Modelling The Fitness Landscapes of a SCRaMbLEd Yeast Genome

Bill Yang
ICOS
School of Computing
Newcastle University
1, Urban Sciences Building
Science Square, Newcastle upon Tyne, UK

Goksel Misirli
School of Computing and Mathematics
Keele University, UK
g.misirli@keele.ac.uk

Anil Wipat
ICOS
School of Computing
Newcastle University
1, Urban Sciences Building
Science Square, Newcastle upon Tyne, UK
orcid.org/0000-0001-7310-4191

Jennifer Hallinan
BioThink
Brisbane, Australia
orcid.org/0000-0002-2860-1022

***Abstract*—The use of microorganisms for the production of industrially important compounds and enzymes is becoming increasingly important. Eukaryotes have been less widely used than prokaryotes in biotechnology, because of the complexity of their genomic structure and biology. The Yeast2.0 project is an international effort to engineer the yeast *Saccharomyces cerevisiae* to make it easy to manipulate, and to generate random variants using a system called SCRaMbLE. SCRaMbLE relies on artificial evolution *in vitro* to identify useful variants, an approach which is time consuming and expensive. We developed an *in silico* simulator for the SCRaMbLE system, using an evolutionary computing approach, which can be used to investigate and optimize the fitness landscape of the system. We applied the system to the investigation of the fitness landscape of one of the *S. saccharomyces* chromosomes, and found that our results fitted well with those previously published. We then simulated directed evolution with or without manipulation of SCRaMbLE, and revealed that controlling the SCRaMbLE process could effectively impact directed evolution. Our simulator can be applied to the analysis of the fitness landscapes of any organism for which SCRaMbLE has been implemented.**

Keywords— SCRaMbLE, yeast, simulation, fitness landscape

I. INTRODUCTION

The Synthetic Yeast 2.0 (Sc2.0) project is an international effort aimed at engineering a eukaryotic genome, that of the Baker's yeast *Saccharomyces cerevisiae*. The project involves eleven institutions from five countries. *S. cerevisiae* is widely recognized as a model organism, is generally regarded as safe, and hence has been studied in considerable detail and is extensively used in industry (Strathern, Jones and Broach, 1982). It is therefore an ideal organism for the genome-scale engineering of a eukaryote (Giaever *et al.*, 2002). The ultimate aim of Sc2.0 is to reconfigure the yeast genome in such a way that it is easier to understand and manipulate, using procedures including the deletion of all known genome destabilizing elements (transposons and sub-telomeric repeat regions); the insertion of symmetrical *loxP* recombination sites (loxPsym) immediately downstream of all non-essential genes; conversion of rarely used stop codons, such as TAG, to the major stop codon TAA, to free up a codon; the watermarking of all protein coding sequences by synonymous base changes, so that they can be identified as synthetic genes by PCR amplification; the removal of all tRNA genes; and the removal of the majority of the 250 introns.

The insertion of the loxPsym sites is of particular importance, since these sites become the locations of genome reshuffling upon the addition of Cre recombinase (Shen *et al.*, 2016). This system is known as SCRaMbLE: Synthetic Chromosome Rearrangement and Modification by loxPsym-mediated Evolution. The loxPsym sites themselves are too short, at only 34 bp, to participate in homologous recombination, so the SCRaMbLE system is only induced by the addition of Cre recombinase (Dymond and Boeke, 2012). When the SCRaMbLE system is induced, not all loxPsym sites will be activated. The stretch of DNA between two active loxPsym sites is referred to as a segment, and may include several ORFs.

The ability to generate multiple variations from a wild-type chromosome, via insertion, deletion, translocation, or inversion of existing genes, means that it is possible to produce thousands or millions of novel genomes. Most of these genomes will, of course, be non-functional, and the Sc2.0 project aims to use directed evolution to select colonies with desirable characteristics. On solid medium, a primary metric for fitness *in vitro* is colony size. Growth in liquid media can also be measured. Fitness *in vitro* is often measured as the ability to produce a substance at enhanced levels.

Directed evolution can be an efficient approach to the identification of desired variants of a wild-type organism. However, by applying only directed evolution, many interesting and potentially useful genotypes will be missed. Further, directed evolution is a time-consuming and wasteful process, which cannot fully explore the genomic richness generated by the SCRaMbLE system.

There is a large body of research on the interaction between evolutionary processes and the fitness landscape generated by individuals in a population (Kauffman and Levin, 1987; Earl and Deem, 2004; Pitzer and Affenzeller, 2012; De Visser and Krug, 2014; Ueda, Takeuchi and Kaneko, 2017). Evolution has been shown to occur more efficiently—that is, more of the possible phenotypes are explored in a shorter time—upon a relatively smooth fitness landscape than on a jagged surface, in which the fitness of one individual is largely unrelated to that of an individual close in genotype (Kvitek and Sherlock, 2011). At present, this more theoretical view of the potential of the SCRaMbLE system is largely ignored, the assumption being that if enough recombinant chromosomes are generated, individuals with desired phenotypes can be identified via screening.

Computational modeling and analysis of the fitness landscape generated by the SCRaMbLE system offers the prospect of identifying system parameters which can produce a smooth fitness landscape, hence improving the efficiency of the directed evolution process, and of improving our understanding of the biology of an important model eukaryote. The stochastic nature of an evolutionary algorithm, combined with genome-wide data on mutations, means that multiple runs can explore different areas of the evolutionary landscape.

In this paper we report the development of a computational model of the SCRaMbLE system, and the fitness landscape generated thereby. The model was parameterized using both genome-scale experimental mutant data and a computational yeast metabolic model. Each set of recombined chromosomes generated from a single chromosome by the SCRaMbLE system is considered as comprising a genetic landscape. Each newly generated chromosome has a genetic distance, *D*, from its parent and all other chromosomes in the population, and a fitness, *f*. By combining these two metrics, a fitness landscape can be constructed and explored. This model allows different configurations of a SCRaMbLEd chromosome to be explored.

II. METHODS

A. Algorithm

Because of the extensive processing and modularization of yeast chromosomes, as described briefly above, it is reasonable to consider each chromosome as a linear vector of genes. We chose to simulate the synthesized right arm of Chromosome IX (synIXR), because it is relatively short (Shen *et al.*, 2016), allowing detailed manual checking of the results, and because *in vitro* data from SCRaMbLE experiments is available for this chromosome, permitting comparison of simulation results with laboratory data. Each simulated SCRaMbLEd chromosome was considered to by hosted in a *strain*, equivalent to an agent in modelling terms.

Although only three ORFs of synIXR are related to the metabolic model, affecting the validation of the fitness function, the accessible experimental data of synIXR SCRaMbLEing facilitated validation of the SCRaMbLE probability (Section Parameterisation).

The simulated chromosome was initialized as a list of strings of 43 segments, reflecting the relative position of each segment in the targeted chromosome. Segments were separated by the loxPsym site immediately after every non-essential ORF. The probablity of a Cre recombinase binding to a breakpoint is *scrProb*. A pseudo-random number generator was used to determine whether a Cre recombinase bound to a breakpoint.

B. Distance metric

The genetic distance between each pair of chromosomes in the population was calculated using the Levenshtein distance (Levenshtein, 1966). This distance metric measures the number of edits required to convert one string into another, using insertion, deletion, or substitution. The Levenshtein distance can be considered to be a measure of the similarity between evolved chromosomes.

In order to determine which of the mutation operations are optimal at any point in the chromosome, we need a cost value for each modification. If we were working with DNA sequence information, the use of a substitution matrix, such as PAM (Schwartz and Dayhoff, 1978) or BLOSUM (Henikoff and Henikoff, 1992), might be appropriate. However, in this simulation, we handled segments as indivisible units, as dictated by the SCRaMbLE system, so all operations were assigned an equal weight of 1.0. This weighting assumes that all operations are equally likely, an assertion which could be modified in the light of experimental data.

Fitness is an abstract concept, and is very difficult to calculate in practice. The fitness function used here was based on two types of data: single gene deletion/overexpression fitness data (experimental fitness) and flux balance analysis results (FBA fitness).

An *in vitro* project previously described used colony size as a measure of fitness (Yoshikawa *et al.*, 2011); SCRaMbLEd chromosomes which produce colonies at least equal in area to those of the wild-type are considered to be fit. However, this approach is clearly infeasible for a simulated system. Another important concept related to fitness is gene essentiality. Of the entire genetic complement of an organism, only some genes are essential for life (Giaever *et al.*, 2002). However, this concept is also fraught with difficulty, particularly for unicellular organisms. Which genes are essential depends largely upon the environment, and an organism grown in rich media is likely to require fewer genes for survival than one grown in minimal media.

Genes do not act in isolation; a gene may be essential only in the absence of one or more other genes (Ulitsky and Shamir, 2007). Synthetic lethality occurs when either of two genes is sufficient for viability alone, but the organism becomes inviable when both genes are knocked out (Ooi *et al.*, 2006). Although most of the research into synthetic lethality has been performed in the context of two-gene interactions, most genes and their products interact with multiple other genes and gene products, in a plethora of ways (Weile *et al.*, 2012). There is a very large body of research into complex genetic networks and their robustness or otherwise in the face of internal and environmental challenges, based largely upon the work of Paul Erdös in the 1950s (Erd, 1959), but blossoming in a genomic context in the early 2000s (Strogatz, 2001; Barabasi and Oltvai, 2004; Farkas *et al.*, 2011), and to which we have contributed (Hallinan, Misirli and Wipat, 2010; Hallinan, James and Wipat, 2011; James *et al.*, 2014). However, because of the nebulous nature of gene essentiality, and the preliminary nature of this work, we chose to apply a naïve definition of essentiality. For our chromosome, genes were deemed to be essential if they were identified as such in any description in the Saccharomyces Genome Database (Cherry *et al.*, 2012). For example, of the 43 segments on synIXR, 7 ORFs (YBL112C, YIR006C, YIR008C, YIR010W, YIR016W, and YIR023W, located on segments 2, 7, 9, 10, 12, and 20 respectively) were essential, and the segments carrying them were identified as essential using this criterion. Any SCRaMbLEd chromosome not carrying all seven essential genes was deemed to be non-viable.

Whilst the absence of essential genes results in the failure of the cell to grow under certain conditions, some genes, especially those encoding enzymes which carry out key processes in metabolic networks, only result in a

reduced growth rate when absent. The contribution of these enzymes, and therefore their genes, can be modelled using genome scale metabolic modelling. Flux balance analysis (FBA) is a common approach to the simulation of metabolic networks (Orth, Thiele and Palsson, 2010). An FBA model determines the flow of metabolites through a given metabolic network, and can be used to predict the growth rate of an organism under a given growth regime. In this work we reconstructed the yeast metabolic network for each of the SCRaMbLEd chromosome variants, taking into account deleted and duplicated enzyme-encoding genes.

The FBA-related fitness was based on the latest consensus yeast metabolic model (Lu *et al.*, 2019), with a constraint file created from simulated SCRaMbLEd results in which ORFs on deleted segments are set to 0. The fitness of the SCRaMbLEd genomes, F_s, was then calculated by running a flux balance analysis, using FlexFlux software (Marmiesse, Peyraud and Cottret, 2015; Lu *et al.*, 2019). The results were normalized with respect to the wild-type fitness, *Fw*, which was calculated by running the no-constraint FBA of the original yeast metabolic model (Eq. 1).

$$Normalized\ FBA\ Fitness = \frac{Fs-Fw}{Fw} \quad (1)$$

Both the single gene deletion fitness data and the single gene over-expression fitness data were obtained from publications (Yoshikawa *et al.*, 2011). We analyzed the distribution of growth rates in the set of mutants from the experimental data, including the deletion and duplication data, to determine how these data could be used to parameterize the fitness function. The variability within the deletion and duplication datasets was found to be high, while the variability between those two data groups was relatively low, with both datasets showing a high proportion of mutants, peaking at a fitness of 0.3-0.35 as measured by the growth rate (Fig. 1). The culture media used for both groups were similar, but differed slightly due to the strategies used for the selection of mutated strains. Using the hypothesis that the fitness distributions of deletion and duplication mutations are similar, we applied quantile normalization to the fitness score of the deleted and duplicated ORFs in the two datasets (Fig. 1). Each quantile normalized fitness score was designated f.

Using this approach, given a SCRaMbLEd genome, an experimental fitness score (EFS) could be calculated by averaging the normalized fitness scores of the deleted or duplicated ORFs.

$$EFS = \bar{X} \quad (2)$$

where x is the normalized fitness of a mutated ORF. Here, x was obtained by comparing f with the median value (0.318 for deletion fitness and 0.295 for duplication).

$$x = f / 0.318 \text{ or } x = f / 0.295 \quad (3)$$

A comprehensive fitness value was then calculated by multiplying the FBA fitness and the EFS. ORFs not included in the metabolic model or wet-lab data were considered to not affect the fitness.

$$Fitness = EFS * normalized\ FBA\ Fitness \quad (4)$$

Overall, the fitness of a strain is calculated by the following equation. Apprently, the fitness of a wildtype strain is 1.

$$Fitness = \bar{X} * \frac{Fs-Fw}{Fw} \quad (5)$$

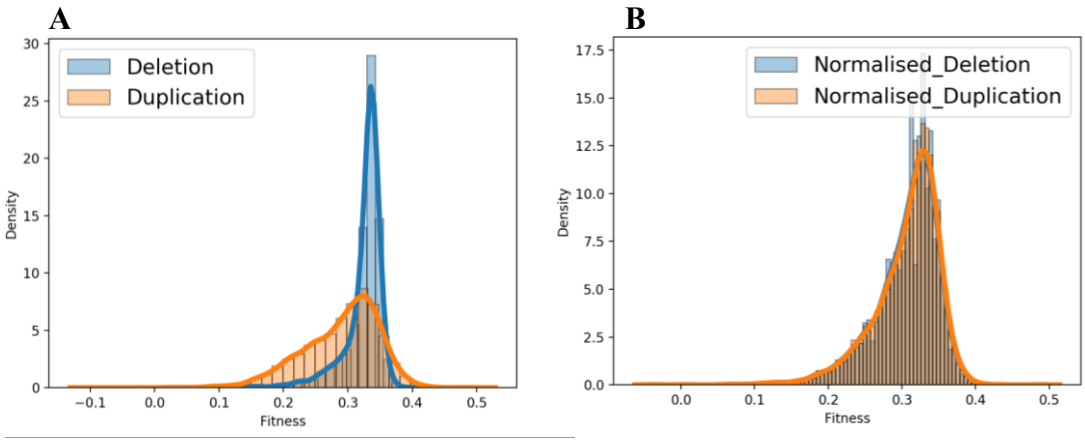


Fig. 1. A) Fitness of a systematically mutated set of yeast strains obtained experimentally by Yoshikawa et al. (2011). There is high variability within the deletion and duplication datasets, but relatively low variability between the datasets, allowing the application of quantile normalisation to both datasets. B) Quantile normalised distribution of single ORF deletion and duplication data. Density: probability density of muations with specific fitness in the dataset; Fitness: growth rate (1/h).

C. Flux balance analysis using FlexFlux

We developed an algorithm to incorporate flux balance analysis when determining fitness values. The algorithm was implemented in Java, and relies on FlexFlux, a steady-state based metabolic network research tool for flux balance analysis (Marmiesse, Peyraud and Cottret, 2015). FBA models were represented in the Systems Biology Markup Language (SBML) (Hucka and others, 2003). The implementation takes a list of deleted genes from the Chromosome class and runs FlexFlux to simulate gene knockout on the latest yeast consensus SBML genome-scale model, yeast_8.3.5 (Lu *et al.*, 2019). First, it creates a text-based constraint file which contains an objective function for maximizing the biomass. Next, the unique IDof every deleted ORF is obtained from the SBML model file. These

IDs are written into the constraint file, and their status is set to "0" to represent deletion. Finally, FlexFlux is called with the constraint file and generates a result document. Some of the deleted ORFs might not be included in the SBML model, indicating that such ORFs are not involved in the well-understood metabolic network. In these cases, these ORFs are not written into the constraint file. If none of the deleted ORFs are included in the SBML model, the method returns wild-type fitness.

D. Parameterisation

The recombinase protein Cre randomly binds to a loxPsym site and initiates SCRaMbLEing We simulated this process using a parameter, scrProb, the probability of a Cre protein binding to a loxPsym site and triggering deletion or duplication. scrProb was estimated based on experimental data. On average, for chromosome synIXR, around six SCRaMbLE events occurred following four hours of induction with 1 M estradiol (Shen *et al.*, 2016). Using this information, we estimated the probability of an event, using a simple simulation to investigate the correlation between *scrProb* and the average number of SCRaMbLE events.

For every scrProb range from 0 to 1.0, with a gap of 0.01, SCRaMbLEing on synIXR was simulated with a pool of 1,000 strains. The number of survivals and the average number of SCRaMbLE events of the 1,000 SCRaMbLEd strains are shown in Fig.s 2 and 3.

The simulation results indicated that when *scrProb* was around 0.3, the number of SCRaMbLE events was about six (Fig.s 2), which is the average number of SCRaMbLE events of surviving strains identified from the experimental data (Shen *et al.*, 2016). *scrProb* could be validated using additional experimental results, which are not currently available. With different *scrProb*, the survival rates of SCRaMbLE strains were different. If *scrProb* = 0.2, the survival rate was 37/1000; while with *scrProb* = 0.4, the survival rate was 10/1000. Hence, given further data about the survival rate, which could be obtained by running a simple wet-lab experiment comparing colony numbers between a SCRaMbLEd culture and a negative control, we could produce a more accurate estimate of the probability of a loxPsym site being involved in a SCRaMbLE event. In this work, we set the *scrProb* to 0.3.

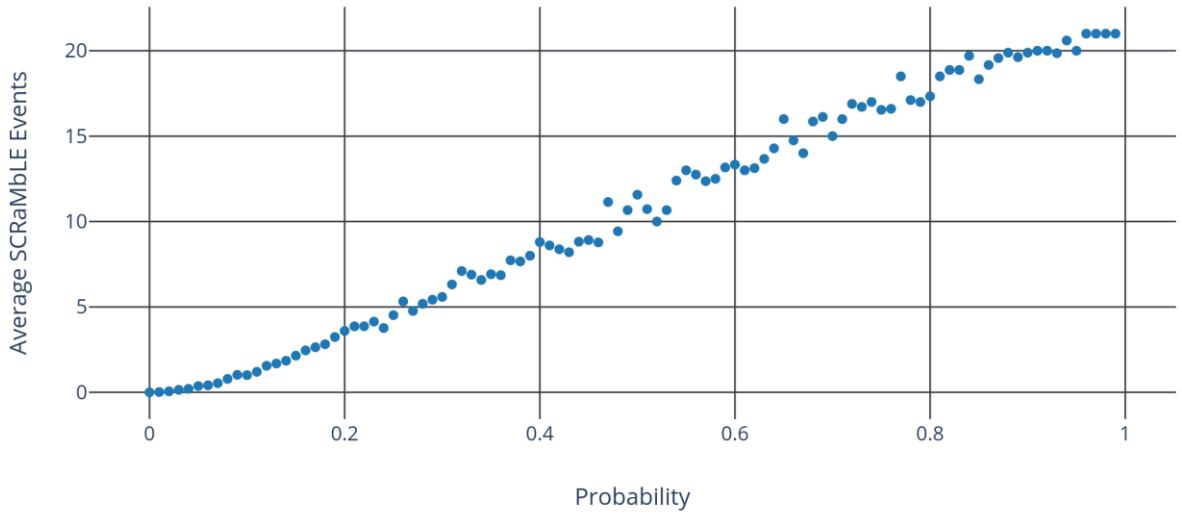


Fig. 2. Single point breaking probability versus the average number of SCRaMbLE events. When the probability of a Cre binding to a loxPsym was around 0.3, the average number of SCRaMbLE events in chromosome synIXR was about six per surviving strain, which is the average number of SCRaMbLE events determined *in vivo* (Shen *et al.*, 2016).

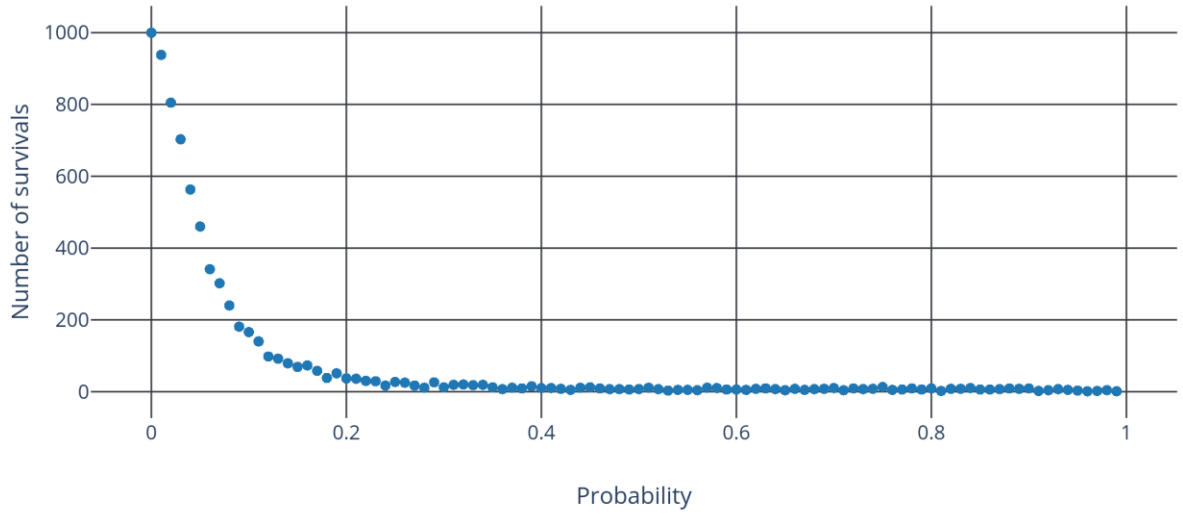


Fig. 3. Single point breaking probability versus the number of survivors of SCRaMbLEing strains with synIXR. The number of strains was set as 1000 before SCRaMbLEing. When the probability of a loxPsym experiencing a SCRaMbLE event was around 0.2 to 0.4, the survival rate was between 37‰ and 10‰.

E. Fitness landscape analysis

We simulated 4,280 strains using a pseudo random number generator. The resulting dataset was used for fitness landscape analysis and further investigation.

Chromosome synIXR has 43 segments, making it difficult to visualize. We therefore used a dimension reduction algorithm, t-Distributed Stochastic Neighbor Embedding (t-SNE), to convert the 43-dimensional input into a two-dimensional array representing the genotype of every genome in the simulation (Hari and Lobo, 2020). t-SNE is a non-linear dimension reduction algorithm, and is implemented by minimizing the Kullback-Leibler divergence between two similarity distributions: the pairwise similarities of high dimensional data points, and the corresponding low-dimensional embedded output points (Van Der Maaten and Hinton, 2008).

The two-dimensional array produced by t-SNE was used as the *x* and *y* axes of the fitness landscape, with the fitness score of each strain as the *z* axis. A tunable parameter of t-SNE, *perplexity*, balances the attention between local and global data by estimating the number of close neighbors of each point in the landscape. t-SNE was optimized by comparing the results produced by our simulator, resulting in a *perplexity* value of 40.

III. Results

In this work we used an *in silico* evolutionary approach to develop a model of the evolution of a population of yeast mutants whose genomes were perturbed using the SCRaMbLE system. We validated the model of deletions by comparing *in silico* and *in vivo* ORF deletions in mutant populations, validated the fitness function by reference to experimental data, and finally analysed the fitness landscape of the populations generated *in silico*, using the model.

A. Deletion patterns in SCRaMbLEd genomes in silico and in vivo.

To evaluate whether the results of the SCRaMbLE simulation were comparable with the wet lab data, we ran a simulation investigating deletion patterns on the circular chromosome synIXR. A random simulation dataset with 80 surviving *in silico* strains was generated, and compared with a wet-lab experimental dataset with 64 surviving *in vitro* strains (Shen *et al.*, 2016) (Fig. 4).

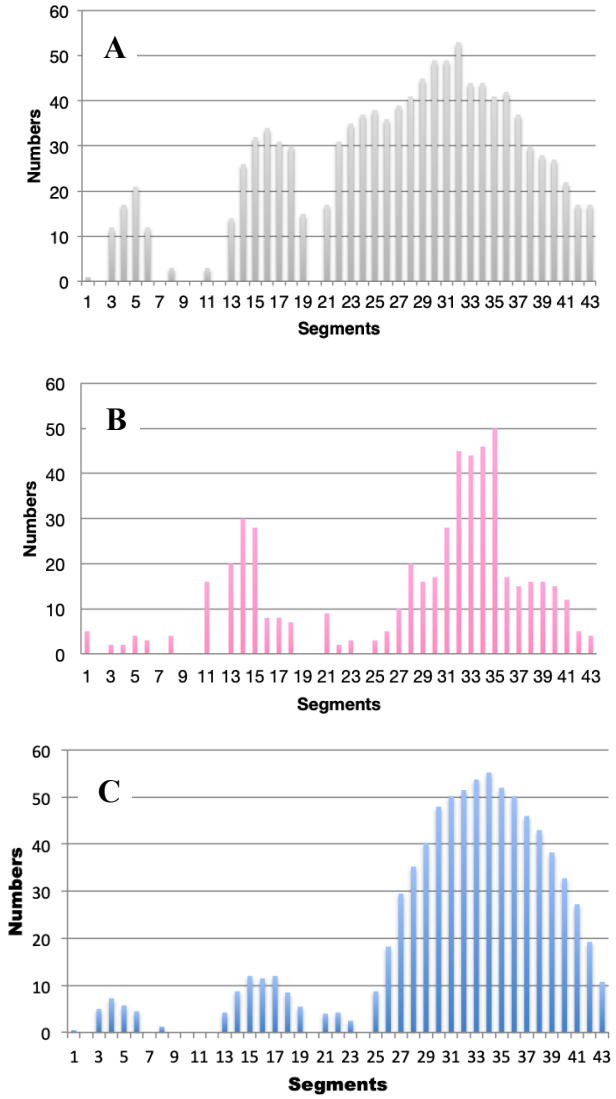


Fig. 4. A) Deletion patterns in simulated *in silico* strains, with a deletion probability of 0.2. B) Deletion patterns obtained *in vivo* by Shen et al. (2016) . C) Simulated deletion patterns with Segment 24 as an essential segment. The Y axis represents the number of SCraMbLEd strains with a specific segment deleted; The X axis represents segments. Each segment is the basic unit of SCRaMbLE, flanked by loxPsym sites.

Comparing Fig. 4A and 4B, we observed similar deletion patterns. Both experimental and simulation data have two deletion patterns, ranging from Segment 13 to Segment 18, and from Segment 20 to Segment 43, respectively. The peaks of these deletion patterns are similar, with around 30 deletions for pattern Segment 13-18 and around 50 deletions for pattern Segment 20-43. However, for the simulation results, the second pattern, between Segments 20 and 43, was much smoother than its experimental counterpart. Combining the fitness score of the simulations by isolating strains with a relatively high fitness score might produce a different perspective, with results more similar to those of the real data. However, since the relevant fitness data was not published with the experimental deletion patterns for synIXR (Shen *et al.*, 2016), we could not make this comparison. The number of deletions of ORFs between Segments 21 and 31 was much higher in the simulation than in the wet lab results. Further data about gene functions, from the Saccharomyces Genome Database (SGD) (Chervitz *et al.*, 1999), suggested that null mutants of Segment 24 (carrying ORF YIR026C) decreased the competitive fitness of growth rate, which may explain the difference described above (Fig. 4).

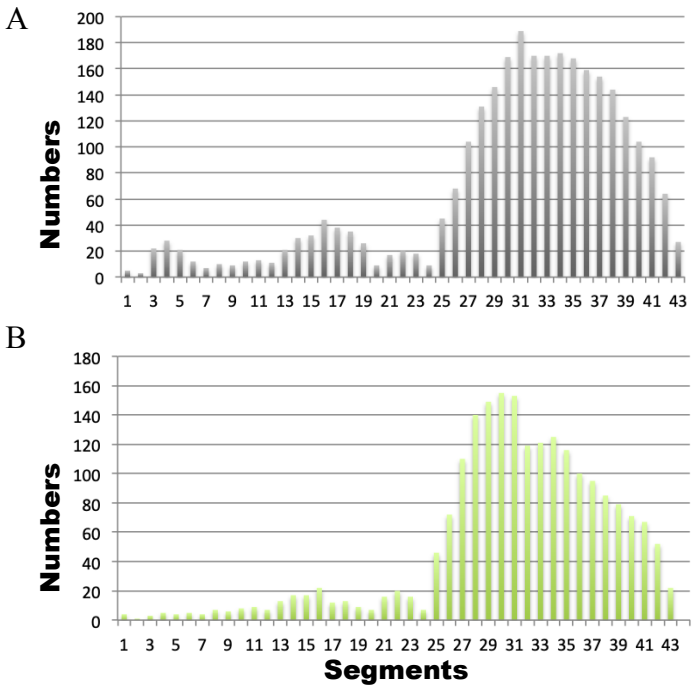


Fig. 5. A) *In silico* deletion patterns of all genomes, and B) high fitness genomes. The Y axis indicates the number of genomes with a related deleted segment in a dataset generated by SCRaMbLE simualation . The most significant difference between all strains and the high fitness strains was observed in the deletion patterns between Segments 3 and 6, which were absent in the high fitness results, suggesting that the ORFs on these segments are related to high fitness.

TABLE I. *IN SILICO* FITNESS VERSUS EXPERIMENTAL FITNESS FOR THREE DELETION MUTANTS AND THE WILD-TYPE STRAIN. THE FITNESS FUNCTION DESCRIBED IN EQ4 WAS USED TO CALCULATE THE FITNESS SCORES OF THE SIMULATIONS.

Strain	Simulation Fitness	Experimental Fitness
Wild-type	1.0	Wild-type
ΔYIR004W	0.886	0.921
ΔYIR005W	0.797	0.783
ΔYIR020C	0.977	1.0

According to the findings of Deutschbauer and co-workers, the ORF YIR026C is vital for strains competing with each other, due to the decreased competitive fitness of the null mutant (Deutschbauer *et al.*, 2005). This observation indicates that Segment 24 is non-essential when there are no other strains competing against it. However, due to its weak competency, the YIR026C null mutant could not survive in competition with other scrambled strains, resulting in the deletion patterns being shifted from Segment 34 to Segment 30. Thus, YIR026C on Segment 24 is an essential ORF in a multicellular consortium. Adding Segment 24 as an essential unit in the simulation produced results similar to the wet-lab results (Fig. 4C).

Strains with a fitness score higher than that of the wild-type strain were selected for further analysis. The most significant differences between high fitness strains was observed in the deletion patterns between Segments 3 and 6, which were absent from the chromosomes of the high fitness strains, an observation which suggests that the ORFs on these segments are associated with higher fitness (Fig. 5).

Together, these results suggest that the SCRaMbLE simulator models deletion events with reasonable accuracy. Since the simulation is based on random numbers, these simulation results provide further evidence that the SCRaMbLE deletion process is largely random, but is constrained by its metabolic and phenotypic effects on the resulting mutant strains.

B. Validation of the fitness function

To validate the final fitness function (Eq. 4) used for the evolutionary process *in silico*, we calculated the fitness of the three single deletion mutant strains: YIR004W on Segment 5, YIR005W on Segment 6, and YIR020C on Segment 18 (Table 1). These ORFs, which are supported by experimental evidence ($p < 0.05$), are all on the synIXR chromosome (Shen *et al.*, 2016). These results (Table 1) were consistent with experimental results.

C. Fitness distance correlation

Fitness distance correlation (FDC) is usually used for optimizing genetic algorithms and analyzing the ruggedness of fitness landscapes. We applied the technique to the

analysis and comparison of the fitness landscapes generated by the simulated SCRaMbLEd yeast strain populations. FDC samples data points on the fitness landscape, and calculates the correlation between the measured fitness and the distance to the global optimal fitness. We used this approach to investigate the topology of the in silico landscape, and for studying the shape and size of the evolutionary search space.

Fitness Dataset 1, with 4,280 strains generated for fitness landscape analysis was used here (Methods Section E). We also constructed a smaller dataset, Fitness Dataset 2, derived from Fitness Dataset 1 by removing strains with mutated genes whose products were not modelled in the metabolic network. When the FDC value ws between -0.15 and 0.15, optimization was difficult, because the fitness landscape was very rough. The FDC coefficient of the 4,280 simulated scrambled genomes in Fitness Dataset 1 was 0.07. However, the FDC rose slightly to 0.09 for FDC Dataset 2, which only included strains with fitness in which a contribution from the FBA contributed to the landscape. Scatter plots (Fig. 6 and Fig. 7) show the structure of FDC Dataset 1 and FDC Dataset 2. For FDC Dataset 2, a cohort of high-fitness strains with chromosomes srparated by relatively high Levenshtein distance could be observed (Fig. 6). There was no significant structure in the correlation of fitness and distance for FDC Dataset 1, in which all mutants were retained (Fig. 7). The results shown in the scatter plots are consistent with the FDC values. The slight difference between Fitness Dataset 1 and Fitness Dataset 2 is probably because, while the whole fitness landscape is rugged, some patterns still exist in the enzyme encoding-genes-mutated subset, since only three ORFs from synIXR are involved in the metabolic network. The fitness landscape had high ruggedness, based on FDC analysis.

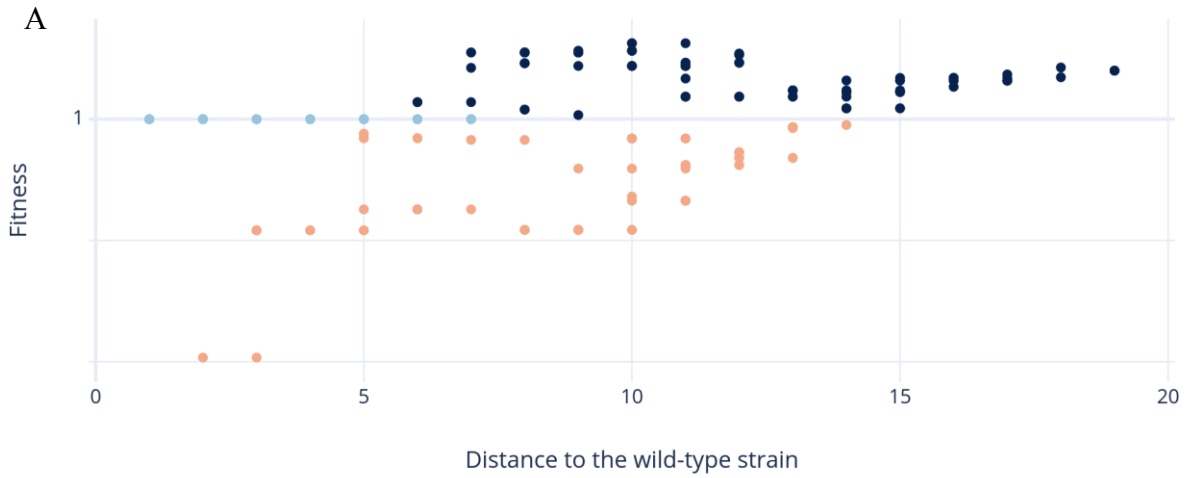


Fig. 6. Scatter plot of the fitness-distance correlation of Fitness Dataset 2, with 237 strains whose mutant gene products featured in the metabolic network. Dark blue: fitness > 1; light blue: fitness = 1; orange: fitness < 1.

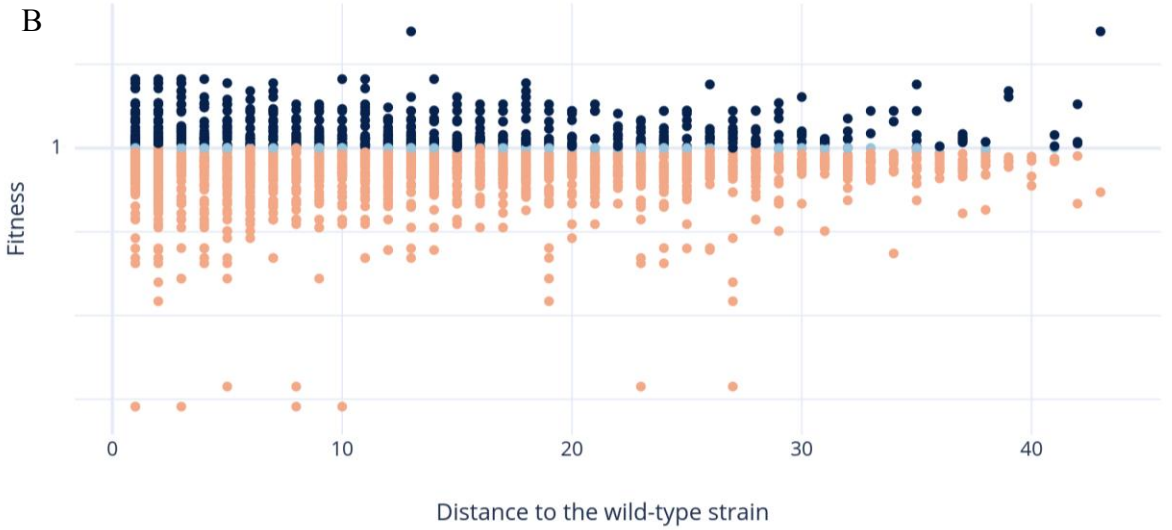


Fig. 7. Scatter plot of the fitness-distance correlation of Fitness Dataset 1 with 4,280 strains with simulated scrambled chromosomes. Mutant strans were included, even if the mutant gene products did not feature in the metabolic model used to calculate the genome fitness, using FBA. Dark blue: fitness > 1; light blue: fitness = 1; orange: fitness < 1.

Fitness Dataset 1 was used for visualizing the fitness landscape (Fig. 8A). Due to the lack of data points for strains with ORFs encoding enzymes, the SCRaMbLE simulator was also used to generate 401 strains with SCRaMbLEd chromosomes with mutated genes whose products contributed to the metabolic network used to evaluate fitness (Fig. 8B). All strains were divided into four groups based on fitness: High, Wild-type, Low, and Dead (not shown). The dimension reduction algorithm t-SNE was used to convert the high dimensional data to two dimensions. The fitness of most of the *in silico* strains was lower than that of the wild-type *in silico* strain. A large number of strains were inviable, due to the deletion of essential genes from their chromosome. In Fig. 8B, clear boundaries can be observed between each group of strains, indicating that there are obvious patterns of chromosomes with mutations in genes contributing to metabolism. Mutations in these ORFs redirect the flux in the FBA model of the metabolic networks, and thus lead to changes in the fitness score. Some ORFs play a key role in the metabolic networks. By altering these key ORFs, the flux of the FBA model changed significantly. Although experimental single-gene deletion and duplication data were integrated into the fitness function (Eq. 4), the fitness scores of strains with and only with mutated ORFs encoding metabolic enzymes were fully dependent on the FBA results of the metabolic model. Since only three ORFs from synIXR were involved in the metabolic model, t-SNE easily captured the key features necessary to distinguish groups with different fitness.

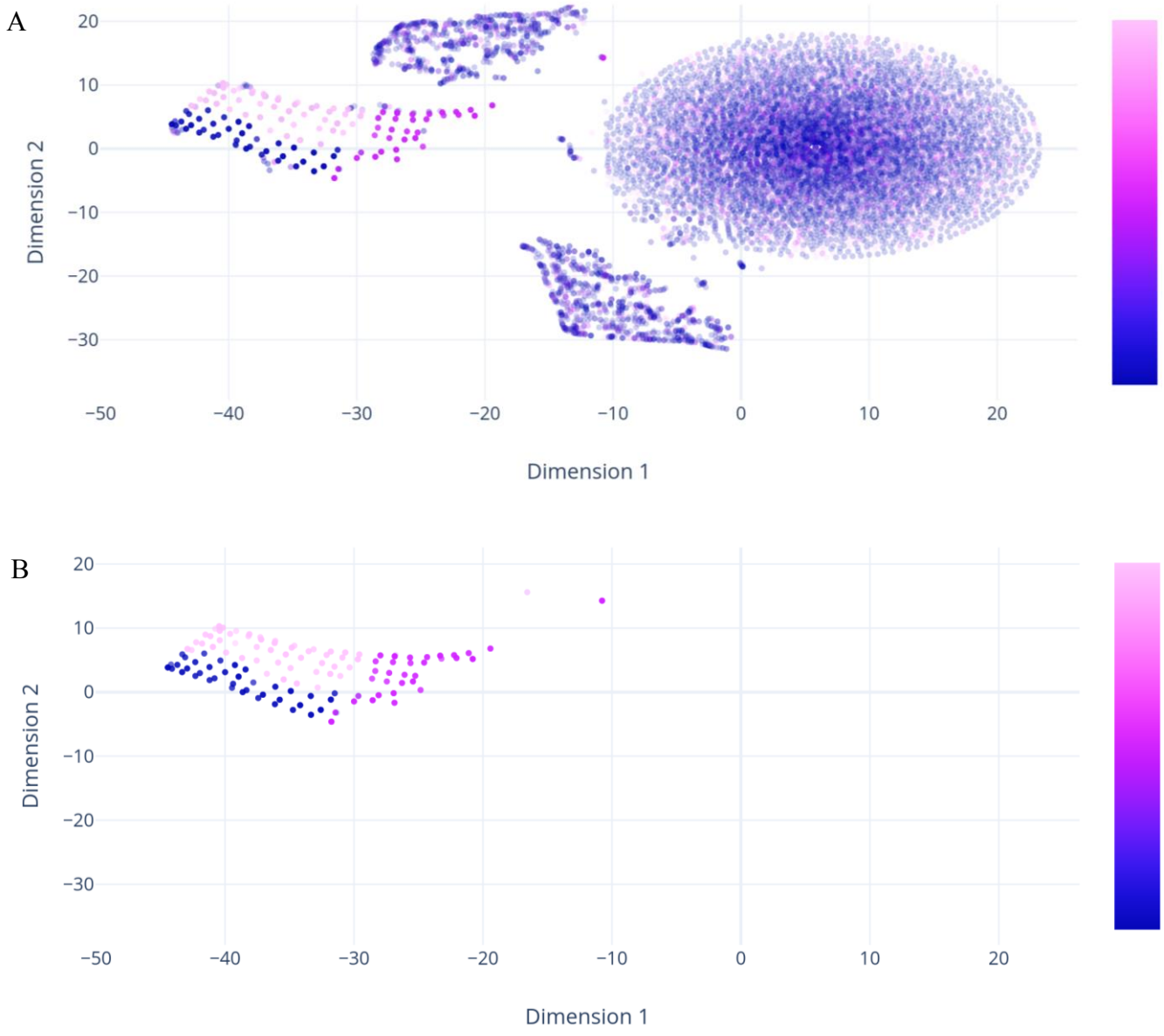


Fig. 8. Visualisation of a fitness landscape of SCRaMbLEing synIXR strains. High fitness (Fitness > Wild-type): Pink; wild-type fitness: purple; low fitness (Fitness < Wild-type). A: the whole landscape of 10000 scrambled strains. There were four clusters in the landscape. No obvious patterns could be observed in most of the fitness landscape. B: 401 scrambled strains with mutated metabolic enzyme-encoded ORFs with clear patterns in terms of fitness score.

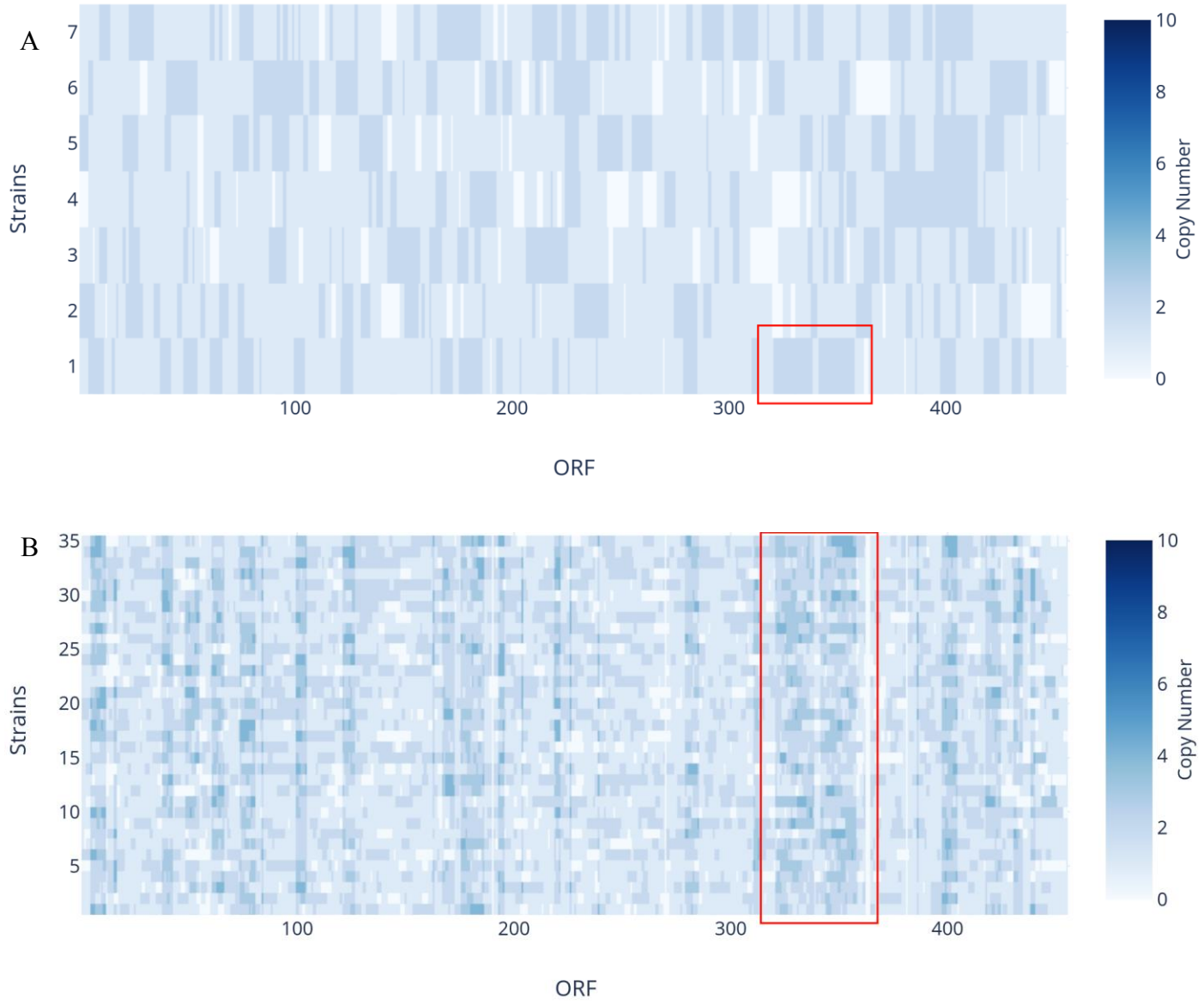
For the other *in silico* strains, including those with non-metabolic-related mutants, no clear patterns regarding fitness were observed (Fig. 8A). These results, together with the results from the FDC analysis, indicated that the fitness landscape of simulated SCRaMbLEd strains showed a high degree of ruggedness, although the enzyme-encoding mutation fitness subset was smoother.

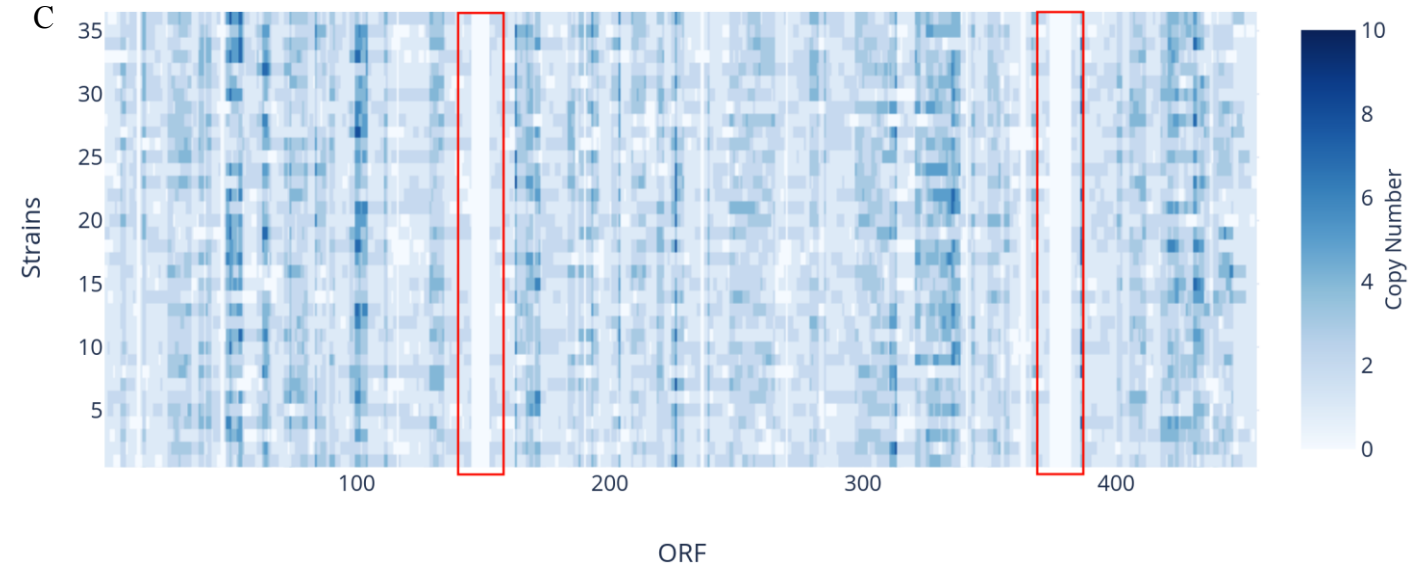
D. Directed evolution of synII

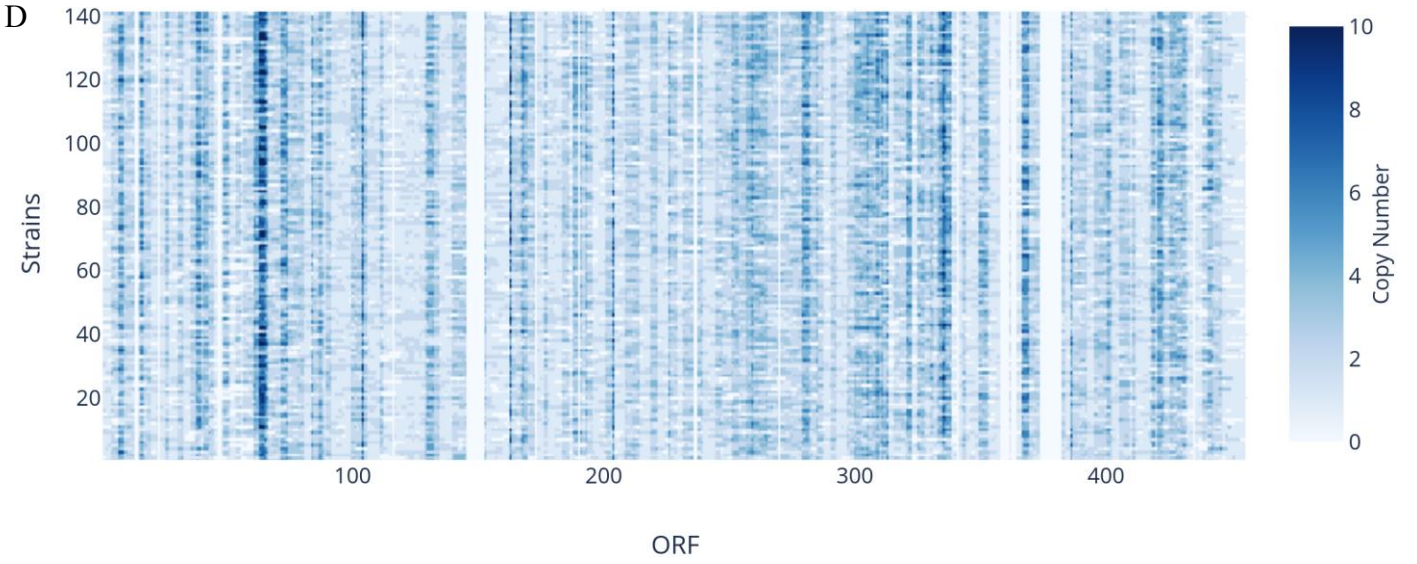
Given the success of the proof of concept simulations of strains with the SCRaMbLEd chromosome synIXR, we explored simulating directed evolution using this modelling approach. Since there were only 43 ORFs on synIXR, which limited the searching space of evolution, we applied the directed evolution mode to another synthetic chromosome synII. SynII has 456 ORFs, which vastly expands the space for directed evolution.

Since synII is a large chromosome, SCRaMbLE was expected to alter essential ORFs with a high probability, leading to more unfit strains with a lower survival rate. For this reason, the number of SCRaMbLEd strains in the simulation was set to 100,000, to ensure that some strains survived. The probability of Cre binding to a loxPsym site, scrProb, was set to 0.3 as for the simulations with the synthetic chromsome synIXR. Whilst synII has about 10 times more loxPsym sites than synIXR, the number of Cre molecules in the cell under experimental conditions was still far in excess of the number of loxPsym sites. Thus, scrProb in the simulation of synII SCRaMbLEing should be similar to the scrProb of synIXR SCRaMbLEing.

Five rounds of directed evolution were simulated with the above settings. There were only seven surviving strains after the first round of SCRaMbLE of 100,000 synII strains, due to the deletion of essential ORFs (Fig. 9). However, the fifth round of SCRaMbLE resulted in 223 survivals from the same number of SCRaMbLEd strains. The major reason was that the chromosomes of these strains gradually gained duplicated copies of essential ORFs during successive rounds of SCRaMbLE, which enabled a strain to survive a deletion event.




C
Strains
35
30
25
20
15
10
5
100
200
300
400
ORF
Copy Number
10
8
6
4
2
0


D
Strains
140
120
100
80
60
40
20
100
200
300
400
ORF
Copy Number
10
8
6
4
2
0

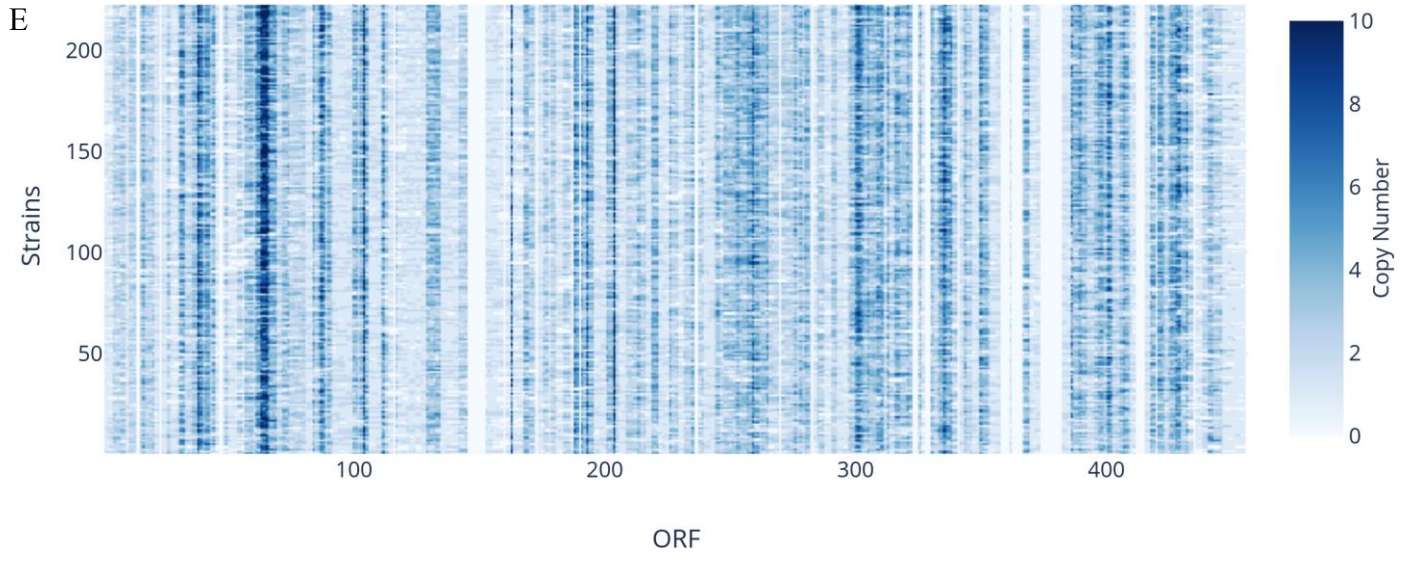

E
Strains
200
150
100
50
100
200
300
400
ORF
Copy Number
10
8
6
4
2
0

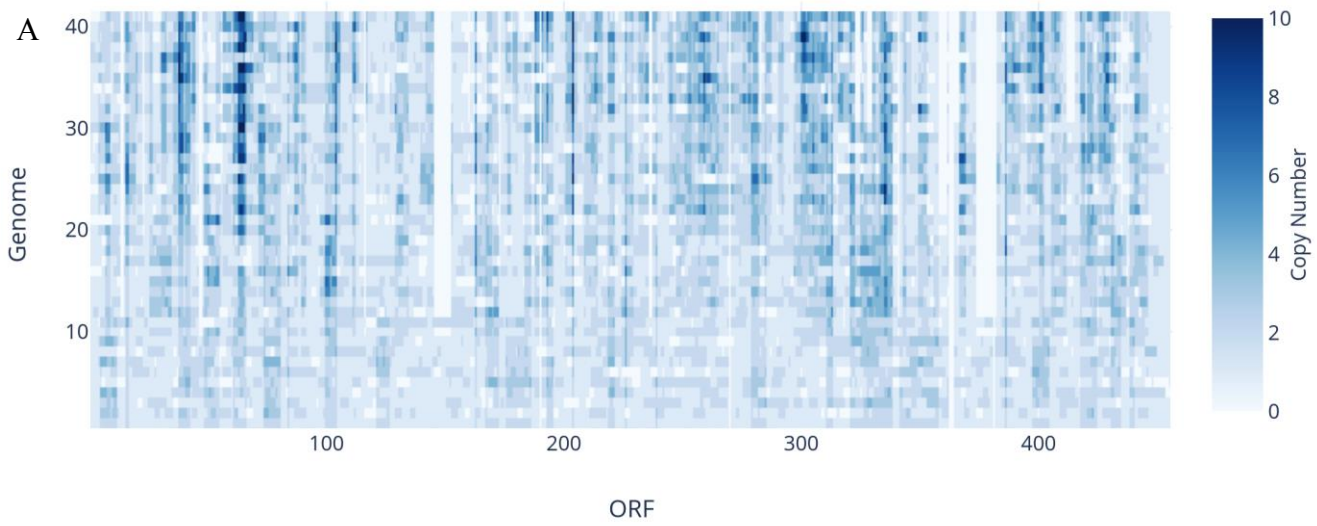
Fig. 9. Heatmap profiles of synII ORFs in five rounds of direcetd evolution. X axis represents the 456 ORFs of synII; Y axis stands for the surviving strains (e.g. there are seven strains resulting from the first round of SCRaMbLE (Fig. 9A)). Directed evolution rounds 1 to 5 are labelled A-E. For each round of directed evolution, the strain with the highest fitness score was picked as the initial strain to be SCRaMbLEd in the following round. In each profile, chromosomes were ranked in descending order of Levenshtein distance from the unscrambled wild-type chromosome (i.e. the bottom chromosomes were more similar to the wild-type chromosome than upper chromosomes). Darker data points represented higher copy numbers, while white areas represent deletion of the corresponding ORF.

Patterns of chromosomal rearrangements could be observed from the profiles of the SCRaMbLED chromosomes (Fig.9). Strain 1 of the first round of SCRaMbLE (Fig. 9A) had the chromosome with the highest fitness score, and was selected as the starting strain for the second round of SCraMbLE. Hence, the chromosomes shown in the heatmap for the second round of SCraMbLE shared similar chromosomal patterns with strain 1. For example, two chunks of duplication (darker zones), inherited from the strain 1 of the first round of SCRaMbLE, lying between ORF300 and ORF400 (marked in Fig. 9A), can also be seen in all daughter strains of the second round (Fig. 9B). However, duplication patterns might not be passed through generations due to potential loss during the directed evolution. While deletion patterns could be maintained if the only copy of an ORF was deleted. For example, the marked deletion pattern from strains of the third round of directed evolution(Fig. 9C) was succeeded in following rounds(Fig. 9D and Fig. 9E).

To obtain further insights from the results of directed evolution, the 10 strains with the highest and lowest fitness at each round were identified, and heatmaps of their chromosomal similarities were produced (Fig. 10). All profiles shared common deletion or duplication patterns, such as the deletion of chromosomal segments between ORF100 and ORF200, and between ORF300 and ORF400. This finding suggested that the fitness landscape is very rugged, so that a minimal alteration to a chromosome could result in a very different phenotype in terms of fitness. Genetic traits from the highest fitness strains were likely to be propagated in subsequent rounds of evolution, while mutations responsible for lowering the fitness of the lowest fitness strains would not be passed on.

Thus, if a pattern of duplication or deletion was established in low-fitness strains and the pattern only appeared in that round of evolution, those mutations may have led to a decrease in fitness. In this case there were some differences (as annotated in Fig. 10B) between the two profiles.

The region indicated by an arrow in Fig. 10B, ranging from ORF49 to ORF54 (YBL071C, YBL070C, YBL069W, YBL068W-A, YBL068W, and YBL067C), was more apparent in the chromosome of low-fitness strains than in high-fitness strains(Fig. 10A). This pattern only appeared on the third round of evolution low-fitness results. This observation indicated that in the third round of evolution an increase in copy number segments ranging from ORF49 to ORF54 may have been linked to a decrease in fitness (Strain 13 to Strain 22).



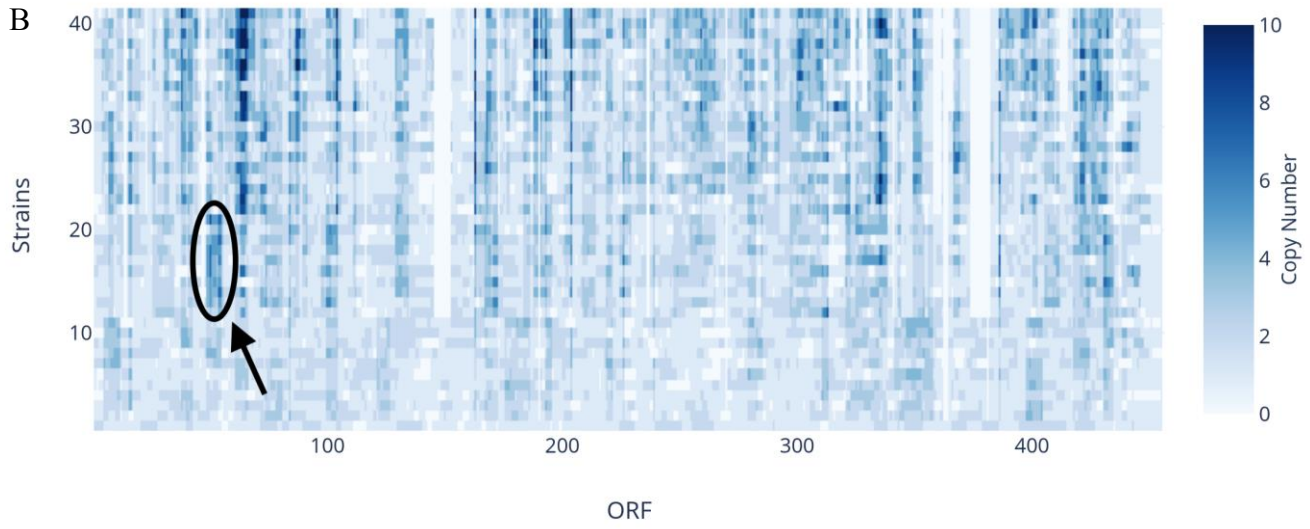


Fig. 10. Chromosomal alignments from synII strains with the highest or lowest fitness from each round of directed evolution. A: From bottom to top: Stacking Round 1 to 5 top highest-fitness strains of each round (only two from Round 1 and ten from Round 2-5). B: From bottom to top: Stacking Round 1 to 5 the lowest-fitness genomes of each round (only two from Round 1 and ten from Round 2-5). The indicated region, ranging from ORF49 to ORF54 (YBL071C,YBL070C, YBL069W, YBL068W-A, YBL068W, and YBL067C), was more apparent in these low fitness strains than in the high-fitness profile(Fig. 10A)

E. Directed evolution with CRISPRi

SCRaMbLE shuffles Sc2.0 genomes in a random way. Whilst, in theory, the whole evolutionary search space could be explored, reaching a specific target is largely based on luck.

To address this problem, we are developing laboratory based genetic tools to improve the ability of controlling SCRaMbLE by altering the probability of occurrence of SCRaMbLE events at ORF sites of interest (unpublished). Regarding blocking specific segments of Sc2.0 from SCRaMbLing, a dCas9-blocking-CRE strategy (or CRISPRi for SCRaMbLE) was designed in previous study by targeting sequences adjacent to corresponding loxPsym sites flanking the segment of interest. Additionally, a novel programmable system called CIRS(CRISPR inspired recombination system) was proposed to improve CRE binding to specific loxPsym sites. Inspired by CRISPR and CIRTS (CRISPR-Cas-inspired RNA targeting system), CIRS uses a redesigned gRNA to direct the gRNA-binding-protein-fused Cre recombinase.

We therefore sought to simulate this system, emulating the effects of targeting CRISPRi and CIRS to block selected recombination events and studying the chromosomal rearrangements in the resulting SCRaMbLEd strains. Here, we describe the simulation results of directed evolution with CRISPRi as an example, and compare them with the previous results of simulation without CRISPRi.

As a proof of concept, a set of ORFs (YBL097W, YBL092W, YBL084C, YBL076C, YBL074C, YBL050W, YBL041W, YBL040C, YBL035C, YBL034C, YBL030C, YBL026W, YBL023C, YBL020W, YBL018C, YBL014C, and YBL004W) was randomly selected as the target of CRISPRi. The probability of Cre binding to adjacent loxPsym sites to these ORFs was set as 0, which represented 100% efficiency of CRISPRi blocking. These ORFs were essential ORFs, and thus the survival numbers were expected to be higher than the previous simulation. Five rounds of directed evolution were performed as for previous simulations. Then the strains with the highest fitness of each round of evolution were selected, along with those from previous simulations without CRISPRi (Fig. 11).

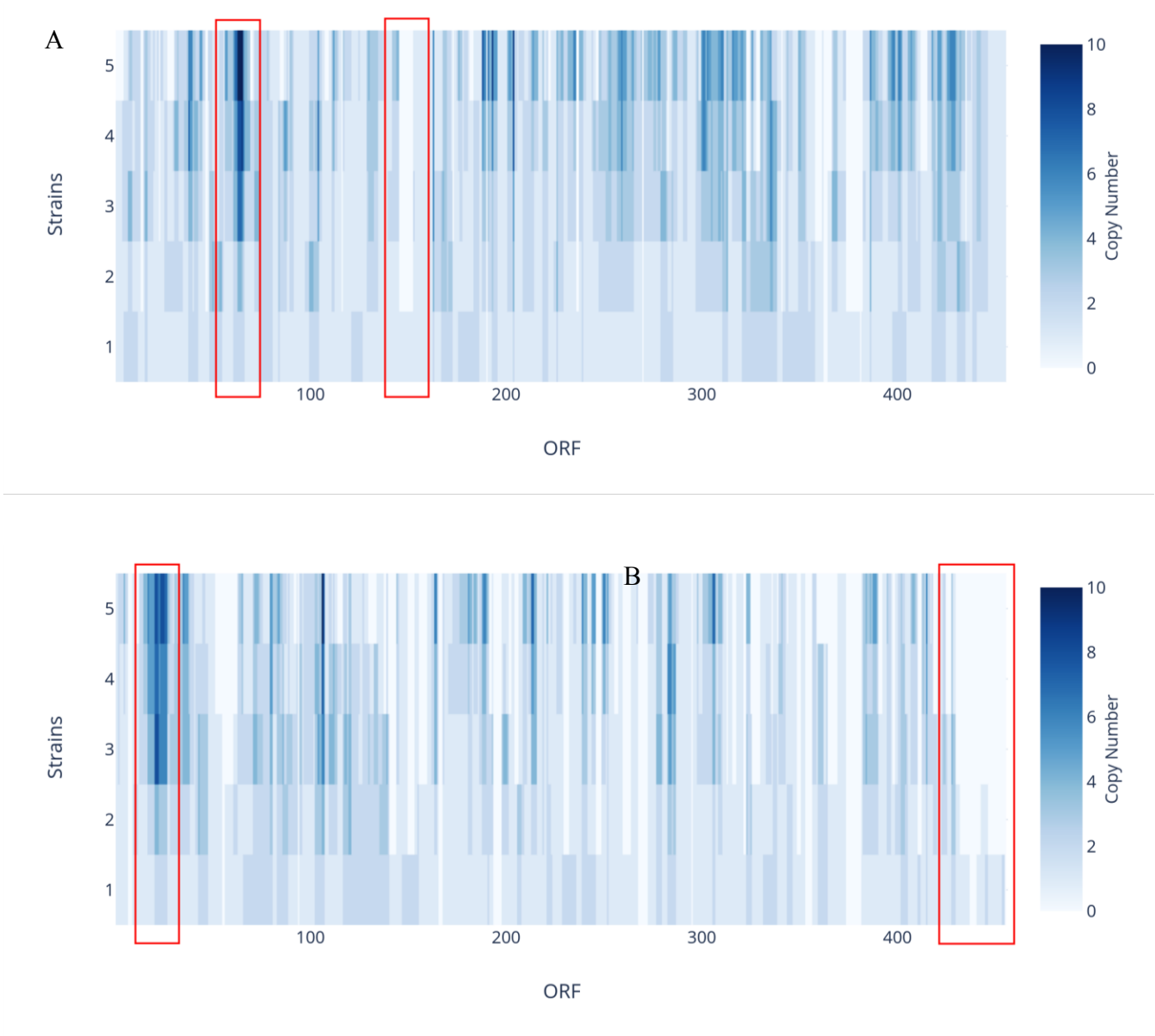


Fig. 11. Profiles of the chromosome from the fittest strains from each round of directed evolution with or without CRISPRi. A: normal directed evolution; B: directed evolution with CRISPRi. Patterns are marked by rectangles.

Patterns of genome conservation could be observed within the comparative heatmaps of the chromosomes from fit strains. For example, in the simulation without CRISPRi, patterns including the duplicated chunk of ORF63-ORF65 (YBL059W, YBL058W, and YBL057C) and the deletion chunk of ORF146-ORF152 (YBR020W, YBR021W, YBR022W, YBR023C, YBR024W, YBR025C, and YBR026C) could be identified (Fig. 11A). In the simulation with CRISPRi, the duplication pattern of ORF17-ORF26 (YBL100C, YBL099W, YBL098W, YBL097W, YBL096C, YBL095W, YBL094C, YBL093C, YBL092W, and YBL091C-A) and the deletion pattern of ORF431-ORF455 were apparent (Fig. 11B). However, although both simulations used the same random seed, the chromosomal profiles of the strains with CRISPRi modified SCRaMbLE did not share many similarities with those of the unmodified SCRaMbLE strains, an observation which suggested that CRISPRi could be used to modify the direction of evolution in the fitness landscape.

With CRISPRi directed evolution, more strains survived during evolution. There were 446 survivals in total for five rounds of normal directed evolution while 846 survivors emerged from the simulation with CRISPRi.

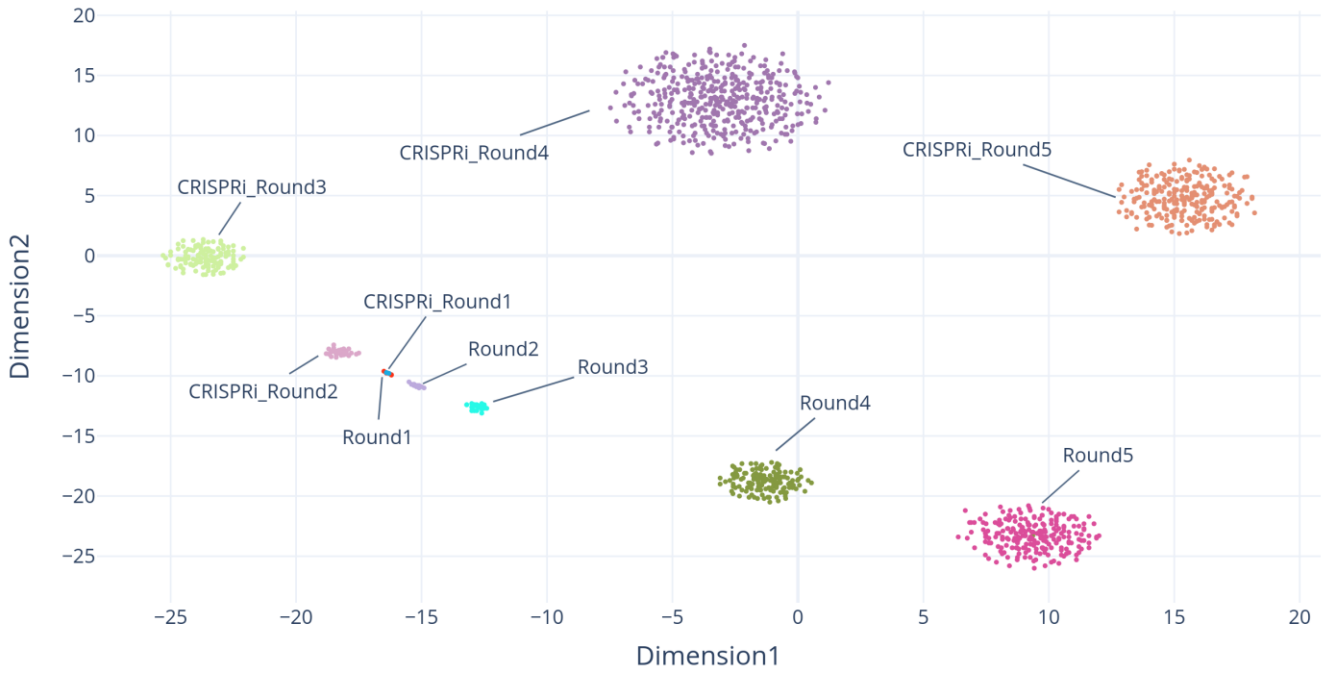


Fig. 12. Chromosomal landscapes of simulated synII SCRaMbLE directed evolution with CRISPRi (CRISPRi_Round 1-5) or without CRISPRi (Round 1-5). X and Y axes represent the dimensions of t-SNE results, whose dimensions were reducted from 456 (number of ORFs on synII) to two. Strains with 456 ORFs were transformed by dimensional reduction to two dimensions. The resulting survivors of each round are indicated with different colors.

To further investigate the ways in which CRISPRi could help explore the fitness landscape, the dimension reduction algorithm t-SNE was applied to both normal simulation and CRISPRi-controlled simulation datasets. The results are shown in Fig.12. The results of the first round of directed evolution, with or without CRISPRi, were all located in the same small area. However, the route of directed evolution diverged after the second round of evolution. This divergence was manifested by the increasing distance between clusters in the landscape. There was a strain from the previous round of evolution in the center of each cluster. Five rounds of normal directed evolution in a t-SNE dimension reduction graph (Fig. 13) showed similar results.

The outliers were the strains with the highest fitness from the previous round, and served to act as the initial strain whose chromosome was SCRaMbLEd in the following round.

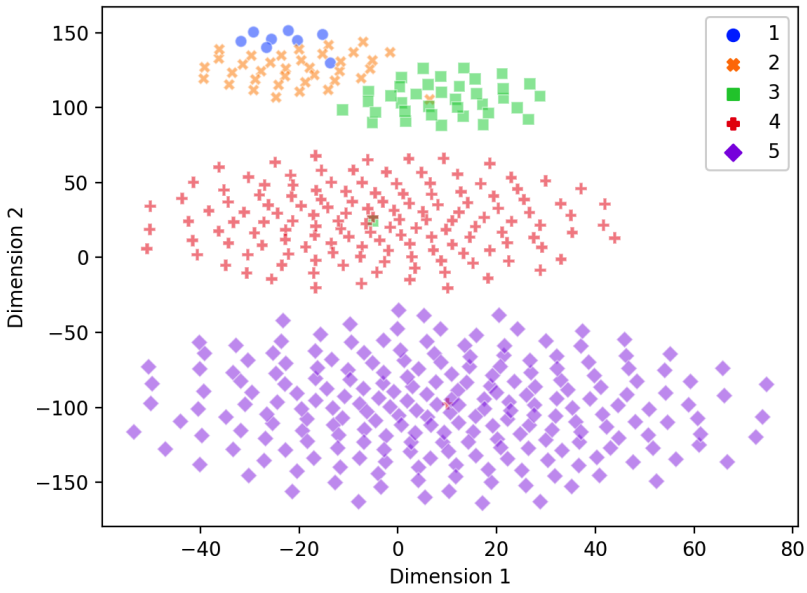


Fig. 13. t-SNE dimension reduction on the chromosomes of strains subject to the standard SCRaMbLE approach. Rounds are labeled as 1 - 5.

The fitness landscapes of both simulation datasets were created based on the same t-SNE dimension reduction results (Fig. 14). No clear patterns could be found in the fitness landscapes, a finding which was consistent with previous findings that the fitness landscapes of SCRaMbLE were rugged. However, red dots, representing high-fitness strains, were apparently more frequent in the clusters of CRISPRi rounds. This finding supported the contention that directed evolution with proper CRISPRi application could achieve high fitness more efficiently.

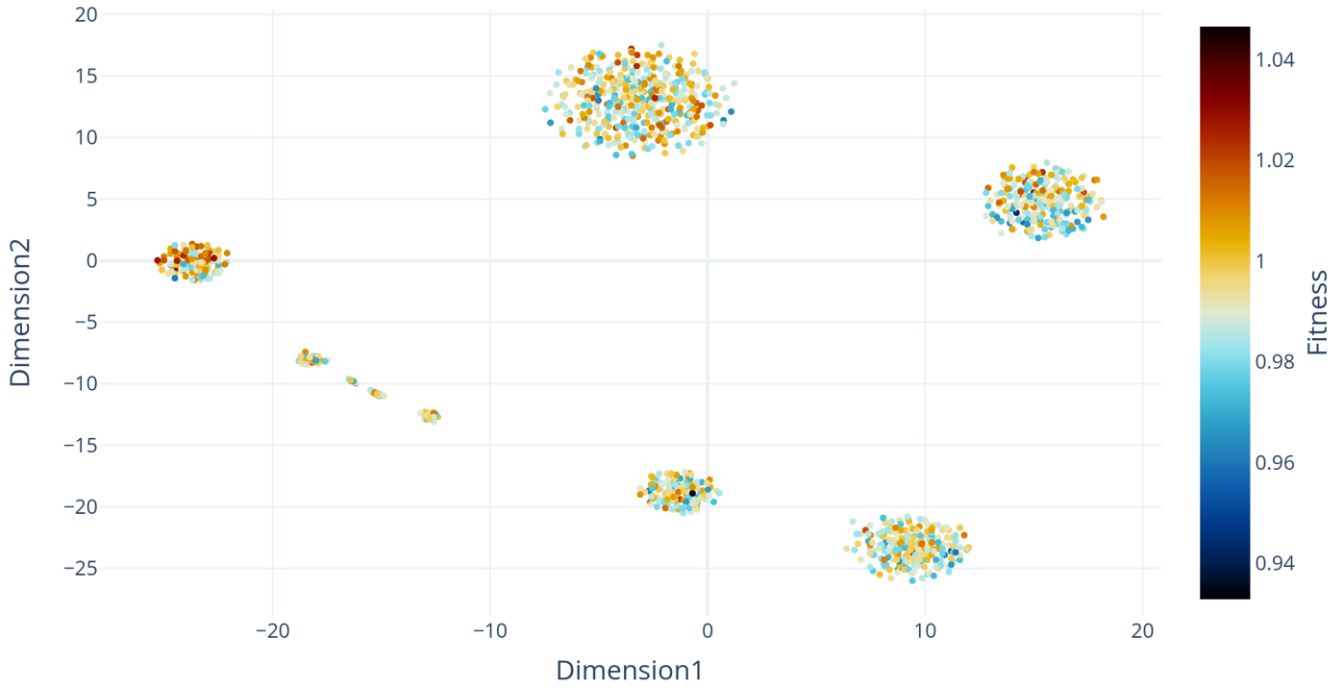


Fig. 14. The fitness landscapes of directed evolution with and without CRISPRi. Same data were used as Fig. 12, but the genomes were colored based on fitness instead.

Collectively, these results indicate that CRISPRi has the potential to guide directed evolution in a specific direction, and also to improve the efficiency of evolution.

IV. Discussion

The aim of the Yeast2.0 project is to produce eukaryotic chromosomes which are easily manipulable, and which can produce millions of variants on the original, naturally evolved, genomes, which can then be searched for genomes that are viable, relatively easily cultivated, and have biological characteristics which are desirable for use in application areas such as the production of drug precursors, biofuels, and industrial enzymes. The choice of *S. cerevisiae* as the first target was due to its known safety, easy culture conditions, and well-studied biology, but this approach could be applied, in principle, to any other eukaryote.

The identification of valuable variants, in the original project, was reliant upon evolution *in vitro*. This approach, while demonstrably valuable, has several drawbacks. The most obvious issue is that evolution *in vitro* requires time, expertise, and technology, and is costly. More fundamentally, however, this use of evolution takes no account of the fitness landscape of the system; it essentially considers each variant as an individual entity, without considering the relationships between variants, their mutations, and their places in the fitness landscape.

An understanding of the characteristics of the fitness landscape generated by the SCRaMbLE system will be of interest from a purely theoretical perspective because, in conjunction with the data produced in the biology laboratories of the consortium members, it provides an unparalleled opportunity to explore a real-world, extensive fitness landscape, and assess our understanding of this process by developing and evaluating simulation approaches. This project also has more directly practical applications. The ability to simulate the fitness landscape generated by the SCRaMbLE system may allow us to investigate the parameters of the system, and identify combinations of parameters which could lead to the generation of smooth fitness landscapes *in vivo*, thereby facilitating the process of artificial evolution *in vitro*, and saving time and money when identifying valuable variants. In the future, it may even be possible to develop genetic circuits to modulate the *in vivo* SCRaMbLE system to bias the evolutionary landscape in an optimal direction through the repression or enhancement of the recombination of particular loxP segments.

In this study, we developed a system for the simulation of SCRaMbLE *in silico*, including metrics for the distance between chromosomes, and for the fitness of the variants. These two metrics allow us to generate a fitness landscape for a SCRaMbLE run with a specific set of parameters. We applied our simulator to a single landscape, identified clear clusters of variants, and evaluated the ruggedness of the landscape of the chromosome we used.

We found that the simulation results of the deletion patterns of synIXR we obtained were consistent with real-world data, a finding which confirms that the SCRaMbLE

process tends to be random. By testing various values of the breakpoint probability of the simulator, we inferred that the real-world probability of a loxPsym site being involved in SCRaMbLE is around 0.3. This value could be narrowed down to a more precise number by running simple wet-lab experiments. We also found that the fitness landscape tends to be rugged, a finding which indicates that we may be able to improve the efficiency of the artificial evolution process by identifying changes that can be made to the system. With the success of modeling the SCRaMbLE of a chromosome and creating a resulting fitness landscape, the simulator was used to model directed evolution by running multiple rounds of SCRaMbLE on the chromosome synII. Patterns of ORFs related to high or low fitness scores were found by comparing heatmaps. The same simulations were executed again with the activation of CRISPRi. By protecting some essential ORFs from SCRaMbLE, CRISPRi sped up the process of finding genomes with high fitness. The results indicate that CRISPRi has the potential to redirect the path of evolution.

The fitness function could be further improved by integrating experimental multiple-gene mutation data, epistasis data, and copy number limitations. The simulation did not set an upper boundary for the copy number of an ORF, because no evidence for such a limitation was found in the literature. However, it is very likely that there would be a limitation on the copy number of an ORF in reality, due to homologous recombination. Thus, genomes with an ORF with more than 10 copies were filtered out in the computational model. With more experimental findings related to the limitation of copy number, the arbitrary limitation could be changed to a more accurate parameter.

The current version of the simulator could be improved in several ways. First, cis-regulatory elements like enhancers can be taken into account. Inverted enhancers would alter the expression of surrounding genes, although they are unlikely to have an impact if the ORF is also inverted. Identifying enhancers in the genome would be difficult, while quantifying their impact on fitness would be even harder. The interaction between replication and transcription could be taken into account in the model. On the one hand, transcription speed could be altered by inversion. Transcription could occur simultaneously with replication. However, the leading strand could be transcribed faster due to the fact that Okazaki fragments on lagging strands need to be connected by DNA polymerase and ligase. Thus, for those transcription units inverted from leading strands to lagging strands during replication, the transcription speed would be lower, and vice versa. The alteration of transcription speed would probably affect the phenotype. On the other hand, inversion might increase the probability of clashes between transcription and the replication fork. For eukaryotes, the speed of replication fork, 18–100 bp/s, is roughly the same as transcription, 2.3 kb/min, which reduces the chance of clashes (Pérez-Ortín, Alepuz and Moreno, 2007; Rogers, 2016; Gispan, Carmi and Barkai, 2017). In addition, most transcription moves in the same direction as the replication fork. Hence, for those transcription units, clashes with the replication fork are likely to occur if they are inverted due to the slower speed and inverted direction of transcription. To model these effects caused by inversion, we will need to locate the origin of replication, and then find parameters to evaluate the above effects.

For larger chromosomes, or high dimensional genomes, due to the number of pairwise similarities needed, the calculation increases exponentially, computation is expensive, and t-SNE is inefficient. A linear dimension reduction algorithm such as Principal Components Analysis could be applied to reduce the dimensionality to under 50, followed by the use of t-SNE (Van Der Maaten and Hinton, 2008). A potential approach would be to use LargeVis (Tang *et al.*, 2016) instead of t-SNE for dimension reduction. LargeVis constructs K-nearest neighbor graphs more efficiently, and uses a principled probabilistic model for graph visualization. Another dimension reduction solution is to use UMAP, which is believed to preserve a better global structure than t-SNE (Becht *et al.*, 2019).

This work will form the basis for an extended simulation study of the SCRaMbLE system. In future work, we will apply the system both to other chromosomes besides synIXR and synII and to individual chromosomes repeatedly, to investigate whether these preliminary results apply to other chromosomes, and to evaluate the extent of variability between the landscapes that can be generated from a single chromosome. As discussed above, one of the most important aspects of the research will be the evaluation of the effects of modification of the parameters on the ruggedness of the landscape. It is highly likely that these parameters do not interact linearly, making it unlikely that optimal parameter settings can be achieved by chance in the laboratory. We also envisage improving the parameterization for our model as new data emerges from future wet-lab SCRaMbLE studies.

In this work, we developed a simulator for the SCRaMbLE system, which has the potential to provide both theoretical and practical insights into this exciting new approach to bioengineering. We also simulated directed evolution based on this simulator with and without rational manipulation of SCRaMbLE, which indicated that the efficiency of directed evolution can be improved with genetic tools.

ACKNOWLEDGMENTS

B.Y. acknowledges funding from the China Scholarship Council (CSC).

REFERENCES

Barabasi, A.-L. and Oltvai, Z. N. (2004) 'Network biology: understanding the cell's functional organization', *Nature reviews genetics*. Nature Publishing Group, 5(2), pp. 101–113.

Becht, E. *et al.* (2019) 'Dimensionality reduction for visualizing single-cell data using UMAP', *Nature biotechnology*. Nature Publishing Group, 37(1), p. 38.

Cherry, J. M. *et al.* (2012) 'Saccharomyces Genome Database: the genomics resource of budding yeast', *Nucleic acids research*. Oxford University Press, 40(D1), pp. D700--D705.

Chervitz, S. A. *et al.* (1999) 'Using the Saccharomyces Genome Database (SGD) for analysis of protein similarities and structure', *Nucleic acids research*. Oxford University Press, 27(1), pp. 74–78.

Deutschbauer, A. M. *et al.* (2005) ‘Mechanisms of haploinsufficiency revealed by genome-wide profiling in yeast’, *Genetics*. Oxford University Press, 169(4), pp. 1915–1925.

Dymond, J. and Boeke, J. (2012) ‘The saccharomyces cerevisiae SCRaMbLE system and genome minimization’, *Bioengineered Bugs*. doi: 10.4161/bbug.19543.

Earl, D. J. and Deem, M. W. (2004) ‘Evolvability is a selectable trait’, *Proceedings of the National Academy of sciences*. National Acad Sciences, 101(32), pp. 11531–11536.

Erd, P. (1959) ‘os and A. R enyi, On random graphs I’, *Publ. Math. Debrecen*, 6, pp. 290–297.

Farkas, I. J. *et al.* (2011) ‘Spectra of" real-world" graphs: Beyond the semicircle law’, in *The Structure and Dynamics of Networks*. Princeton University Press, pp. 372–383.

Giaever, G. *et al.* (2002) ‘Functional profiling of the Saccharomyces cerevisiae genome’, *nature*. Nature Publishing Group, 418(6896), pp. 387–391.

Gispan, A., Carmi, M. and Barkai, N. (2017) ‘Model-based analysis of DNA replication profiles: predicting replication fork velocity and initiation rate by profiling free-cycling cells’, *Genome research*. Cold Spring Harbor Lab, 27(2), pp. 310–319.

Hallinan, J. S., James, K. and Wipat, A. (2011) ‘Network approaches to the functional analysis of microbial proteins’, *Advances in microbial physiology*. Elsevier, 59, pp. 101–133.

Hallinan, J. S., Misirli, G. and Wipat, A. (2010) ‘Evolutionary computation for the design of a stochastic switch for synthetic genetic circuits’, in *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*, pp. 768–774.

Hari, A. and Lobo, D. (2020) ‘Fluxer: a web application to compute, analyze and visualize genome-scale metabolic flux networks’, *Nucleic Acids Research*. Oxford University Press, 48(W1), pp. W427--W435.

Henikoff, S. and Henikoff, J. G. (1992) ‘Amino acid substitution matrices from protein blocks’, *Proceedings of the National Academy of Sciences*. National Acad Sciences, 89(22), pp. 10915–10919.

Hucka, M. and others (2003) ‘The systems biology markup language ({SBML}): a medium for representation and exchange of biochemical network models.’, *Bioinformatics*, 19(4), pp. 524–531.

James, K. *et al.* (2014) ‘Integration of gene expression data with interaction and annotation data reveals patterns of connection between primary Sjogren’s syndrome associated genes and immune processes’, *Rheumatology*. Oxford University Press, 53, pp. i136--i136.

Kauffman, S. and Levin, S. (1987) ‘Towards a general theory of adaptive walks on rugged landscapes’, *Journal of theoretical Biology*. Elsevier, 128(1), pp. 11–45.

Kvitek, D. J. and Sherlock, G. (2011) ‘Reciprocal sign epistasis between frequently experimentally evolved adaptive mutations causes a rugged fitness landscape’, *PLoS Genet*. Public Library of Science, 7(4), p. e1002056.

Levenshtein, V. I. (1966) ‘Binary codes capable of correcting deletions, insertions, and reversals’, in *Soviet physics doklady*, pp. 707–710.

Lu, H. *et al.* (2019) ‘A consensus S. cerevisiae metabolic model Yeast8 and its ecosystem for comprehensively probing cellular metabolism’, *Nature Communications*. doi: 10.1038/s41467-019-11581-3.

Van Der Maaten, L. and Hinton, G. (2008) ‘Visualizing data using t-SNE’, *Journal of Machine Learning Research*.

Marmiesse, L., Peyraud, R. and Cottret, L. (2015) ‘FlexFlux: Combining metabolic flux and regulatory network analyses’, *BMC Systems Biology*. doi: 10.1186/s12918-015-0238-z.

Ooi, S. L. *et al.* (2006) ‘Global synthetic-lethality analysis and yeast functional profiling’, *TRENDS in Genetics*. Elsevier, 22(1), pp. 56–63.

Orth, J. D., Thiele, I. and Palsson, B. Ø. (2010) ‘What is flux balance analysis?’, *Nature biotechnology*. Nature Publishing Group, 28(3), pp. 245–248.

Pérez-Ortín, J. E., Alepuz, P. M. and Moreno, J. (2007) ‘Genomics and gene transcription kinetics in yeast’, *TRENDS in Genetics*. Elsevier, 23(5), pp. 250–257.

Pitzer, E. and Affenzeller, M. (2012) ‘A comprehensive survey on fitness landscape analysis’, *Recent advances in intelligent engineering systems*. Springer, pp. 161–191.

Rogers, S. O. (2016) *Integrated molecular evolution*. Crc Press.

Schwartz, R. M. and Dayhoff, M. O. (1978) ‘Detection of distant relationships based on point mutation data’, *Evolution of protein molecules (eds. H. Matsubara and T. Yamanaka)*, pp. 1–16.

Shen, Y. *et al.* (2016) ‘SCRaMbLE generates designed combinatorial stochastic diversity in synthetic chromosomes’, *Genome Research*. doi: 10.1101/gr.193433.115.

Strathern, J. N., Jones, E. W. and Broach, J. R. (1982) *Molecular biology of the yeast Saccharomyces*. Cold Spring Harbor Laboratory.

Strogatz, S. H. (2001) ‘Exploring complex networks’, *nature*. Nature Publishing Group, 410(6825), pp. 268–276.

Tang, J. *et al.* (2016) ‘Visualizing large-scale and high-dimensional data’, in *Proceedings of the 25th international conference on world wide web*, pp. 287–297.

Ueda, M., Takeuchi, N. and Kaneko, K. (2017) ‘Stronger

selection can slow down evolution driven by recombination on a smooth fitness landscape', *PloS one*. Public Library of Science San Francisco, CA USA, 12(8), p. e0183120.

Ulitsky, I. and Shamir, R. (2007) 'Pathway redundancy and protein essentiality revealed in the Saccharomyces cerevisiae interaction networks', *Molecular systems biology*. John Wiley \& Sons, Ltd Chichester, UK, 3(1), p. 104.

De Visser, J. A. G. and Krug, J. (2014) 'Empirical fitness landscapes and the predictability of evolution', *Nature Reviews Genetics*. Nature Publishing Group, 15(7), pp. 480–490.

Weile, J. *et al.* (2012) 'Bayesian integration of networks without gold standards', *Bioinformatics*. Oxford University Press, 28(11), pp. 1495–1500.

Yoshikawa, K. *et al.* (2011) 'Comprehensive phenotypic analysis of single-gene deletion and overexpression strains of Saccharomyces cerevisiae', *Yeast*. doi: 10.1002/yea.1843.

Conflict of Interest

The authors declare no conflicts of interest.