Chapter 22

Bulk-Fitness Measurements Using Barcode Sequencing Analysis in Yeast

Claire A. Chochinov and Alex N. Nguyen Ba

Abstract

The use of DNA barcodes for determining changes in genotype frequencies has been instrumental to increase the scale at which we can phenotype strain libraries by using next-generation sequencing technologies. Here, we describe the determination of strain fitness for thousands of yeast strains simultaneously in a single assay using recent innovations that increase the precision of these measurements, such as the inclusion of unique-molecular identifiers (UMIs) and purification by solid-phase reverse immobilization (SPRI) beads.

Key words DNA barcodes, Bulk fitness assays, Strain fitness determination, Unique molecular identifiers

1 Introduction

Characterizing strain fitness is a crucial aspect of genetic studies and evolution. Methods to quantify relative fitness have traditionally relied on crude measurements such as side-by-side spot dilution assays [1] and growth-curve fitting [2]. More recently, methods involving the mixing of the two strains of interest [3] (competitive fitness assays) have become popular for their ease and flexibility. These techniques rely on measuring the relative change in frequency of the strains over time by using benign markers such as fluorescence [4] which can accurately track the frequency of each strain by flow cytometry. Although these methods all produce similar results [5], the advantage of competitive fitness assays is that it is amenable to parallelization using DNA barcoding technologies [6, 7]. By transforming strains of interests with random nucleotide barcodes, a *bulk* competitive fitness assay can be performed by DNA sequencing (Fig. 1). This procedure is inexpensive

Supplementary Information The online version contains supplementary material available at [https://doi.org/10.1007/978-1-0716-2257-5_22].

Frédéric Devaux (ed.), *Yeast Functional Genomics: Methods and Protocols*,
Methods in Molecular Biology, vol. 2477, https://doi.org/10.1007/978-1-0716-2257-5_22,

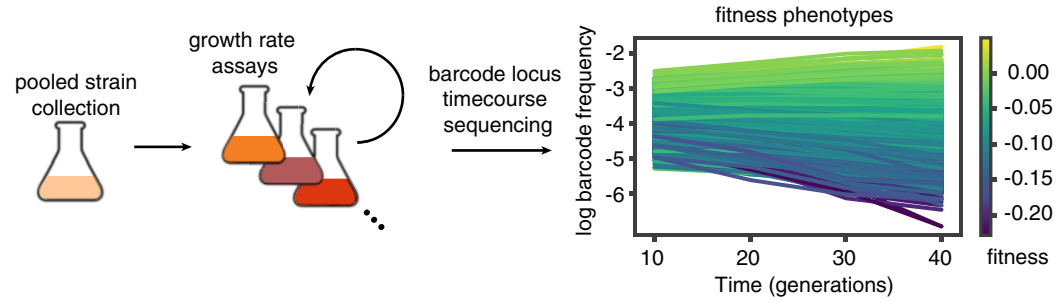

Fig. 1 Schematic of the experiment. Pooled barcoded strains are grown in bulk in a desired environment and passaged for a few generations by repeated serial dilution. Barcode locus sequencing at each passage is used to obtain strain frequencies and infer fitness

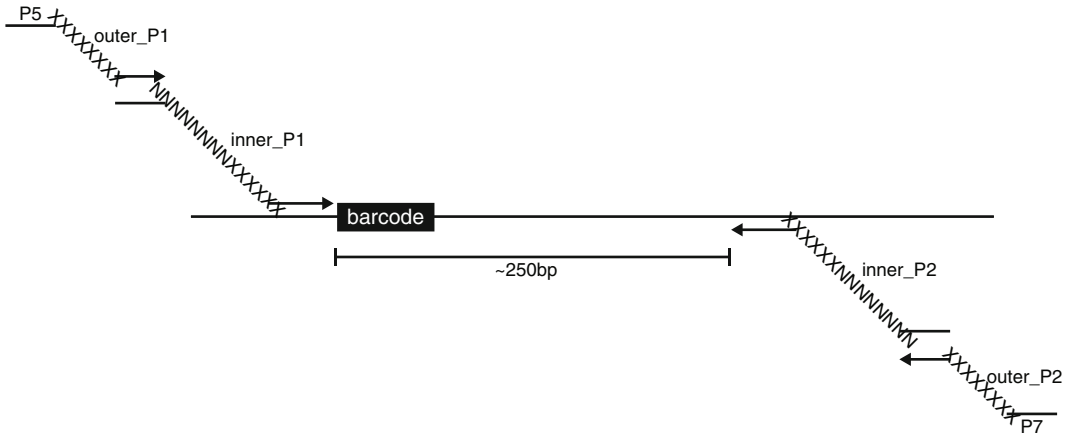

Fig. 2 Schematic of the amplification reaction. Two-stage PCR is used to amplify the barcodes from pooled strains. The first stage, using inner_P1 and inner_P2 primers is a limited cycling reaction, which adds unique-molecular identifiers (UMIs) to amplification products. Following purification by solid-phase reversible immobilization (SPRI beads), a second amplification using outer_P1 and outer_P2 is used to attach Illumina adapter sequences

(at the time of writing, the cost of these pooled competitive fitness measurements can be as low as 0.2 cent per strain, per environment) and efficient for large panels of strains: we routinely perform these assays for >100,000 strains simultaneously in multiple conditions with replicates, with each set of assays taking at most 1 h per day over a week. At this scale, we obtain measurements with errors in competitive fitness of approximately 0.1–1%, which is similar to what has been observed in pairwise assays using fluorescence.

We describe here an experimental protocol and analysis for these bulk competitive fitness assays using barcoded yeast, which we chose due to its long history in systematic genetics studies [8]. We incorporate several innovations to obtain more accurate measurements, including a refinement on the enzymes and reagents used for the procedure and the introduction of unique-molecular identifiers to prevent PCR exponential noise [9] (Fig. 2). This

protocol can be easily adapted to different species, by changing the growth media and genomic DNA extraction protocol. The experimental scale can also be modified, for example, by performing these growth assays and genomic extractions in microtiter plates. Finally, any phenotype that can be measured by the relative change in frequency of the strains of interest can be assayed using this method.

2 Materials

Prepare all solutions using molecular biology grade water. All reagents can be stored at room temperature unless noted.

2.1 Yeast Genomic DNA Extraction Using Silica Mini-Preparative Columns

1. Yeast lysis buffer: 1 M Sorbitol, 100 mM Sodium Phosphate buffer, pH 7.4, 10 mM EDTA, 0.5% SB3-14 (3-(*N*,*N*-dimethylmyristylammonio)propanesulfonate), 200 μg/mL pancreatic Rnase A, and 20 mM DTT, and 5 mg/mL Zymolyase 20T. Made in 15-mL falcon tube aliquots by mixing the non-enzymatic components first, and stored in a −20 °C freezer after mixing. The reagent is thawed at room temperature and mixed by vortex before use (*see* **Notes 1**–**3**).
2. Binding buffer: 100 mM MES, pH 5, 4.125 M Guanidine thiocyanate, 25% isopropanol, 10 mM EDTA (*see* **Notes 4** and **5**).
3. Wash buffer 1: 0.85 M Guanidine thiocyanate, 25% isopropanol, 10 mM EDTA.
4. Wash buffer 2: 80% Ethanol, 10 mM Tris–HCl, pH 8 (*see* **Note 6**).
5. Elution buffer: 10 mM Tris–HCl, pH 8.5.
6. Silica mini-preparative columns and collection tubes (*see* **Note 7**).

2.2 Two-Step PCR

1. Q5 polymerase kit: containing 5× Q5 polymerase buffer, and Q5 polymerase enzyme. Stored at −20 °C, thawed on ice before use.
2. KAPA HiFi PCR kit: containing 5× KAPA HiFi polymerase buffer, 10 mM dNTP mix, and KAPA HiFi polymerase enzyme. Stored at −20 °C, thawed on ice before use (*see* **Note 8**).
3. 10 mM dNTP mixture. Stored at −20 °C, thawed on ice before use.
4. AMPure XP beads or equivalent. Stored at 4 °C, warmed to room temperature, and vortexed before use.

Table 1
List of primers

Primers for first-stage PCR (containing UMIs and optionally an index)	
Inner_P1	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG NNNNNNNN XXXXXX YYYYYYYYYYYYYYYYYYY
Inner_P2	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG NNNNNNNN XXXXXX YYYYYYYYYYYYYYYYYYY
Primers for second-stage PCR (containing optionally an index)	
Outer_P1 (with Illumina P5)	AATGATACGGCGACCACCGAGATCTACAC XXXXXXXX TCGTCGGCAGCGTC
Outer_P2 (with Illumina P7)	CAAGCAGAAGACGGCATACGAGAT XXXXXXXX GTCTCGTGGGCTCGG

Y: nucleotides representing the priming sequence
X: nucleotides representing an optional index
N: nucleotides representing random unique molecular identifiers

5. Two sets of primers at 10 μM: one priming to the genome flanking the barcode locus and containing the unique-molecular identifiers, a DNA sequence handle and optionally a multiplexing index, and another set priming on the DNA sequence handles that also contain the Illumina adapter sequences that bind to the flowcell, and optionally a multiplexing index. The unique-molecular identifiers should be random nucleotides of equal base frequency and the total length should be about 16 bp (or 8 bp on each primers). We recommend multiplexing indexes of different lengths (6–11 bp), but many options are possible (*see* Table 1 and **Note 9**).
6. AMPure XP beads washing buffer: 85% Ethanol, 10 mM Tris–HCl, pH 8.

2.3 Growth Media

1. YPD liquid media: 20 g/L peptone, 10 g/L yeast extract, 20 g/L glucose. Autoclave first the peptone and yeast extract mix in 90% of the final desired volume and then add sterile glucose concentrated solution and sterile water up to the final volume.
2. Growth environment media: media should be prepared similarly to YPD and be free of particulate (*see* **Note 10**).

2.4 Softwares

1. Python version 3.6 or more recent.
2. Python module Regex [10].
3. gcc version 6 or more recent.
4. tsl robin_map hash table package [11].

3 Methods

3.1 Growth Assays

1. Barcoded yeast cells form a population that can be used to inoculate fresh YPD media to define a start culture. In general, the number of cells in YPD at saturation is approximately 10^8 cells/mL, and this number is used as a reference for seeding this start culture. Inoculate at least 1000 times more cells than the number of barcodes in the populations and grow for about 5–7 generations. For example, if the number of genotypes whose fitness is being measured is 10^4, then ten million cells (100 μL) can be inoculated into approximately 5 mL of YPD to be grown for 24–48 h on a rotating culture drum or on a platform shaker. Larger assays can be scaled appropriately (*see* **Notes 11** and **12**).
2. When yeasts from the start culture are fully saturated, passage at least 1000 times more cells than the number of barcodes to the growth media of interest (e.g., a stressful growth media) into two parallel cultures to produce technical replicates. The inoculation volume to final volume should be at a fraction of 1/32 to 1/1024, and grown for 24 or 48 h (*see* **Notes 13–15**).
3. Continue passaging the culture at least twice (and up to ten times) for no more than 50 total doublings and aliquot a portion (we usually do at least the inoculation volume, but 1.5 mL is typical) of the saturated culture for DNA extraction at each re-inoculation. Pellet the cells to remove the supernatant. The pellet can be processed immediately for DNA extraction (*see* Subheading 3.2) or stored at −20 °C (*see* **Notes 16** and **17**).

3.2 Genomic Extractions of Yeasts

1. To the cell pellet, add 100–200 μL of yeast lysis buffer. Generally, 100 μL of yeast lysis buffer is sufficient to lyse over 10^8 yeast cells. Vortex well to resuspend, and place at 37 °C with mild shaking for at least 30 min and up to 2 h.
2. Add four volumes of binding buffer to the lysed cells (if 100 μL of yeast lysis buffer was used, add 400 μL of binding buffer) and vortex extensively. The lysed cell solution should clarify significantly indicating DNA release. If some cells remain un-lysed, centrifuge at 16,000 × *g* for 1 min to pellet them.
3. Pipet the clarified solution to a silica mini-preparative column fitted in a collection tube and centrifuge at 16,000 × *g* for 1 min. Discard the flowthrough.
4. Pipet 400 μL of wash buffer 1 onto the silica mini-preparative column and centrifuge at 16,000 × *g* for 1 min. Discard the flowthrough.

5. Pipet 600 μL of wash buffer 2 onto the silica mini-preparative column and centrifuge at 16,000 × *g* for 1 min. Discard the flowthrough.
6. Centrifuge at 16,000 × *g* for 3 min to dry the column.
7. Place the mini-preparative column into a clean 1.5-mL microcentrifuge tube, add 50 μL elution buffer to the center of the column and centrifuge at 10,000 × *g* for 2 min. Discard the column and store the eluted DNA, which is typically at 20 ng/μL, at −20 °C until further use.

3.3 Two-Step PCR Amplification of Barcodes

1. Prepare a 20 μL PCR reaction using the Q5 polymerase: 10 μL of genomic DNA (approximately 200 ng of DNA, or the DNA belonging to about 10^7 yeast cells, *see* **Note 18**), 4 μL 5× Q5 polymerase buffer, 0.4 μL 10 mM dNTP mixture, 1 μL each of primers priming around the barcode and containing the UMIs, 0.2 μL Q5 polymerase and water to 20 μL. Cycle as follows:
   (a) 98 °C for 1 min.
   (b) 98 °C for 10 s.
   (c) 50 °C for 10 s (or 5 °C lower than the T_m of the primers).
   (d) 72 °C for 30 s.
   (e) Repeat **steps 2–4**, twice (*see* **Note 19**).
   (f) 72 °C for 1 min.
   (g) 4 °C indefinitely.
2. Use AMPure XP beads to purify the reaction at a 1.25× bead ratio (*see* **Notes 20** and **21**):
   (a) Add 25 μL AMPure XP beads and vortex well (25 μL = 1.25 × 20 μL).
   (b) Place PCR tube on a magnet to collect the beads.
   (c) Pipette out the supernatant.
   (d) Add 180 μL AMPure XP beads washing buffer. The beads can be washed by taking the tube off the magnet, rotating the tube, and replacing it back on the magnet.
   (e) Pipette out the washing buffer and let dry at room temperature for about 2 min. Ensure that there is no remaining washing buffer at the bottom of the PCR tube.
   (f) Add 33 μL water to the beads and vortex well.
   (g) Place the PCR tube on a magnet to collect the beads.
   (h) Pipette the supernatant to a fresh PCR tube and store at 4 °C if not proceeding to the next step immediately (*see* **Note 22**).
3. Prepare a 50 μL PCR reaction using the KAPA polymerase: 33 μL of the AMPure XP cleaned reaction made in **step 2** in

this section, 10 μL 5× KAPA HiFi polymerase buffer, 1 μL 10 mM dNTP mixture, 2.5 μL each of primers containing the Illumina adapters binding to the flowcell and 1 μL of the KAPA HiFi enzyme. Cycle as follows:

(a) 98 °C for 1 min.
(b) 98 °C for 10 s.
(c) 62 °C for 10 s (or the T_m of the primers).
(d) 72 °C for 30 s.
(e) Repeat **steps 2–4**, 34 times.
(f) 72 °C for 2 min.
(g) 4 °C indefinitely.

4. Use AMPure XP beads to purify the reaction at a 0.8× bead ratio (*see* **step 2** in this section).
5. The cleaned PCR should be run on an agarose gel and visualized for successful reaction (*see* Fig. 3 and **Note 23**). It is ready to be quantified, pooled, and sequenced with an Illumina sequencing platform (*see* **Notes 24** and **25**).

3.4 Parsing Sequencing Reads into Barcodes

1. Parse the demultiplexed sequencing reads using the python fuzzy regex package [10] (*see* **Note 26**). The following python functions may be used as example for the described protocol. Prepare a python script resembling the following (adapting the barcode and UMI lengths, indexes, priming sites, and downstream sequences for your needs). This script will parse the barcode, the UMI, and the optional inline index for the first read and second read.

```
import regex

inner_indexes =
["XXXXXX","XXXXXXX","XXXXXXXX","XXXXXXXXX","XXXXXXXXXX","XXX-
XXXXXXXX","XXXXXX","XXXXXXX","XXXXXXXX","XXXXXXXXX","XXXXXXX-
XXX","XXXXXXXXXX"]

inner_indexes_2 =
["XXXXXX","XXXXXXX","XXXXXXXX","XXXXXXXXX","XXXXXXXXXX","XXX-
XXXXXXXX","XXXXXX","XXXXXXX","XXXXXXXX","XXXXXXXXX","XXXXXXX-
XXX","XXXXXXXXXX"]

def OR(inner_indexes):
  return "("+"|".join(["("+index+")" for index in inner_in-
dexes])+")"
UMI = "(.{8})"
barcode = "(.{16})"
inner_P1_priming_site = "YYYYYYYYYYYYYYYYYY"
```

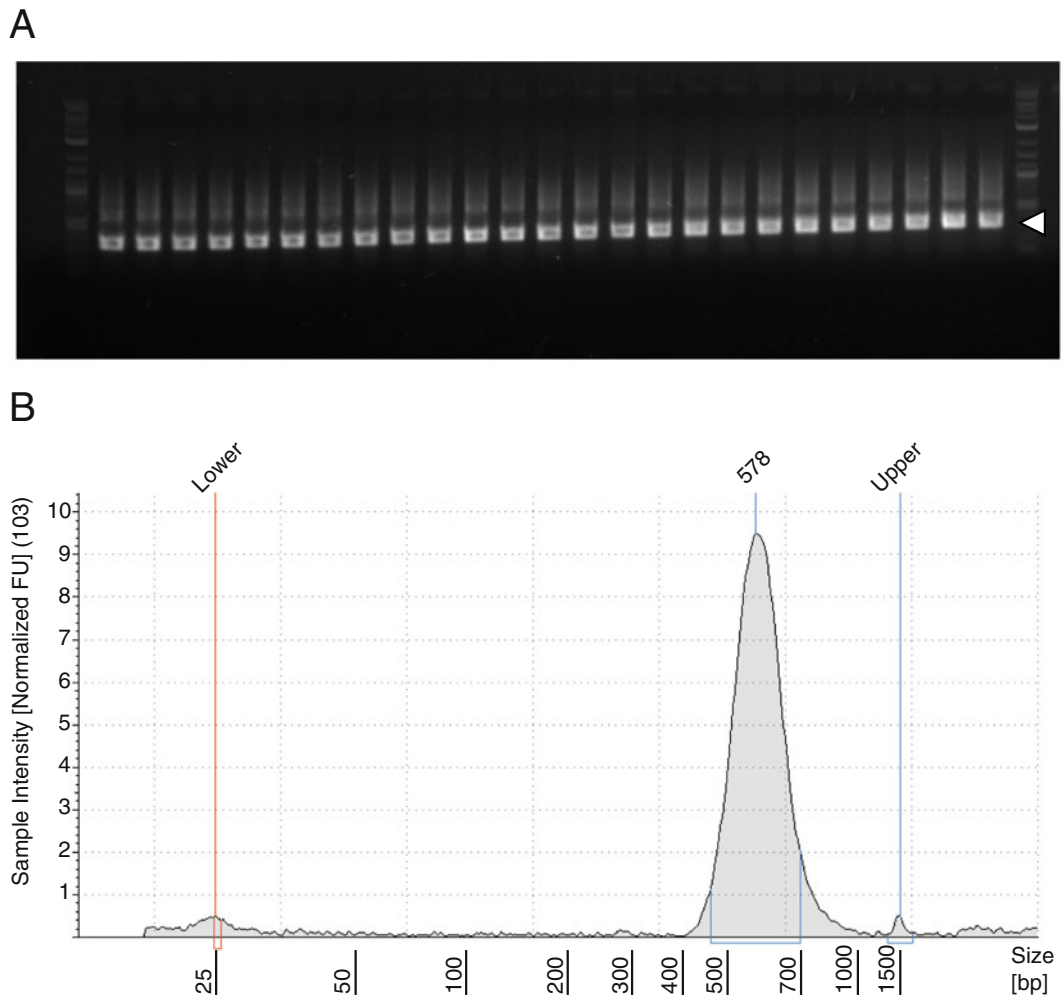

Fig. 3 Successful reactions. (**a**) Agarose electrophoresis of the final two-stage PCR reaction for 25 parallel reactions. The white triangle indicates the desired band, with expected fragment size of approximately 290 bp. 1 kb Plus DNA ladder from New-England Biolabs in first and last lane, with major bands representing 500 bp, 1000 bp, and 3000 bp. A faint smear or band at two times the fragment size can also be observed. (**b**) Agilent TapeStation analysis of the pooled two-stage PCR reaction from **a**. Measured fragment size by the TapeStation frequently shows a peak of two times the expected size

```
inner_P2_priming_site = "YYYYYYYYYYYYYYYYYY"
sequence_downstream = "ZZZZZZZZZZZZZZZ"

barcode_locus = (
 UMI +
 OR(inner_indexes[0:12]) +
 inner_P1_priming_site + "{e<=3}" +
 barcode +
 sequence_downstream + "{e<=3}")
```

```
barcode_regex_object = regex.compile(barcode_locus)
inner_p2_regex = ( UMI + OR(inner_indexes_2[0:12]) + inner_P2_priming_site + "{e<=3}")
inner_p2_regex_object = regex.compile(inner_p2_regex)
for line in sys.stdin:
 reads = line.split("\t")
 m = barcode_regex_object.match(reads[0])
 m2 = inner_p2_regex_object.match(reads[1])
 if m and m2:
  UMI_val = m.groups()[0] + m.groups2()[0]
  index_val = m.groups()[1] + m.groups2()[1]
  bc_val = m.groups()[2]
   print(str(UMI_val) + " " + str(index_val) + " " + str(bc_val))
```

2. Run the above script using the following commands to parse one pair of demultiplexed read files:

```
paste <(gunzip -c reads_illumina_1.R1.gz) <(gunzip -c reads_illumina_1.R2.gz) | python above_script.py >init_parsed_1.txt
```

3. If using inline indexes, it may be useful here to demultiplex the files further. This command will only output reads corresponding to the desired index (replacing the XXXXXX with the indexes of choice).

```
grep -P "\tXXXXXX\t" init_parsed_1.txt >parsed_1.txt
```

4. Once barcodes are parsed, they should be "error-corrected." When the list of barcodes is known (there is a dictionary), such as in the yeast deletion collection, correction of barcodes is done by using the lowest Levenshtein distance to the dictionary up to two mismatches. This can be done efficiently using a software package such as FLAMINGO [12] that uses N-grams or error-correcting tries such as used in short-read mapping strategies [13]. We provide a code in the Supplementary information that can efficiently perform this operation. Compile the software and prerequisites:

```
g++ error_correct_from_dict.cpp -o error_correct_from_dict -O3 -std=c++0x -I$HOME/local/include -lpthread -fopenmp -D_GLIBCXX_PARALLEL
```

This software has the prerequisite tsl robin_map freely available [11] which should be unpacked in your $HOME/local/include directory.

5. Run the software, which will error correct by default at two Levenshtein distance and discard reads that do not map to known barcodes (*see* **Notes 27** and **28**).

```
error_correct_from_dict -bc 3 -dict dictionary.txt parsed_1.
txt >corrected_1.txt
```

The -bc flag specifies the column in the parsed_1.txt file that contains the barcode of interest, and the dictionary.txt file contains a single column of known barcodes.

6. After error correction, barcode counts are obtained by counting their distinct unique molecular identifiers. This can be performed efficiently by keeping unique lines:

```
sort -u corrected_1.txt >unique_1.txt
```

7. The above process should be repeated for each time-point and each replicate assay. Counts can be obtained by simply counting each barcode for each time-point, and frequencies can be obtained by dividing the counts by the total counts for each time-point. A function such as this one is useful as it will provide counts based on the barcode field only (replacing the cut flag with the proper column):

```
cut -f3 unique_1.txt | sort | uniq -c | awk '{printf("%s\t%s
\n",$2,$1)}' >counts_1.txt
```

3.5 Fitness Inference

1. Following this parsing step, the data useful for fitness inference is the number of reads for each barcode, at each time-point, and for each replicate assays. We define the relative selection coefficients, s, based on the deterministic continuous model for logistic growth of alleles against each other with no frequency dependence. The general formula for the change in frequency of an allele over time:

$$\frac{\mathrm{d}f_i}{\mathrm{d}t} = f_i(s_i - \overline{s}).$$

Where we define:

$$\overline{s} = \sum_i f_i(t)s_i$$

as the population mean fitness at time t.

If we define $f_i(0)$ as the initial frequency of lineage i, the solution to the above equations is:

$$f_i(t) = \frac{f_i(0)e^{s_i t}}{\sum_j f_j(0)e^{s_j t}}$$

The fitness for every lineage is obtained by maximizing the likelihood of observed count data:

$$likelihood = C * \Pi\, \mathrm{f_i(t)}^{\mathrm{n_i(t)}}$$

where C is a constant factor, and $n_i(t)$ is the number of reads for lineage i at time t.

For two time-points, the relative fitness of a strain (s_i) compared to a reference strain ($s_j = 0$) is the change in logarithmic ratio of frequencies over time and is solved as:

$$s_i = \frac{\log\left(f_i(t_2)/f_j(t_2)\right) - \log\left(f_i(t_1)/f_j(t_1)\right)}{t_2 - t_1}.$$

For more than two time-points, the solution for multiple alleles is a joint-inference problem and numerical methods to maximize the likelihood of the observed count data can be used. We provide here a program that will do the full numerical solve, (*see* **Notes 29–31**) for a discussion on faster approximations that are roughly equivalent to each other and are useful when the number of reads is sufficiently large as in described in this protocol. A correlation over 0.99 between these approximations and a full numerical solve is typical when barcodes have at least 20 reads per time-point (*see* **Note 32**).

First, assemble counts into a tab-delimited text file with the following format:

```
BC 10 20 30 40 50
NNNNNNNNNNNNNNNN 1534 1079 602 346 101
NNNNNNNNNNNNNNNN 1061 1573 2200 3044 3509
```

which consists of a first header row which describes the following fields. There should be more than two time-points, and the header row should have integer values labeling the generations. The following rows should have the barcode (BC) as the first field, and integer values for the counts at each time-point. Ensure that all rows have the same number of fields. Ensure that barcodes are unique and do not appear multiple times throughout the file. Remove barcodes that have only a single time-point with non-zero counts: their fitness cannot be estimated.

2. Compile the program as follows, which has the same prerequisite as the error_correct script:

```
g++ -std=c++11 -O3 -I$HOME/local/include BFA_solve.cpp -o
BFA_solve -fopenmp
```

3. Run the software, including both the replicates of the same assay (replicates are not required and fitness values can be estimated independently, but estimation is better with replicates for lower frequency strains):

```
BFA_solve arranged_counts_1.txt arranged_counts_1_rep.txt
>fitness_1.txt
```

4. The software describes the following row header:

```
Lineage s stderr(s) f0_array
```

Following rows have the barcode, the fitness, the estimated error on this fitness, and the estimated starting frequencies of the barcode (when running the software with replicates). A reference strain barcode is chosen based on the highest read counts. To adjust all fitness values based on a wanted reference barcode, subtract the fitness values of all strains by the fitness of the reference barcode.

Some statistics are also given by the software, which describe the fit of the data to the model.

The software may take a long time to converge depending on the count values and the number of strains to estimate. We suggest using the noted approximations in cases where the software takes an unacceptably long time to converge.

4 Notes

1. The yeast lysis buffer can be made using 1 mg/mL of Zymolyase 100 T instead, but this reagent has been found to be more expensive.
2. The detergent (SB3-14) can be made as a 10% stock solution and dissolved with the help of moderate heat. Avoid mixing too vigorously as bubbles easily form.
3. The enzymatic solution is cloudy when thawing and turns into a brown, transparent liquid above approximately 10 °C. Using the lysis buffer when cloudy does not harm the reaction. The lysis buffer is stable upon multiple freeze-thaws.
4. The binding buffer may be more easily made from a concentrated solution of 6 M guanidine thiocyanate. Stir on a warm plate when dissolving as dissolving guanidine thiocyanate is endothermic.
5. It is important that the final pH of the solution be about 5 as DNA binding to silica preparative column is more efficient in acidic conditions.

6. The concentration of ethanol should remain above 75%, so it is important to seal the bottle containing the washing buffers accordingly.
7. Experience has shown that mini-preparative silica columns of different brands or of different types have different binding properties, such as binding capacities and fragment length retention, but that this protocol will yield sufficient high-quality DNA in all instances. Use columns that have specifications for large DNA fragments when possible. Alternatively, magnetic silica beads can also be used.
8. The two different polymerases used here have been optimized for sensitivity and library yield.
9. We use 12 inline inner indexes, which in combination with the Illumina outer indexes facilitates multiplexing large number of assays. The inline indexes should be of different length to prevent homogenous base compositions for each Illumina sequencing cycle.
10. We usually add ampicillin at 100 μg/mL to reduce the occurrence of bacterial contamination.
11. Barcodes should ideally uniquely represent one strain or genotype.
12. The barcoded yeast population can be maintained indefinitely in 5–20% glycerol at −80 °C, thawed at room temperature and vortexed before inoculation.
13. Any media that can sustain yeast growth can be used here, but for convenience it is recommended that the doubling time of the average yeast population in this media be no more than 8 h. In all cases, it is best to calibrate the media such that yeast do not spend more than 12 h at saturation as cell lysis is greatly reduced on saturated cultures.
14. The number of cells passaged per barcode (N) dictates the effect of genetic drift on fitness measurements [14]. To measure relative fitness with average errors in the order of 1% requires at least 100 cells passaged per barcode (and preferably much more). Increasing this number of cells passaged does not however decrease this error bound indefinitely as errors such as media batch variation, new mutations occurring in the assay, and inaccuracies in sampling time tend to dominate. In practice, we hardly obtain measurements more precise than 0.1%.
15. The inoculation ratio determines the number of doublings (or generations) until saturation, which can be determined by the number of doublings to carrying capacity. For example, an inoculation fraction of $1/2^5$ leads to 5 generations per day. The desired ratio is dependent on the fitness variance of the

population, with a higher relative volume of inoculum to fresh media being useful for large fitness variances, and a lower relative volume being convenient for small fitness variances, which can be determined empirically. Large fitness variance leads to higher rate of adaptation (the mean population fitness increases faster), which greatly reduces the frequency of several low-fitness lineages, and, as explained in **Note 12**, the error in fitness measurement is directly related to the number of cells passaged. Low fitness variance, on the other hand, only leads to deterministic changes in lineage frequencies over long periods of time. Experience has shown us that 5–10 generations per day usually lead to adequate fitness determinations. Because higher relative inoculation volume dilutes the stressful media, cells can be washed with fresh media prior to inoculation.

16. We typically passage 5–7 times total. The higher sampling and sequencing frequency leads to more accurate fitness measurements because errors in frequency determination by sequencing are reduced by the regression on multiple time-points.
17. Passaging for fewer than 50 generations limits the effect of new beneficial mutations within the growth assay.
18. The reaction will usually work well if the number of templates in the reaction is at least 10^6 (approximately 10 ng for yeast) and up to 1 μg of total template DNA.
19. The limited cycling ensures that true DNA barcode frequency in the genomic DNA is not distorted by exponential amplification, as duplicate UMI–barcode pairs will be removed bioinformatically. We have at times increased the number of cycles here and have yet to see large discrepancies.
20. AMPure XP bead ratios to use can be determined empirically or with useful tables that show retention of desired fragments to the magnetic beads [15]. In practice, we use ratios that maximize the retention of desirable fragments.
21. In many cases, the addition of carrier molecules such as highly purified rRNA from another species such as bacteria (at a final 50 ng/μL carrier RNA) can dramatically improve yield of this purification. However, we found that this sometimes leads to excess of primer dimers.
22. The PCR product is not visible at this stage if visualized on an agarose gel due to the limited cycling.
23. When the library is visualized by electrophoresis using the Agilent TapeStation and occasionally in agarose gels, we frequently observe one peak corresponding to the fragment of desired size, and a second peak at twice the size of the desired fragment. Sometimes the second peak is the only visible peak

on the TapeStation. We interpret the second peak to be due to incoherent migration of non-complementary DNA fragments (so-called "bubble" products) frequently observed in overamplified libraries. These do not interfere with downstream measurements.

24. We quantify each reaction with a Qubit dsDNA HS kit or equivalent, pool at equimolar concentration and quantify the final pool again.
25. The accuracy of the fitness measurement is strongly dependent on the number of reads or the sequencing coverage per barcode per time-point; however, the assay is less sensitive to the sequencing depth than the number of cells passaged per barcode given that the number of sampling time-points is relatively high. We usually aim for at least 100 reads per barcode per time-point, and it is usually not useful to go beyond the number of cells passaged per barcode, or beyond the number of genomic DNA fragments per reaction (which is typically around 10^7 total per time-point).
26. Sequencing reads containing barcodes can be parsed and demultiplexed with a variety of bioinformatics solutions including in-house scripts (e.g., *see* [16] or [17] for pipelines that includes error correction, chimera detection, duplicate read removal and frequency parsing) or using published methods designed for this task [18]. The example here is meant to show the procedure using the simplest pipeline possible and a variety of modifications depending on the primer and experimental designs are anticipated.
27. When barcodes are unknown, a clustering step can be used to first build this dictionary. In practice, we assume that well-represented barcodes that are all two Levenshtein distance from each other belong to the dictionary.
28. The error rate of barcode reads is dependent on the Illumina platform used and the quality of the data obtained. At optimal clustering on the NextSeq platform, about 0.1% of barcode reads of 16 bp will have one mismatch to a known barcode, and 0.01% will have two mismatches.
29. For the approximations: obtain the average s_i from all possible pairs of time-points using the above formula. This is approximately the same as performing simple linear regression on the logarithm of observed ratio of frequencies against the number of generations.
30. The two-point approximation cannot be used when a barcode is not observed for a time-point (no reads map to the barcode). Either the point should be ignored, or the second

approximation that simply performs linear regression on the logarithm of ratios should be used. Finally, the full joint-inference algorithm provided here accounts for this.

31. The approximations given here ignore the increased noise at lower sequencing read counts. Thus, data points at low read counts should have lower weights in the regression because they do not accurately estimate the actual lineage frequencies. The mentioned joint-inference algorithm accounts for this.
32. This fitness model can be made more accurate by incorporating genetic drift which becomes important when the number of cells at the passaging bottleneck is low [16, 19].

References

1. Petropavlovskiy AA, Tauro MG, Lajoie P, Duennwald ML (2020) A quantitative imaging-based protocol for yeast growth and survival on agar plates. STAR Protoc 1: 100182. https://doi.org/10.1016/j.xpro.2020.100182
2. Hall BG, Acar H, Nandipati A, Barlow M (2014) Growth rates made easy. Mol Biol Evol 31:232–238. https://doi.org/10.1093/molbev/mst187
3. Lenski RE, Rose MR, Simpson SC, Tadler SC (1991) Long-term experimental evolution in Escherichia coli. I. Adaptation and divergence during 2,000 generations. Am Nat 138:1315–1341
4. Thompson DA, Desai MM, Murray AW (2006) Ploidy controls the success of mutators and nature of mutations during budding yeast evolution. Curr Biol 16:1581–1590. https://doi.org/10.1016/j.cub.2006.06.070
5. Wiser MJ, Lenski RE (2015) A comparison of methods to measure fitness in Escherichia coli. PLoS One 10:e0126210. https://doi.org/10.1371/journal.pone.0126210
6. Shoemaker DD, Lashkari DA, Morris D et al (1996) Quantitative phenotypic analysis of yeast deletion mutants using a highly parallel molecular bar-coding strategy. Nat Genet 14: 450–456. https://doi.org/10.1038/ng1296-450
7. Venkataram S, Dunn B, Li Y et al (2016) Development of a comprehensive genotype-to-fitness map of adaptation-driving mutations in yeast. Cell 166:1585–1596.e22. https://doi.org/10.1016/j.cell.2016.08.002
8. Giaever G, Nislow C (2014) The yeast deletion collection: a decade of functional genomics. Genetics 197:451–465. https://doi.org/10.1534/genetics.114.161620
9. Lundberg DS, Yourstone S, Mieczkowski P et al (2013) Practical innovations for high-throughput amplicon sequencing. Nat Methods 10:999–1002. https://doi.org/10.1038/nmeth.2634
10. Barnett M regex: Alternative regular expression module, to replace re. https://pypi.org/project/regex/. Accessed 8 Apr 2021
11. Goetghebuer-Planchon T (2021) Tessil/robin-map. https://github.com/Tessil/robin-map/. Accessed 8 Apr 2021
12. FLAMINGO Package (Approximate String Matching) Release 4.1. http://flamingo.ics.uci.edu/releases/4.1/. Accessed 8 Apr 2021
13. Li H, Durbin R (2009) Fast and accurate short read alignment with burrows-wheeler transform. Bioinformatics 25:1754–1760. https://doi.org/10.1093/bioinformatics/btp324
14. Desai MM, Fisher DS (2007) Beneficial mutation selection balance and the effect of linkage on positive selection. Genetics 176:1759–1798. https://doi.org/10.1534/genetics.106.067678
15. Lundin S, Stranneheim H, Pettersson E et al (2010) Increased throughput by parallelization of library preparation for massive sequencing. PLoS One 5:e10029. https://doi.org/10.1371/journal.pone.0010029
16. Nguyen Ba AN, Cvijović I, Rojas Echenique JI et al (2019) High-resolution lineage tracking reveals travelling wave of adaptation in laboratory yeast. Nature 575:494–499. https://doi.org/10.1038/s41586-019-1749-3

17. Johnson MS, Martsul A, Kryazhimskiy S, Desai MM (2019) Higher-fitness yeast genotypes are less robust to deleterious mutations. Science 366:490–493. https://doi.org/10.1126/science.aay4199
18. Zhao L, Liu Z, Levy SF, Wu S (2018) Bartender: a fast and accurate clustering algorithm to count barcode reads. Bioinformatics 34: 739–747. https://doi.org/10.1093/bioinformatics/btx655
19. Levy SF, Blundell JR, Venkataram S et al (2015) Quantitative evolutionary dynamics using high-resolution lineage tracking. Nature 519:181–186. https://doi.org/10.1038/nature14279