sim 0.1 - 0.2$). Therefore, errors in estimation of order 0.5 (from the mean fitness estimate, coverage ratio, and ξ_i) propagate up to errors in β_i of order 2.5-5 at late times. More detailed analysis and measurements would need to be conducted to yield a more quantitative estimate for β_i and its uncertainty.

In previous experiments, the estimated values of β_i started off low (around 4), but reached values as high as 15 at later times when there were more mutants in the population. Since our experiments start off with a relatively high mutant fraction, our results are at least roughly consistent with previous work. Large values of β_i suggest that there is high variability when the populations are low: i.e. variations in viability (surviving stationary phase), lag phase (time delay to start growth after dilution), and in the first rounds of division.

Replicate-replicate correlation As an independent test of the consistency of the noise model, we examined the correlation between replicates in the same batch and compared to the inferred within-replicate noise parameters κ_i. Specifically, we looked at the sample standard deviation of the log slope $\sigma_i = \ln(r_{i+1}/R_{i+1}) - \ln(r_i/R_i)$. The log slope was chosen since its variance can be shown to be

$$\mathrm{Var}(\sigma_i) \approx \frac{\kappa_i}{r_{i+1}} \tag{14}$$

if our additive noise model holds with our definition of κ_i.

The final row of Figure S2 shows the sample standard deviation $\delta\sigma_i$ of the log slopes plotted against the number of reads at the second timepoint of the cycle. The plots show the $r_{i+1}^{-1/2}$ scaling as expected for a wide range of reads.

We can use the distribution of $\delta\sigma_i$ to fit a κ parameter, and compare it to the expected value. The 3 curves with the scaling $r_{i+1}^{-1/2}$ show 3 different fits. The within replicate κ from the variance of Z_i is shown in red. In blue is the inference $\hat{\kappa} = E[\sqrt{r_{i+1}}\delta\sigma_i]^2$ (between replicate estimate). Black is the theoretical minimum value that the noise parameter could take if there was only read noise ($\beta_i = 0$).

For the first pair of timepoints, κ from within a replicate is larger by $\frac{C_{i+1}}{C_i}$ compared to the between replicate κ. This is expected since the first measurement is common for all replicates in a single batch (see **Method Details**). By late times, both estimates of κ are close to the being pure read noise.

Multiplicative noise regime For each batch and timepoint, we roughly fit a frequency independent part of the noise by averaging $\delta\sigma_i$ at high ($\sim 10^3$) read number (green line). We then modify the noise parameter κ_i to be

$$\tilde{\kappa}_i = \kappa_i + \alpha_i^2 r_{i+1} \tag{15}$$

The multiplicative noise varies little across timepoints and experiments, typically $\sim 10\%$/cycle corresponding to uncertainties in estimates of s of $\gtrsim 1\%$ per generation. For typical values of κ_i, the crossover between multiplicative noise and read noise occurs at $\sim 10^3$ reads. The multiplicative noise constant increases with time in batch 2, which is dominated by mutants at late times.

Fitness assay

Basic procedure We used the fitness assay outlined earlier to calculate fitnesses and error estimates for each lineage, between every pair of timepoints, replicates and batches. For each replicate, we combined estimates across timepoints into an overall fitness estimate $\bar{s}$ by an inverse variance weighted sum:

$$\bar{s} = \left(\sum_i \hat{s}_i/\mathrm{Var}(s_i)\right) / \left(\sum_i 1/\mathrm{Var}(s_i)\right) \tag{16}$$

$$\mathrm{Var}(\bar{s}) = \left(\sum_i 1/\mathrm{Var}(s_i)\right)^{-1} \tag{17}$$

This method gives the correct weighting for the max posterior estimate of the mean of a collection of gaussian random variables with equal means and unequal variances.

We averaged over timepoints within a replicate to obtain the fitness values reported for each replicate. We averaged once more across replicates and batches to obtain the fitness values reported in the main text.

Distribution of fitness effects Figure S3 shows the distributions of fitness effects from all the replicates and batches. The two colors correspond to the haploids and the diploids respectively. As can be seen, almost all the lineages are either very close to neutral relative to the ancestor, or diploid. Both the neutral haploids and diploids tend to have low coverage at late times, which gives broad peaks (typically in the 2% range) compared to fitness differences ($1.5\% - 3.5\%$). Due to systematic variation between batches, the diploid and main haploid peaks are well resolved only for some batches.

Replicate-replicate fitness correlation To test the data against our noise model we examined the replicate-replicate fitness correlation. High fitness lineages tend to have lower errors due to higher read counts. Their errors are dominated by the multiplicative noise. We can see that the errors inferred via the noise model are very similar to the observed variation between replicates.

Figure S4 shows all of the replicate-replicate correlations. The scale of the inferred error bars is consistent with the scale of the differences in fitness between replicates, but systematic differences are clearly noticeable. Batch 1 shows good correlation across all replicates. Batches 2 and 3 show systematic deviations of both the diploid and high fitness lineages. Some of the differences in batch 2 explained by the low coverage timepoints in batch 2, replicate 2. The low coverage leads the inference to be dominated by a single slope. Batch 3 replicate 3 looks systematically different from the other replicates in batch even at high coverage timepoints.

Cross-batch fitness correlation We next examine the correlations between batches, and found that these were worse than within-batch correlations. Figure S5 shows fitness-fitness correlations between the best replicates in each batch. (This is in contrast with Figure 2 in the main text, which compares the averages over each batch.)

Systematic differences between batches for specific mutation classes Both the diploid and high fitness lineages exhibited systematic differences across batches. While the between replicate deviations were in the $1-2\%$/generation range, the between batch differences were as high as 5%/generation.

The last panel in Figure S5 compares the fitnesses across replicates and batches for the *GPB2*, *PDE2*, and diploid classes. Estimated error bars from the fitness assay are plotted. The fitnesses within a batch correlate well, with most deviations occurring in replicates with low coverage timepoints. The overall batch-batch systematics are different for different types of mutations. For most pairs of classes, the relative ordering does not change. However, some like *PDE2* and *GPB2* switch order in the different batches. These differences suggest that the systematic deviations are not merely an artifact of the fitness estimation algorithm and thus cannot be consistently corrected for statistically.

Within the best replicates, there is a very narrow spread of all but one of the *GPB2* mutant lineages, and all but one of the *PDE2* lineages. This suggests that the intrinsic precision and potential accuracy of the barcode fitness assay is $\lesssim 1\%$.

Testing for differences in fitness effect between mutant classes The systematic cross-batch differences informed how we tested for differences in fitness effects between different mutation classes. We first carried out a number of ANOVA tests, for differences between genes, mutation types, and paralogs.

To test if gene identity was at all significant, we treated the batch as a categorical variable and still ended up with a $P < 10^{-16}$. For our tests of fitness difference of particular pairs, we carried out tests separately for each batch. We averaged over all timepoints and replicates within a batch to get a single fitness per lineage.

Within each batch we tested the hypothesis that the means of the fitness distributions for different classes were different. We found significant differences in fitness between *IRA1* and *GPB1*, *IRA1* and *IRA2*, *GPB1* and *GPB2*, and found that mutation type made a significant difference in the fitness of *IRA1* mutants (test was not significant for mutation type in *IRA2*). For the diploids, we found that a third copy of chromosome 11 gave significant fitness benefit, but a third copy of chromosome 12 did not. Additional coding mutations

did not significantly change the fitness of diploids.

We also tested the null hypothesis that the distribution of *IRA1* mutants was the same as the distributions of *PDE2* and *GPB2* mutants. We used the non-parametric Kolmogorov-Smirnov (KS) test to test for any difference between distributions. The KS test compares the CDFs of two empirical distributions, and compares the largest gap between them (which is distributed in a way independent of distribution). The fact that the data pass the KS tests as well give us confidence that our results are not due to noise-modeling assumptions.

The results are also robust to changes in the fitness inference algorithm. If we instead use a weighted log-linear regression, choosing $\hat{s}$ by

$$\hat{s} = \arg\min_s \sum_i \frac{(\log(r_{i+1}/R_{i+1}) - \log(r_i/R_i) + \mu_i - s)^2}{\kappa_i/r_{i+1}} \tag{18}$$

the fitness estimates change by $\sim 1\%$ per generation at most. The differences in distributions and relative orderings of fitnesses persist.

DATA AND SOFTWARE AVAILABILITY

All Illumina sequencing data (for both the whole-genome sequencing and the fitness measurement assays) can be found under NIH BioProject PRJNA310010. The software repository for the barcode counting code can be found at `https://github.com/sunthedeep/BarcodeCounter`