gUMI-BEAR, a modular, unsupervised population barcoding method to track variants and evolution at high resolution

Short Title: A modular barcoding method to track variants and evolution

Shahar Rezenman[1*], Maor Knafo[1*], Ivgeni Tsigalnitski[1], Shiri Barad[1], Ghil Jona[2], Dikla Levi[2], Orly Dym[3], Ziv Reich[1§] and Ruti Kapon[1§]

[1]Dept. of Biomolecular Sciences, [2]Life Sciences Core Facilities, [3]The Dana and Yossie Hollander Center for Structural Proteomics

Weizmann Institute of Science Rehovot, Israel 7610001

*These authors contributed equally to this work

[§]Corresponding Author

ruti.kapon@weizmann.ac.il

ziv.reich@weizmann.ac.il

Abstract

Cellular lineage tracking provides the means to observe population makeup at the clonal level, allowing exploration of heterogeneity, evolutionary and developmental processes and individual clones' relative fitness. Its use, however, is limited because existing methods are highly specific, expensive, labour-intensive, and, critically, do not allow the repetition of experiments. To address these issues, we developed gUMI-BEAR (genomic Unique Molecular Identifier Barcoded Enriched Associated Regions), a modular, cost-effective method for tracking populations at high resolution. We first demonstrate the system's application and resolution by applying it to track millions of *Saccharomyces cerevisiae* lineages growing together across multiple generations, revealing fitness differences, lineage-specific adaptations to environmental changes and subtle dynamic shifts. Then, we demonstrate how gUMI-BEAR can be used to perform parallel screening and optimisation of virtually any number of gene variants, thus enabling unsupervised identification of individuals optimised for particular tasks. Comparison between multiple, identical libraries allowed us to reveal the interplay between stochastic and deterministic outcomes in this experiment.

Introduction

With the advent of single-cell techniques, it became clear that significant heterogeneity exists within populations of seemingly identical cells[1–6]. However, how individual adaptations translate into success or failure, or how population dynamics are affected by individual variations, is not yet clear. A number of experiments are beginning to shed light on these questions by employing procedures to quantify a population's clonal makeup[6–13]. Using these methods, multiple lineages can be distinguished by the introduction of DNA barcodes, short segments of DNA inserted into cells' genomes and passed on to their progenies. By sequencing the barcodes, one can capture a snapshot of the clonal composition of a growing population at any given moment. Successive records can be used to trace the prevalence of individual lineages through time[14], offering insight into adaptation dynamics and fitness changes as they occur *in the context of evolving populations* rather than in each strain on its own. Cellular barcoding has thus become invaluable for understanding core evolutionary processes such as clonal interference[15], differentiation of organs[16] and heterogeneity development in cancer[1]. Its utilisation is, however, limited because it is generally expensive, labour-intensive, highly specific to each biological system, and current modes of application do not lend themselves to repetition - a fundamental requirement in experimental research[17]. Another ability lacking from current barcoding techniques, which prevents them from being used to study particular adaptations, is the ability to *directly* single out specific clones for further study. Currently, if particular clones are found of interest, based on the dynamics of the population, the procedure for distinguishing them for further study requires first separating single clones using other methods such as streaking or sorting and then sequencing the

barcode of each colony or well to identify the variant. This procedure can be particularly cumbersome if one is looking for clones that are present at low frequencies as the probability of finding a clone goes as its percentage in the population. Thus, a clone present below 1% requires (on average) more than 100 colonies to be sampled and sequenced, a feat that is both formidable and expensive.

We set out to simplify and streamline the barcoding process so that it can be applied to a broader range of studies by incorporating the following capabilities, (1) generation of identical library duplicates, (2) direct variant identification, (3) precision and modularity in the insertion point of the barcode and in the number of participating lineages and cells and, (4) integration into different organisms. In this paper, we describe the method we devised to achieve these goals that we named gUMI-BEAR (genomic Unique Molecular Identifier-Barcode-Enriched Associated Regions) and show how it can be utilised to accomplish high-resolution lineage tracking at a relatively low cost as well as to directly test millions of variants of a gene in the context of a growing population. We note that this method can be applied by any lab versed with standard molecular biology techniques.

Results

Building a barcoded library

gUMI-BEAR is based on the unsupervised genomic barcoding of each cell in an initial population with a construct, the gUMI-box, containing a unique 24 bp sequence, the barcode, along with auxiliary elements required for successful transformation and ones that simplify downstream preparation of the samples for sequencing (see Fig 1 and

Methods). Thus, the DNA segment that is inserted into the genome upon transformation includes, in addition to the gUMI, the Illumina READ1 sequence that renders the segment ready for attachment of Illumina P5 adapters after the experiment. A generic linker sequence, designed to aid in the attachment of elements that will be necessary post-experiment, is placed downstream of the gUMI. To prepare the library for sequencing, we use two-step PCR amplification where a UMI and a READ2 sequence are added to each amplicon in the first, short-cycle step, endowing each barcode that enters the analysis pipeline with a tag (Fig 1g inset) and preparing it for index attachment. The tag provides a means to avoid biases that may be introduced in the multiple cycles required to produce enough DNA for sequencing from small samples and increases the accuracy of lineage frequency quantification.

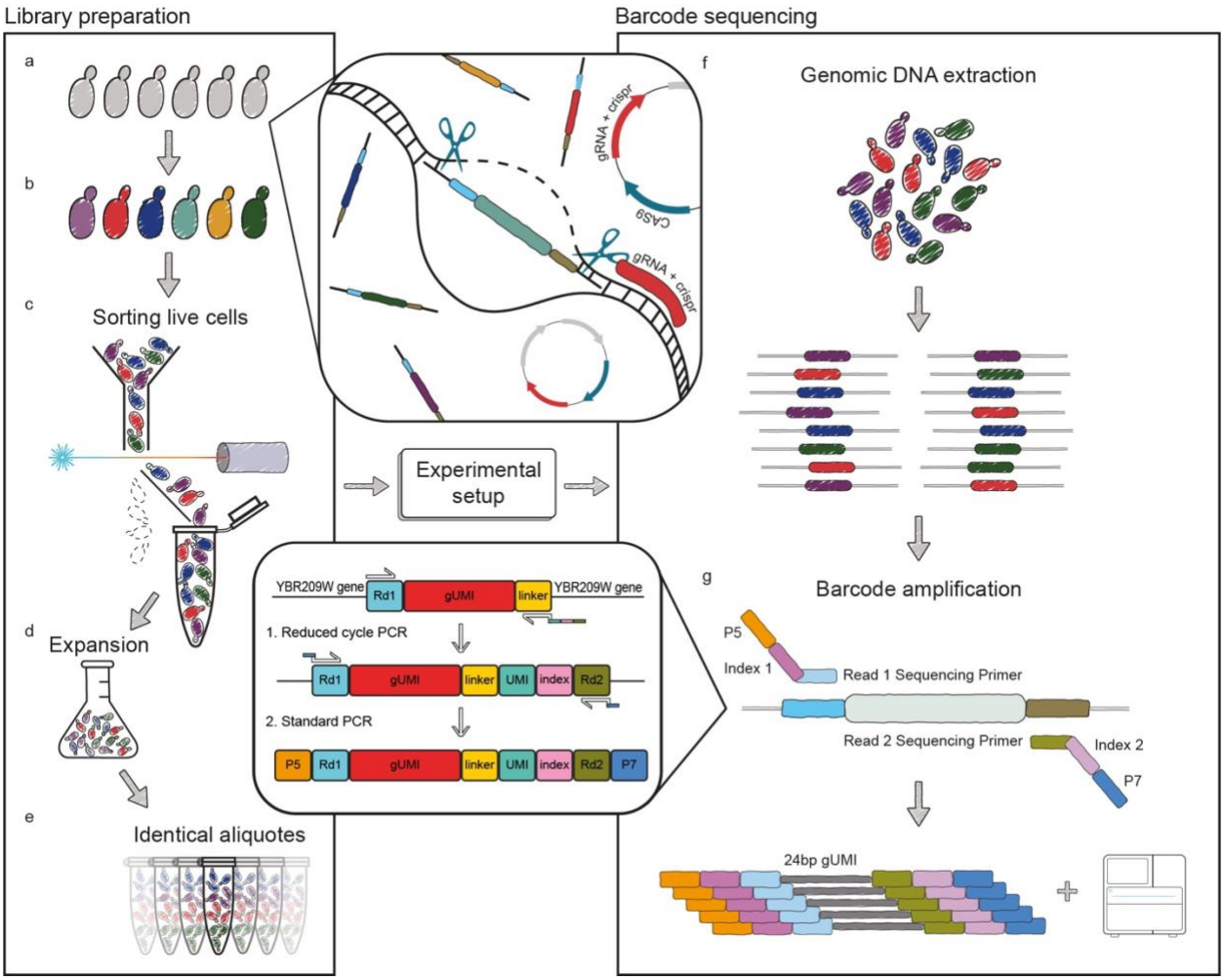

Fig 1. Schematic illustration gUMI-BEAR

An illustration demonstrating the method in its entirety, from library construction through experimental design, ending in the preparation of libraries for deep-sequencing. **(a, b)** An initial cellular population is barcoded by inserting unique 24 bp sequences into its genomes (inset). (**c)** Following recovery from transformation, live cells are sorted from the population (**d)** are allowed to grow, and (**e)** the resulting population is divided into aliquots containing equal numbers of each barcoded lineage, which are subjected to the chosen experiment after which (**f)** the cells undergo DNA extraction. (**g)** A two-step polymerase chain reaction (PCR) targeting the barcode region (inset) enables amplification of each lineage-associated variant (top) and subsequent sequencing of the barcode library (bottom).

To build a modular system that could be applied to a host of organisms as well as to any genomic location, we decided to incorporate the gUMI using CRISPR double-strand breakage of the integration site[18,19] together with the yeast homology-directed repair (HDR) mechanism. Although the latter is usually sufficient for transformation, we found that the number of successful transformants increased 20-fold (Fig 2b) with the addition of CRISPR, which in effect, allows almost any size library to be prepared. Equally important, this allowed us to easily change the locus of integration by simply replacing the gRNA in the pCAS vector and the homologous arms of the donor DNA while leaving the gUMI intact, thus satisfying our demand for flexibility in application.

Having the ability to produce a virtually unlimited number of barcoded cells, we next wanted to control the number of unique clones that we start each experiment with. This was achieved by sorting out the desired number of live cells from the culture at a time when the population had recovered from the transformation but had not yet started to divide (Fig 1c). At this time point, we can expect *uniquely* barcoded cells that have the potential to divide and thus produce distinct, identifiable lineages. To determine this time point, we performed time-resolved tests where, following transformation, we measured optical density (OD) and the percentage of dead cells in the culture (see Methods). We found that after 20–24 h of recovery, the percentage of dead cells started to decrease while the OD began to increase (Fig 2c). We thus identified 24 hours as the inflection point and the optimal time for cell sorting for all subsequent library preparations.

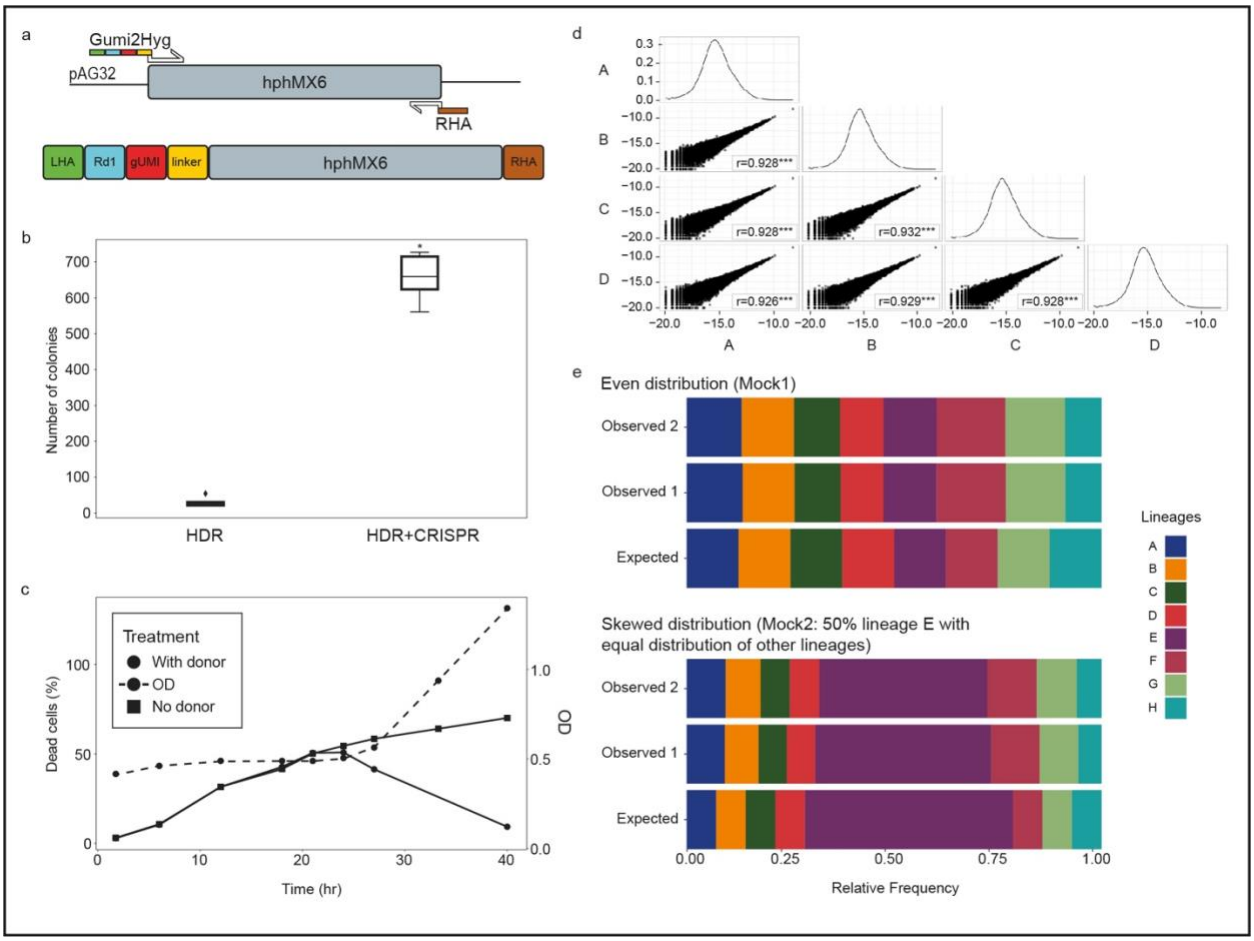

Fig 2. Construction of the barcoded library

(a) Schematic illustration of the donor DNA construction process described in Methods. (**b)** Cells were transformed with donor DNA with (white) or without (grey) a pCAS vector carrying the CRISPR/Cas9 machinery. The box plots present the number of colonies formed following transformation for five repeats of each treatment. The difference between the distributions is significant (t-test, p = 3.89e-8). (**c)** Percentage of dead cells (left y-axis) for a culture grown following transformation with (circles) and without (squares) donor DNA. Samples were taken for live/dead quantification using propidium iodide staining at intervals during the 40-h following transformation. The right y-axis presents optical density (OD; dashed line) measurements for the culture. **(d)** Four replicates of a library, A–D, were deep-sequenced to reveal their population composition and the log-transformed frequencies of the lineages are plotted against each. Pearson correlations were computed between all replicates and the Pearson coefficient, r, is presented at the bottom of each comparison. The diagonal presents the log-transformed frequency distribution for each

replicate (***$p < 0.0001$). (**e)** Bar plot of expected and observed lineage frequencies for two control populations, Mock1 (upper panel) and Mock2 (lower panel). Eight lineages were isolated from single colonies, and their barcodes were identified using Sanger sequencing. Their DNA was extracted and mixed in known ratios to create the Mock1 and Mock2 control populations. Each population was sequenced on two separate occasions, and the results underwent the same analysis pipeline to reveal its population composition. Each colour represents a single lineage according to the legend to the right of the graph.

One of the important prerequisites we set for the system was to be able to produce identical copies of libraries to allow comparisons between experiments. This entails allowing enough time for the cell population to expand sufficiently to produce a designated number of copies of each barcoded yeast to enable subsequent division into aliquots, while still minimising expansion time to decrease the probability of acquiring mutations before the start of the experiment. To assess the effect of various library preparation parameters on lineage makeup, we created seven mini-libraries that varied in the number of transformations performed, the number of cells sorted and the expansion times after transformation and sorting (Supplementary Table 1). The compositions of the populations in the resulting mini-libraries were determined by deep-sequencing the barcodes of four random aliquots from each. As expected, the number of viable lineages directly correlated with the number of cells sorted, although the precise number of lineages in the population was ten times lower than the number of lineages sorted. The distribution of barcodes and the similarity between aliquots were unaffected by transformation parameters or sorting (Supplementary table 1), demonstrating the robustness of the process.

Since we wanted to maximise the number of lineages available to us for the proof-of-concept experiments, all mini-libraries were pooled into a final library, which was then

divided into 100 aliquots for further studies. To analyse the barcode composition of the resulting library, we sequenced four of the aliquots. We found that these contain nearly identical barcode contents distributed log-normally, as shown in Fig 2d. Almost all 26,000 lineages were observed in all four samples (two were missing from one sample), with lineage frequencies exhibiting high between-aliquot correlations ($r = 0.990–0.999$; $p < 0.001$).

To verify that both the experimental procedure and the analysis pipeline (as described in Methods) produce an accurate description of the population, we constructed two communities consisting of different, known proportions of yeast from eight lineages whose barcodes had already been revealed by Sanger sequencing. Fig 2e presents the expected population structures along with the ones obtained by sequencing their barcodes on two separate occasions. In Mock1, which included equal proportions of all eight lineages, and in Mock2, with a non-uniform distribution, 50% of which is made up of lineage F and the remaining by equal proportions of the residual seven lineages, we observed an average error of 1.8%, with lineage F showing the greatest error in Mock2 (7.5%). The fact that errors were not consistent between the two mock populations and that results are almost identical when comparing the two different sequencing runs indicates that the variations from the expected structure are due to minor inaccuracies in the mock populations' construction process rather than sequencing errors.

Applying the gUMI-BEAR method to track evolutionary dynamics

Having verified our ability to expose the lineage makeup of populations, we set out to test the capabilities of our system in an experimental setting that allows diverse lineage

dynamics to occur along many generations. During a 44-day experiment, cells were grown in turbidostats, in YPD, and temperatures were varied three times, under the assumption that such changes would alter the fitness of lineages differentially and induce significant changes in the composition of the population[6,12,14]. We initially grew the population at 30 °C for two days, raised the temperature to 39 °C for eleven days, then to 41 °C for 27 days, finally reducing it back to 39 °C for the remaining four days of the experiment. To capture time-resolved population dynamics, samples were collected throughout the experiment and sequenced to determine the clonal makeup of the population at each time point.

To reveal the dynamics of individual lineages, we calculated the frequency of each lineage at each time point and used it to plot traces of lineage abundance as a function of time (Fig 3a and b). To test the sequencing depth used in the analyses and see that we were not under-sampling, which would make the results inaccurate, or oversampling, which would increase sampling costs unnecessarily we constructed rarefaction curves for both the number of unique lineages detected and their distribution. As seen in Fig 3c and 3d, both parameters plateau quickly at a low percentage of reads, indicating our sequencing depth was sufficient. Since the distribution of lineage frequencies is more uniform at the early stages of the experiment (Fig 3d, inset), the number of reads at which it reaches the plateau increases with time but is still well below our sequencing depth.

Using our analyses, we were able to follow a population that initially consisted of 26,000 (Fig 3c) distinct lineages distributed log-normally (Fig 3d, inset) and reveal frequency changes of individual lineages. The size of the population dropped to 16,000 after the first day, after which the population underwent several cycles of clonal

"takeovers" in which distinct lineages rose in frequency to become dominant (Fig 3a, b), with other lineages going down or remaining at low frequencies. The three temperature shifts induced sharp changes in population structure, the most dramatic manifestation of which was the immediate decline in abundance of the lineages that had been most successful under the previous condition, indicating that they had lost fitness relative to others. Surprisingly, this also occurred when we increased the temperature from 39 °C to 41°C, implying that adaptation to high temperatures does not necessarily occur gradually. We clustered the lineages based on their temporal trajectories and revealed five behavioural clusters. The clusters that displayed clonal takeovers almost always consisted of a few lineages, whereas another cluster was composed of most of the lineages that participated in the experiment and which occurred at low frequencies throughout. Notably, despite the high level of stress, ~2500 of the lineages (~15%) were not diluted out and remained in the population throughout the entire 44 days of the experiment (~128 generations). These results indicate that temperature changes can produce significant fitness gains for a small number of clones, allowing them to reach high frequencies. However, these differences in fitness are not enough to take over the population completely, allowing less-fit lineages to survive and rise to higher frequencies when conditions become more favourable for them.

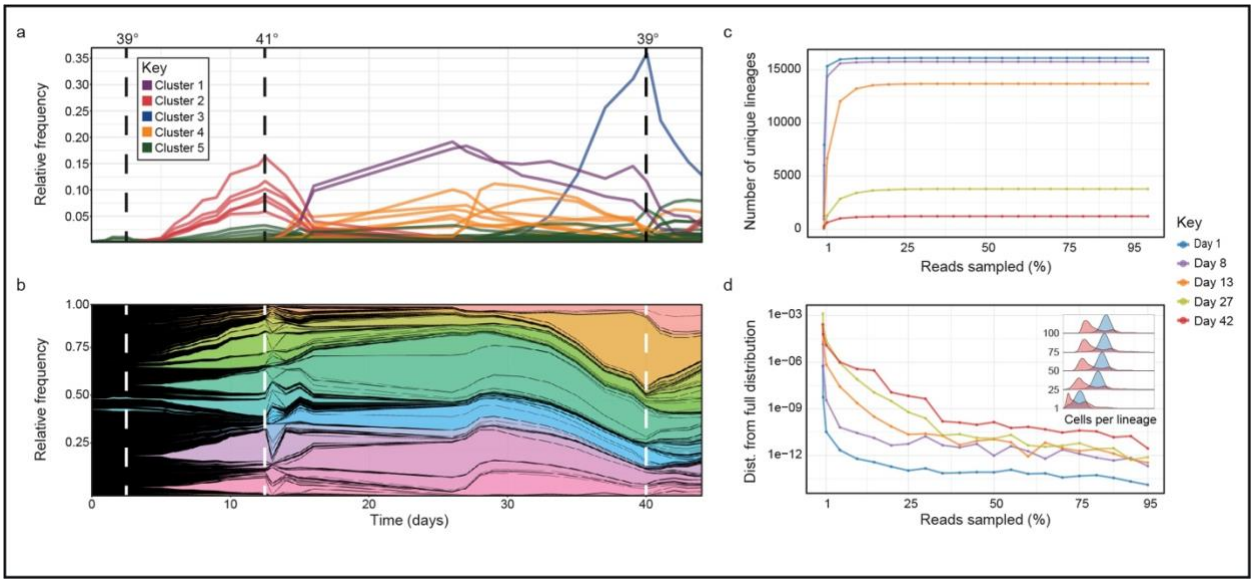

Fig 3. Applying the gUMI-BEAR method to track evolutionary dynamics

(a, b) Lineage trajectories throughout a 44-day experiment in which temperatures were altered (vertical dotted lines) to induce fitness fluctuations in the population. (**a)** Each lineage is represented by a line. Colours were assigned to each lineage based on K-means clustering performed on their trajectories throughout the experiment (as seen in the key). (**b)** Muller plot where each lineage is represented by a different colour and the width of the coloured region at each timepoint is proportional to the lineage's relative frequency. (**c, d)** Rarefaction curves for the number of lineages (**c**) and lineage distributions (**d**) obtained by sampling an increasing number of reads at five time points (depicted by the different colours as defined in the key). (**d)** The accuracy of the distribution is quantified by the Wasserstein distance between the distribution obtained with a reduced number of reads and the original distribution (based on 100% of the reads). The inset shows a ridge plot of the frequency distribution of barcodes in the population quantified by sampling an increasing number of reads (y-axis, percentage of population sampled) from Day 1 (blue) and Day 44 (red).

Applying the gUMI-BEAR method to track gene variants

Another implementation we envisioned for our method is in the study of gene variants. The mode of application, in this case, is to conjugate gUMI-boxes containing different

barcodes to specific gene variants. This creates an identity between barcode and variant and allows to use barcode incidence as a proxy for the abundance of each variant in a population. Thus, the fitness of millions of variants can be tested together, in one experiment, in the evolutionary setting in which they compete with each. Because barcodes are short and are sequenced as a whole, inaccuracies inherent to the assembly of gene ORFs from short reads are avoided, allowing the method to be implemented on any gene length with any number of mutations. Importantly, the reverse-barcode can be used as a PCR primer to directly isolate variants of interest for further analysis and study. We chose *Hsp82* to demonstrate these capabilities because the protein it encodes has chaperone-like activity and affects fitness gain and loss via variability control[20,21]. Hence, a collection of mutants is likely to elicit a population of yeast cells with various fitness levels. Because we were essentially not limited in the number of variants we tested and to avoid bias in their construction, we used a random mutagenesis kit (Genemorph II by Agilent part number 200550, see Materials and Methods) at conditions that assured 0–4.5 mutations/kb. The resulting variants were conjugated to the gUMI-box, creating a donor molecule ready to be transformed into the native locus in the yeast genome, replacing the wild-type *Hsp82* (Fig 4b and Methods). Identical aliquots of the library were prepared in a manner similar to that described above.

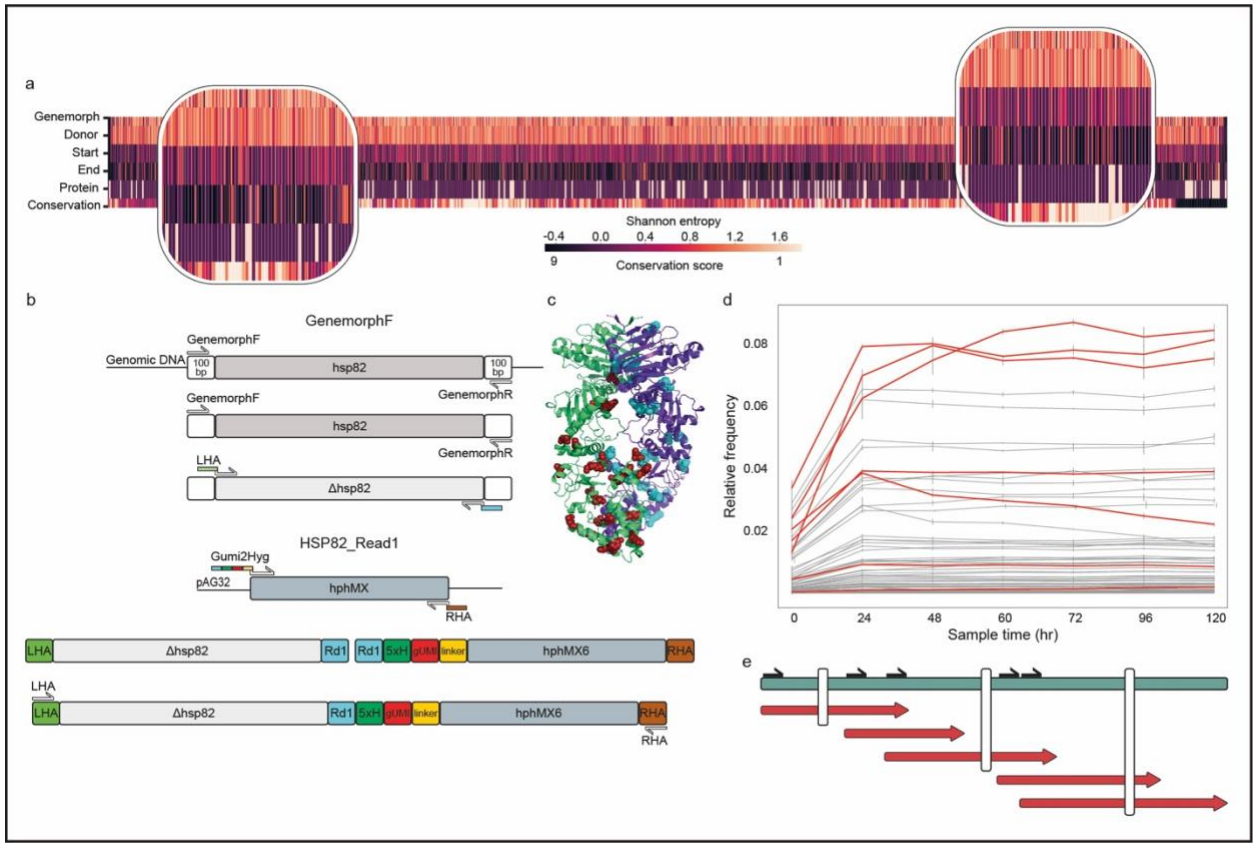

Fig 4. Applying gUMI-BEAR to track gene variants

(a) Heatmap showing entropy values (4 upper bands, as per the colour bar) for mutations in each nucleotide in the Δ*Hsp82* variants for various steps in the library construction process. For each position, mutational proportions were normalised to sequencing depth and Shannon's entropy was calculated (dark to bright represents low to high entropy). From top to bottom: The Genemorph panel refers to *Hsp82* variants created by insertion of random mutations using PCR with the Genemorph II kit. The Donor panel refers to the final donor molecule before transformation (mean entropy of three replicates). The panels labelled Start and End show the population at the start of the experiment, after transformation of the donor molecules and selection for successful transformants (mean of three replicates) and at the end of the experiment, respectively. The Protein panel shows mutational occurrence in the translated protein obtained by long-read sequencing of full-length variants (mutations are identified in white). The Conservation panel shows the conservation score calculated using the ConSurf[18] server (dark for conserved and bright for variable regions). The insets show two zoomed regions exhibiting different patterns where both conserved and variable regions can have non-synonymous mutations. (**b)** Donor construction process as described in the

Methods section. (**c)** The Hsp82 protein dimer (shown in green and purple). All mutations found by long-read sequencing are shown as red and cyan balls in the Hsp82 dimer. The image was produced with the Protein Homology/Analogy Recognition Engine (Phyre2). (**d)** Changes in relative lineage abundances within the yeast population over time. Each line represents the mean relative abundance of a lineage in the four replicates as a function of time, with error bars showing the standard deviation between replicates. Red lines mark the seven lineages whose *Hsp*82 variant was amplified using the gUMI sequence as a reverse primer and then Sanger-sequenced. (**e)** A scheme presenting Sanger sequencing using five forward generic primers distributed along the *Hsp82* ORF with a specific gUMI reverse complementary primer that was used to obtain a constructed full-length sequence of the variant (Red). The constructed variants are compared to the full-length sequence of the variant achieved by long-read sequencing (Loop Genomics) in green.

To test the effect of the assembly process on the distribution of mutations, we deep-sequenced amplicons from two stages in the assembly process: after mutagenesis of the gene and after assembly of the final donor molecule containing the gUMI-box. A comparison of the sequences of the *Hsp82* variants emerging from the mutagenesis step and of the donor molecules revealed a decrease in the number of variants after integration of the ORF. This is to be expected, as the variant-creation process is performed in a sequence agnostic manner, whereas genomic integration has biological constraints and, as such, eliminates some mutations. We then calculated the distribution of Shannon's information entropy of mutated nucleotides per site (see Fig 4a and supplementary Fig 4). The mutated ORF within the donor molecules displayed the same entropy distribution in three independent assembly replicates, confirming that the process of integration into the gUMI-box was robust. Likewise, this distribution of mutational entropy was identical to that of the molecules emerging from the Genemorph process showing that integration

into the gUMI-box during the donor DNA assembly process does not introduce bias (Fig 4a)

For variant screening, we used batch mini-cultures to grow four replicates that were sampled and diluted at a rate of 1:100 every 12 hours, for a total of 120 hours. Samples were sequenced to yield the relative abundance of each lineage as a function of time (Fig 4d). We found the distribution of lineage frequencies to be heavily skewed throughout the experiment, with 50 lineages comprising almost all (99%) of the population. Of these, lineages in ranks 25 - 50 were each present at a relative abundance of 1% or less and all were successfully tracked by the gUMI-BEAR method. The frequency of some lineages increased or decreased throughout the experiment while the abundance of most lineages remained constant, indicating equal relative fitness. The trajectory of the most successful lineages, namely those ranked among the top 50 in terms of their abundance, was reproducible between replicates, as indicated by the high correlations between different replicates ($r = 0.89 \pm 0.034$) and the small standard deviation (Fig 4d). This, in turn, signifies that the dynamics were determined by the variants that entered the experiment and not by adaptations that occurred throughout it, which could differ between repeats.

To demonstrate the capability of the system to identify variants of interest, we chose seven lineages for further analyses. As a first step, we used their reverse gUMI sequences as PCR primers to isolate specific variants. We were able to sequence a variant ranked as low as the 47^{th} rank in a sample taken after 108 hours, constituting 0.19% of the population, demonstrating how even low-frequency lineages can be directly isolated from a mixed population (Fig 4 d, e). The complete sequences of the Hsp82

variants were determined by Sanger sequencing using five primers distributed along the gene (Fig 4e). A comparison of the sequences obtained in this manner with sequences from full-length third-generation sequencing (Loop Genomics) confirmed the quality of variant sequencing (Fig 4e). Note that using present methods, 500 single cell colonies, on average, would need to be sequenced to single out a variant present at this low percentage.

Next, we compared the distribution of mutations at the different stages of the experiment. When we compared short reads obtained from the initial yeast library, after transformation and expansion, with long reads from the population after it had undergone 120 hours of evolution (Fig 4a), there was an overall decline in entropy, indicating that the experiment began with numerous mutations holding reduced fitness which were eliminated as growth continued. In the final library, we also found regions where no mutations had occurred, presumably because mutations in these areas reduce fitness significantly. The process of growing a population that harbours random mutations in a gene thus provided a biological filter, at evolutionarily relevant conditions, that eliminated lethal or function-harming mutations while allowing the propagation of mutations that were neutral or that provided a fitness advantage. Thus, we were able to highlight the fittest variants possible within this mutation rate without exerting any bias.

Next, we translated the variants into amino acid sequences and compared these with evolutionary conservation scores obtained using ConSurf[22]. We found a significantly higher frequency of mutations in areas considered variable. Nevertheless, some mutations were found in highly conserved positions (Fig 4a). To gain insights into the structural role of mutations in full-length variants, we entered the 50 sequences, obtained

from the long-read sequencing, into the web interface of the Protein Homology/Analogy Recognition Engine (Phyre2) portal to model the structures of these proteins[23]. Overall, we found 39 mutations that are shown in the Hsp82 homodimer model as red and cyan balls for the two Hsp82 monomers (Fig 4c). These analyses (described in more detail in the supplementary results) did not reveal a structural basis for a variant's success, illustrating that our method can identify beneficial mutations that are not based on known mechanisms and would otherwise have remained hidden.

Discussion

In this work, we describe a single-cell barcoding method that is versatile, easy to implement, and cost-effective, thus lending itself to a host of experimental scenarios. These traits are achieved due to a combination of features: Firstly, we used CRISPR/cas9 coupled with the cellular HDR to integrate the barcode containing gUMI-box, thereby achieving a high yield of transformants which in turn allows the construction of libraries of virtually any size. Furthermore, the gUMI-box integration site can easily be changed by simply designing a new gRNA and a corresponding donor DNA. Secondly, we perform a sorting step following recovery from the transformation to obtain precise control over the number of lineages participating in the experiment. Importantly, our method allows the assembly of multiple identical replicates of the same initial population, opening the way for comparisons between experiments. This, in turn, makes it possible to apply barcoding towards investigating the role of randomness and determinism in highly complex biological systems[19,24,25]

The gUMI-box, together with the simple and efficient protocol we applied to prepare gUMI-barcoded sequences for deep-sequencing (Fig 1g), eliminates the need for multiple preparation steps and, at the same time, provides high resolution and accuracy and supports the use of small samples.

We presented two applications to demonstrate the potential and adaptability of the gUMI-BEAR method. The first application exhibited its scale and precision in tracking ~16,000 lineages through 30 timepoints over a period of 44 days and revealed detailed evolutionary dynamics resulting from the complex environmental regime. We obtained an intricate picture of the rich population dynamics that arise as multiple lineages adopt numerous modes of exploration and exhibit varying adaptabilities to changing environmental conditions. Attaining such a picture through standard population studies, such as bulk RNA sequencing, whole-genome sequencing, or plating assays, would be financially and or analytically impractical.

In the second experiment, we used gUMI-BEAR to explore gene-variant fitness and dynamics of populations with induced fitness variations. By coupling a gUMI to each variant, we were able to scan a large number of random mutations produced *in vitro* in an evolutionarily relevant context in which they compete with each other, eliminating the need for either fitness testing of monoclonal variants or costly in-depth sequencing. Due to our ability to produce identical libraries, we were able to observe largely deterministic dynamics for this population. We used reverse-gUMI sequences as PCR primers to target specific variants and then used Sanger sequencing to identify the variation without the need for laborious work to single out variants and expensive long-read sequencing. Our method thus allows in-depth exploration of millions of variants, all living and competing

as a population, while also providing the ability to explore interesting variants, regardless of their frequency, using simple techniques such as PCR and Sanger sequencing.

We note that the gUMI-BEAR method can be utilised for variant screening in more complex systems consisting of multiple genes by adding a short sequence to identify each ORF. Under such a configuration, the relevant genes can be multiplexed to create an array of randomly mutated genes interacting with each other to enable the study of complex enzymatic pathways.

We designed gUMI-BEAR to be an easy-to-use and robust method to track changes in population makeup and variant fitness. Previously available methods[1,3,10] to track such changes were not scalable and were difficult to deploy on new lines and modalities or more costly. A library can be created using the gUMI-BEAR method in a few days and our results show that it provides an accurate, cost-effective, and versatile means of tracking subtle changes that propagate into major population shifts in a way that can easily be adapted to other biological systems.

Acknowledgements

We thank Shira Holand for the graphic design, Yonatan Nutkewitz for helpful discussions and help with experiments and Hadas Keren-Shaul for help with next-generation sequencing.

Methods

The methods described below were common to both of the experiments conducted, except where otherwise noted.

Construct assembly

gUMI-box (Genomic Unique Molecular Identifier)

The gUMI-box contains a 24 bp random sequence that serves as a transmissible barcode for each lineage in the population. The gUMI is flanked by the read 1 Illumina generic sequence on one side and a 15 bp generic sequence called a linker on the other side. The linker is used as a target sequence for DNA amplification in the deep-sequencing library preparation step, such that two-step PCR produces a library containing all elements necessary for sequencing on Illumina platforms (Fig 1g). An additional hygromycin-resistance cassette (HGmx6; obtained from the vector pAG32[26]) is added downstream of the linker sequence to allow selection and to prevent contamination (Fig 2a). As a final step, two homologous arms are added to the gUMI-box upstream and downstream of the genomic integration site. All elements are combined into one linear DNA construct, the gUMI-box, which is then transformed into a population of yeast cells.

Donor Construction

For a population exhibiting no initial fitness variations

Donor DNA was built using two primers containing 50 bp sequences inserted upstream (left homologous arm; LHA) and downstream (right homologous arm; RHA) of the integration site in the *YBR209W* locus, which encodes a putative protein of unknown function and a non-essential gene[27]. The forward primer consists of the specific LHA and most components of the gUMI-box. At its 3' end, it contains a complementary sequence

to the HGmx6 cassette from pAG32. The reverse primer consists of the specific RHA and a sequence complementary to the end of the HGmx6 cassette. PCR using these two primers produces the full linear DNA used for transformations in later steps (Fig. 2a). The double-strand break site was determined using CHOPCHOP[19]. The pCAS plasmid was cloned by restriction-free (RF) cloning with two primers containing the gRNA sequence (5`-TAGAGCGTCAATCAAGAAAG-3`) as described previously[13].

For use in tracking gene variants

Six-step PCR was utilized to obtain a full-length donor DNA comprising 5'-LHA-d*Hsp82*-gUMI-box-RHA-3'. *HSP82* was amplified from *S. cerevisiae* BY4741 gDNA to obtain a wild-type copy of the gene. Random mutations were introduced using the Genemorph II random mutagenesis kit (Agilent, 200550). To achieve a low mutation rate (0–4.5 mutations/kb), 1 µg of the template DNA was amplified using this kit in a 30-cycle PCR protocol. A series of PCRs containing the gUMI-box and the *ΔHsp82* were performed to create the full-length construct (Fig 4b and Supplementary methods).

Two pCAS plasmids were cloned by RF cloning, each with two primers containing the gRNA sequence for the cleavage site upstream (5'-CAAACAAACACGCAAAGATA-3`) and downstream of the *HSP82* gene (5'-AGCTGACACCGAAATGGAAG-3`)

Transformation

For tracking the evolution of a population exhibiting no initial fitness variations

S. cerevisiae BY4741 yeast cells were transformed with donor DNA (5 µg) and cloned pCAS (2 µg) according to standard protocols[19,28]. Following a three-hour recovery in 450 µL of YPD media, ten aliquots were mixed and incubated for 20 h in 100 mL YPD broth with hygromycin B (300 µg/mL) at 30 °C at 220 rpm. To accurately determine the number of lineages involved in each experiment, the cells were stained with propidium iodide[27] to mark perforated cells and the desired number of viable cells was sorted according to the library parameters given in Supplementary Table 1. The library was grown in 100 mL YPD with Hygromycin B and was allowed to propagate for 32 h (~16 generations). This step was followed by the division of the library into 100 identical aliquots that were then immediately frozen in 50% Glycerol at -80 °C (Fig 1a).

For use in tracking gene variants

Cells were transformed as described above. Following recovery, three aliquots were incubated for 20 h in 10 mL YPD with Hygromycin B at 30 °C at 220 rpm to select viable transformants. The library was diluted 100-fold and was allowed to propagate for an additional 30 h. This step was followed by the division of the library into 100 identical aliquots that were immediately frozen in 50% Glycerol at -80 °C.

Experimental outline

Evolution of the population exhibiting no initial fitness variations

The experiment was performed in a turbidostat at a constant volume of 100 mL YPD supplemented with Hygromycin B (200 µg/mL). Genetic drift was kept to a minimum by maintaining the concentration of the culture within a narrow rate, eliminating large dilution steps. The AU (absorbance unit) value was kept within a constant, narrow range of 0.9–1.1. To test the robustness of the gUMI-BEAR method and its response to changing environments, cells were grown for a total of 44 days under varying conditions designed to induce significant changes in population composition as follows: The initial temperature of the culture was set to 30 °C. After 48 hours, the temperature was increased to 39 °C for eleven days, 41 °C for 27 days and reduced back to 39 °C for the remainder of the experiment (Fig 3a, b). A total of 31 samples were taken throughout the experiment to capture detailed population dynamics.

Evolution of a population incorporating gene variants

Four library aliquots were mixed to create a homogeneous population and to minimise minor differences between aliquots. The mix was evenly distributed to four replicates that were grown in 100 mL Erlenmeyer flasks containing 25 mL YPD with Hygromycin B (200 µg/mL). Environmental conditions were maintained at a constant temperature of 30 °C and constant mixing at 220 rpm on a magnetic stirrer. The experiment spanned five days such that every 12 hours, as cells approached the end of their exponential growth phase, a sample was taken, and the culture was diluted 1:100.

Sample preparation and sequencing

DNA was extracted using the MasterPure Yeast DNA purification kit (Epicentre, MPY80200), and concentrations were measured by Qubit. A sub-sample (150 ng) was taken from each sample to serve as a template for four-cycle PCR in which the barcode region was linearly amplified, and the internal index and UMI were added (Fig 1g). Five samples from this reaction were pooled on a Qiagen PCR clean-up column (Cat. 28104) and eluted in a 50 µL elution buffer. These samples were further cleaned by 1.8× Ampure XP beads (Cat. A63881). From each pool, 10 µL were used for 12–17 cycles PCR using Illumina indexed primers. Samples were purified using 1.8× Ampure XP beads, and their quality was evaluated using Agilent Tape-station (Cat. 5067-5584) and Qubit concentration measurements (Cat. Q32853). All samples were diluted to a concentration of 0.5 ng/µL, and 10 µL of each diluted sample was used to create the final library. This sample was further diluted to a final concentration of 2 nM in a final volume of 100 µL. The library was sequenced at the Weizmann Institute of Science's G-INCPM unit on an Illumina NovaSeq platform with the following cycles: 29|10|10|26.

Preliminary analysis of raw data

Raw data were filtered to remove low-quality reads. Demultiplexing of samples was performed in two parts using Illumina indices and the internal indices. In order to count lineage frequency and determine which barcodes originated from a cell and which from sequencing mistakes, the CD-HIT[29,30] algorithm was used in two steps. In the first, seed sequences are created from samples taken only from the first time point of the experiment. Next, sequences from each specific time point were compared with the seed sequences list (using the cd-hit-est2d function). In each step, we allowed up to a two bp

mismatch before assigning the sequence to a new seed. The CD-HIT function clst2txt.pl was used to create lineage count data frames for each sample. The final output of this process contained many clusters of sequences that were each classified as a single lineage for further analysis. In the final step of the preliminary analysis, custom, in-house R-code was used to pool together all sequences from the same cluster with the same UMI to one cell. The final output of the abovementioned pipeline accurately measured the number of actual, single cells originating from a specific lineage for every time point after correction of PCR biases. For full script see rezenman/gUMI-BEAR github page

1. Shaffer, S. M. *et al.* Rare cell variability and drug-induced reprogramming as a mode of cancer drug resistance. *Nature* **546**, 431–435 (2017).
2. Locke, J. C. W., Young, J. W., Fontes, M., Hernández Jiménez, M. J. & Elowitz, M. B. Stochastic pulse regulation in bacterial stress response. *Science* **334**, 366–9 (2011).
3. Ackermann, M. A functional perspective on phenotypic heterogeneity in microorganisms. *Nat. Rev. Microbiol. 2015 138* **13**, 497–508 (2015).
4. Raser, J. M. & O'Shea, E. K. Noise in Gene Expression: Origins, Consequences, and Control. *Science* **309**, (2005).
5. Pilpel, Y. Noise in Biological Systems: Pros, Cons, and Mechanisms of Control. in *Methods in molecular biology (Clifton, N.J.)* vol. 759 407–425 (2011).
6. Nguyen Ba, A. N. *et al.* High-resolution lineage tracking reveals traveling wave of adaptation in laboratory yeast. *Nature* **575**, 494 (2019).

7. Oliver, K. M., Russell, J. A., Moran, N. A. & Hunter, M. S. Facultative bacterial symbionts in aphids confer resistance to parasitic wasps. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 1803–7 (2003).
8. Birgy, A. *et al.* Origins and breadth of pairwise epistasis in an α-helix of β-lactamase TEM-1. *bioRxiv* 2021.11.29.470435 (2021) doi:10.1101/2021.11.29.470435.
9. Levy, S. F. *et al.* Quantitative evolutionary dynamics using high-resolution lineage tracking. *Nature* **519**, 181–186 (2015).
10. Rodriguez-Fraticelli, A. & Morris, S. A. In preprints: the fast-paced field of single-cell lineage tracing. *Development* **149**, dev200877 (2022).
11. Cvijović, I., Nguyen Ba, A. N. & Desai, M. M. Experimental Studies of Evolutionary Dynamics in Microbes. *Trends Genet. TIG* **34**, 693–703 (2018).
12. Blundell, J. R. & Levy, S. F. Beyond genome sequencing: Lineage tracking with barcodes to study the dynamics of evolution, infection, and cancer. *Genomics* **104**, 417–430 (2014).
13. Jasinska, W. *et al.* Chromosomal barcoding of E. coli populations reveals lineage diversity dynamics at high resolution. *Nat. Ecol. Evol. 2020 43* **4**, 437–452 (2020).
14. Barrick, J. E. *et al.* Genome evolution and adaptation in a long-term experiment with Escherichia coli. *Nature* **461**, 1243–1247 (2009).
15. Park, S. C. & Krug, J. Clonal interference in large populations. *Proc. Natl. Acad. Sci.* **104**, 18135–18140 (2007).

He, Z. *et al.* Lineage recording in human cerebral organoids. *Nat. Methods* **19**, 90– 16. 99 (2022).

17. Lobkovsky, A. E. & Koonin, E. V. Replaying the tape of life: Quantification of the predictability of evolution. *Front. Genet.* **3**, 246 (2012).
18. Ryan, O. W. *et al.* Selection of chromosomal DNA libraries using a multiplex CRISPR system. *eLife* **3**, e03703 (2014).
19. Ryan, O. W., Poddar, S. & Cate, J. H. D. CRISPR–Cas9 Genome Engineering in *Saccharomyces cerevisiae* Cells. *Cold Spring Harb. Protoc.* **2016**, pdb.prot086827 (2016).
20. Cowen, L. E. & Lindquist, S. Hsp90 Potentiates the Rapid Evolution of New Traits: Drug Resistance in Diverse Fungi. *Science* **309**, (2005).
21. Cote-Hammarlof, P. A. *et al.* The Adaptive Potential of the Middle Domain of Yeast Hsp90. *Mol. Biol. Evol.* **38**, 368–379 (2021).
22. Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N. & Sternberg, M. J. E. The Phyre2 web portal for protein modeling, prediction and analysis. *Nat. Protoc. 2015 106* **10**, 845–858 (2015).
23. Stewart-Ornstein, J., Weissman, J. S. & El-Samad, H. Cellular noise regulons underlie fluctuations in Saccharomyces cerevisiae. *Mol. Cell* **45**, 483–93 (2012).
24. Koonin, E. V. & Wolf, Y. I. Is evolution Darwinian or/and Lamarckian? *Biol. Direct* **4**, 42 (2009).
25. Raj, A. & van Oudenaarden, A. Nature, nurture, or chance: stochastic gene expression and its consequences. *Cell* **135**, 216–26 (2008).
26. Addgene: pAG32. https://www.addgene.org/35122/.
27. Fisk, D. G. *et al.* Saccharomyces cerevisiae S288C genome annotation: A working hypothesis. *Yeast* **23**, 857–865 (2006).

28. CLONTECH. *Yeast Protocols Handbook FOR RESEARCH USE ONLY Yeast Protocols Handbook*. www.clontech.com (2009).

29. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).

30. Li, W., Fu, L., Niu, B., Wu, S. & Wooley, J. Ultrafast clustering algorithms for metagenomic sequence analysis. *Brief. Bioinform.* **13**, 656–668 (2012).