Experimental assay of a fitness landscape on a macroevolutionary scale

Victoria O. Pokusaeva[1+], Dinara R. Usmanova[2+], Ekaterina V. Putintseva[3,4+], Lorena Espinar[3,4], Karen S. Sarkisyan[3,4], Alexander S. Mishin[5], Natalya S. Bogatyreva[3,4], Dmitry N. Ivankov[3,4], Guillaume J. Filion[3,4], Lucas B. Carey[4*], Fyodor A. Kondrashov[1,3,4,6*]

[1]*Institute of Science and Technology Austria, Am Campus 1, 3400 Klosterneuburg, Austria.*
[2]*Department of Systems Biology, Columbia University, New York, 10032, NY, USA.*
[3]*Bioinformatics and Genomics Programme, Centre for Genomic Regulation (CRG), 88 Dr. Aiguader, 08003 Barcelona, Spain.*
[4]*Universitat Pompeu Fabra (UPF), 08003 Barcelona, Spain.*
[5]*Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry, Miklukho-Maklaya 16/10, 117997 Moscow, Russia.*
[6]*ICREA, 23 Pg. LluísCompanys, 08010 Barcelona, Spain.*
[+]*Equal contribution*
[*]*Authors for correspondence*

Characterizing the fitness landscape, a representation of fitness for a large set of genotypes, is key to understanding how genetic information is interpreted to create functional organisms. Here we determined the evolutionarily-relevant segment of the fitness landscape of His3, a gene coding for an enzyme in the histidine synthesis pathway, focusing on combinations of amino acid states found at orthologous sites of extant species. The His3 fitness landscape is dominated by synergistic epistasis. For 80% of substitutions, the effect on fitness depended on the genetic background. Furthermore, at 67% of sites, substitutions are under sign epistasis, having both strongly positive and negative effect on different genetic background. 46% of sites were under reciprocal sign epistasis. Sign epistasis affected few genotypes but involved interaction of multiple sites which makes multidimensionality of the His3 fitness landscape essential, with the number of dimensions of the landscape approximately one quarter of the sequence length.

Predicting function and fitness of organisms from their genotypes is the ultimate goal of many fields in biology, from medical genetics to systems biology to the study of evolution[1-5]. Among the conceptual frameworks for understanding the genotype to phenotype connection is the fitness landscape, which assigns a fitness (phenotype) to every possible genotype (sequence) of a gene or genome under consideration[4,6]. The recognition of the importance of the fitness landscape stimulated the development of a variety of theoretical approaches to its description, including its general shape and epistatic interactions between alleles, a key property which determines the complexity of the fitness landscape[7-10]. Before the advent of next-generation sequencing, experimental assays of the fitness landscape were few and could not address the issue at the sequence level (e.g. [11]). Recently, large-scale experimental assays described the shape of the fitness landscape a few mutations away from a local fitness peak[12-21]. Also, some assays involving a smaller number of genotypes considered combinations of mutations with established functional[22-27] or evolutionary[28-33] significance.

Empirical evidence of the nature of large-scale fitness landscapes mostly comes from the study of genotypes incorporating random mutations[13-21] (also see [4] for review), a majority of which were deleterious[14,18-21,34]. Thus, our present knowledge of fitness landscapes is primarily driven by the study of deleterious mutations and their interactions, although local adaptive trajectories have also been considered[2,4,27,35-37]. Deleterious mutations were found to engage in synergistic epistasis, whereby the joint effect of multiple mutations was stronger than the sum of their individual effects[11,14,15,17-21,27]. Furthermore, sign epistasis was mostly rare[5,17-21,27,38], although some of these conclusions differ from study to study (i.e. see [4,38]).

Unfortunately, there are fundamental limitations to assaying the fitness landscape on a large or macroevolutionary level by random mutation libraries. The number of genotypes underlying the fitness landscape is the combinatorial set of all amino acids across the length of the protein[4,6]. For example, for the 220 amino acid protein coded by His3 gene in *Saccharomyces cerevisiae*, the fitness landscape is a 220 dimensional genotype space with 20^{220} different possible sequences. Such immense spaces are intractable computationally to say nothing of an experimental assay. Fortunately, it may not be necessary to survey all genotypes to study the evolutionary-relevant section of the fitness landscape. Because a vast majority of mutations in protein sequence are deleterious[34], a randomly sampled sequence is non-functional, especially a protein sequence[39,40], although this limitation may be less severe for DNA[41].

Here we propose an evolutionary approach for assaying fitness landscapes on a macroevolutionary scale in a high-throughput manner avoiding random sampling of mostly non-functional sequences. We reasoned that the functionally and evolutionarily relevant section of the fitness landscape can be represented by the combination of extant amino acid states, those found in extant species. This approach, applied previously on a limited scale[28-33] mitigates the problem of chartering a prohibitively large fitness landscape while highlighting the relationships between evolutionarily-relevant genotypes **(Figure 1a)**. Crucially, substitutions that have been fixed in evolution are fundamentally different from random mutations, the former representing the driving force of molecular evolution being either neutral or beneficial in at least some genetic contexts while the latter, in contrast, are mostly deleterious and are primarily relevant on a microevolutionary scale. Therefore, current empirical data do not shed much light on the impact of interactions between substitutions that fixed in the course of evolution by natural selection. Combinations of extant amino acid states allow one to assay a much wider functionally relevant area of the sequence space than approaches based on random mutagenesis of a single sequence (**Figure 1b,c**).

Estimating fitness of evolutionary-relevant genotypes

We studied His3, a gene coding for imidazoleglycerol-phosphate dehydratase (IGPD, His3p), an enzyme essential for histidine synthesis. In a multiple alignment of His3 orthologues from 21 yeast species we identified 686 extant amino acid states (**Supplementary Data File 1**), which were evenly distributed across the His3p structure (**Figure 1d**). These 686 substitutions, which occurred over the course of ~400 million years of yeast evolution[42] (**Figure 1b,c**), correspond to ~10^{83} sequences, even a tiny fraction of which would be too many to survey. Thus, we sectioned His3 into 12 independent segments such that the full combinatorial set of substitutions that have occurred in His3 during yeast evolution comprised 10,000-100,000 genotypes per segment (see **Methods** and **Extended Data Figure 1a**). The 12 segments were of similar length, constrained by the molecular methods employed for library construction (see **Methods**), and covering a diverse range of secondary structure and functional elements. For each of the 12 segments of His3 we performed an independent experiment surveying its fitness landscape. For each segment we used degenerate oligonucleotides to construct genotypes consisting of combinations of amino acids present in extant His3 sequences, and determined the fitness conferred by these genotypes by expressing them in a Δ*his3* strain of *S. cerevisiae* and measuring the rate of growth **(Extended Data Figure 1b)**. This way we assayed the fitness landscape of the genotype space that was traversed over the course of the last ~400 million years of evolution[42].

Across 12 experiments, we measured fitness for a total of 4,170,503 genotypes (956,648 unique amino acid sequences) with high accuracy. Of these, 452,709 consist solely of combinations of extant amino acid states from His3 orthologues, while the remaining genotypes incorporate other amino acid substitutions (**Supplementary Table 1** and **Methods**). For each segment we measured fitness for 60% - 99.8% of all possible genotypes from the combinatorial

103 set of selected extant amino acid states found in 21 yeast species and a smaller fraction of
104 combinations found across all domains of life (**Supplementary Table 1**), characterizing the
105 evolutionary relevant fitness landscape (**Figure 1b**). For example, for segment 3 consisting of 17
106 amino acid sites, for 11 sites more than one extant amino acid state was selected from the yeast
107 species (L145=2, L147=2, Q148=3, K151=2, V152=2, D154=3, L164=3, E165=4, A168=2,
108 E169=4, A170=4), with the full yeast combinatorial set consisting of 2*2*3*2*2*3*3*4*2*4*4 =
109 55,296 genotypes out of which we determined the fitness for 48,198, or 87% of the possible yeast
110 extant states combinations in our library.
111
112 A substantial proportion of combinations of extant amino acid states led to genotypes with
113 low fitness (**Figure 2, Figure 3a,b, Extended Data Figure 2**). This observation could be
114 explained by i) some extant amino acids having a universally deleterious effect, ii) some amino
115 acid states exerting a negative effect on fitness because of intergenic interactions with other *S.*
116 *cerevisiae* genes, or iii) by epistatic interactions between the extant amino acid states within His3[45].
117 We exclude the possibility that some extant amino acid states had a universally deleterious effect
118 because no extant amino acid states were present only in unfit genetic backgrounds (**Figure 3c**).
119 We exclude the possibility that some extant amino acid states disrupt intergenic interactions
120 because the complete His3 coding sequences from extant species fully complemented a His3
121 deletion in *S. cerevisiae* (**Extended Data Figure 3**). Thus, the observed genotypes with low fitness
122 can only be explained by epistatic interactions among extant amino acid states in the same or
123 different segment of the His3 gene. Remarkably, more than 80% of substitutions between extant
124 amino acid states had substantially different effects on fitness in different backgrounds (**Figure**
125 **3d**). In contrast, only 20% of amino acid substitutions that occurred in His3 evolution are truly
126 neutral, in the sense that do not exert strong influence on fitness in any genetic background. More
127 than half of the universally neutral substitutions were observed in the disordered region of the
128 protein (48/78). Thus, the His3 fitness landscape across the 12 segments was strongly influenced
129 by epistasis on a macroevolutionary scale, whereby the impact of extant amino acid state on fitness
130 being dependent on the background in which it occurs[43-45]. An epistatic fitness landscape is rugged
131 in the sense that evolving genotypes must avoid fitness valleys that emerge through deleterious
132 combinations of amino acid states that may also be found in fit genotypes[29,30,46-48]. Characterizing
133 the ruggedness and the mechanisms that determine the underlying epistasis becomes the primary
134 challenge in understanding the fitness landscape of His3.
135
136 **Unidimensional epistasis of the His3 fitness landscape**
137
138 The ruggedness of the fitness landscape can be characterized by the complexity of
139 underlying epistatic interactions. If the epistatic interactions are high-order[49-51] then the fitness
140 landscape can only be projected in a complex multidimensional space, i.e. fitness can only be
141 described by a multidimensional function of the genotype[52]. Alternatively, if epistasis is
142 unidimensional the fitness landscape can be described as a function of an intermediate variable,
143 called the fitness potential[53], determined by additive contribution of amino acid states.
144
145 To quantitatively determine how well fitness differences between genotypes can be
146 explained by unidimensional epistasis we used a deep learning approach. We used a sequential
147 neural network architecture composed of three layers. Each neuron in the architecture performs a
148 linear transformation of its input and then applies a nonlinear (sigmoid) function. Hence, by
149 utilizing one neuron in the first layer, we obtain a linear combination of the contributions of each
150 amino acid state to fitness potential, which is then non-linearly mapped to fitness by the three
151 layers of the neural network architecture. (see **Methods**; **Extended Data Figure 4**). Ten segments
152 were described by a sigmoid function in which organismal fitness remains constant with
153 decreasing fitness potential and then is rapidly reduced to lethal after a certain threshold (**Figure**
154 **4a**). The ability of the cliff-like threshold fitness function[54] to predict fitness from genotype varied

between the His3 segments from near perfect (R2=0.99) in segment 7, to relatively poor (R2=0.66) in segment 5 (**Extended Data Figure 5**).

Multidimensional epistasis of the His3 fitness landscape

Many genotypes were predicted poorly by any function of a unidimensional fitness potential (**Extended Data Figure 5a**), and many substitutions can both increase or decrease fitness depending on the background **(Figure 3d)**. Theoretically, the same deep learning approach could be used to detect non-additive multidimensional interactions by including pairwise (n-wise) terms as features of the samples in the binary feature matrix (see **Methods** and **Extended Data Figure 4**). However, the number of pairwise features is the square of the number of individual features and even ~100,000 unique genotype fitness measurements were too few to provide a reasonable sampling of pairwise interactions. Thus, the neural network could not be used to estimate the prevalence of multidimensional epistasis with the available number of data points. To determine the contribution of multidimensional forms of epistasis to the ruggedness of the His3 fitness landscape we identified instances of sign epistasis: substitutions between extant amino acid states that were strongly beneficial or strongly deleterious (change in fitness of > 0.4 in absolute value) depending on the background in which they occurred[55]. Some of these instances may be due to miscalled fitness of few genotypes, therefore, we confirmed pairs of extant amino acid states to be under sign epistasis only when sign epistasis was a significantly predictable function of the genetic background (see **Methods**).

An example of sign epistasis is the substitution C141S in the second segment that had an opposite effect on fitness depending on amino acid at site 143 (I, V or T). The substitution I143T in turn exhibits sign epistasis depending on the amino acid at site 163 (F, I, V or L) **(Figure 4b,d)**. These epistatic interactions can be represented by a graph in which nodes represent a pair of extant amino acid states at a specific site and nodes are connected by edges if strong sign epistasis has been detected between them (C141S - I143T - I163F) **(Figure 4d)**. We found that 86 out of 128 (67%) sites in our library exhibit sign epistasis and 46% (59/128) exhibit reciprocal sign epistasis. Most sites interacted by sign epistasis with multiple other sites (**Figure 4e,f**) demonstrating that, although sign epistasis effects few genotypes, it leads to a fitness landscape that requires multiple genotype dimensions for a proper characterization.

The dimensionality of the fitness landscape can then be determined by the dimensionality of the graph of all pairs of sites under sign epistasis. The Euclidean dimensionality of the fitness landscape was calculated as the fractal dimensionality of the graph of interacting pairs[56], which ranged from two to seven (**Extended Data Figure 6**). Mathematically, the dimensionality of the fitness landscape corresponds to the minimal number of dimensions in which the epistatic interactions shown in (**Extended Data Figure 6**) can be drawn with the lines having equal lengths (representing a Hamming distance of 1) and without intersections. Thus, strong sign epistasis at affected sites leads to a high dimensionality of the landscape, at least seven in segment 5, i.e. it is necessary to consider the simultaneous interaction of seven sites to accurately predict the fitness of genotypes consisting of extant amino acid states at these sites[50]. Consequently, high dimensionality decreases the ability of an additive unidimensional model to predict fitness of some genotypes from sequence **(Figure 4f, inset)**.

Inference of inter-segmental epistatic interactions in the His3 gene sequence

Epistasis may be caused by interaction among positions within a segment (intra-segmental epistasis) or by interaction of the segment with the rest of the *S. cerevisiae* His3 sequence (inter-segmental epistasis). The contribution of inter- versus intra-segmental interactions can be

decoupled. Given two fit genotypes (eg: ABC & abc in one His3 segment), any unfit intermediate states (eg: aBc in the same His3 segment) must be due to intra-segmental epistasis because the rest of the protein remains constant. For each segment we took as a measurement of intra-segmental epistasis all pairs of fit genotypes and calculated the proportion of unfit intermediate genotypes as a function of the Hamming distance between the two fit genotypes. We then compared this proportion with the total proportion of all unfit genotypes as a function of Hamming distance from *S. cerevisiae*, a measurement that includes both inter- and intra-segmental epistasis. We found three times more inter-segmental than intra-segmental epistasis **(Extended Data Figure 7)**, likely because a single segment provides a much smaller target space for interactions than the entire His3 protein. The proportion of sites under epistatic interactions increased exponentially with Hamming distance (**Extended Data Figure 7**), analogous to Orr's snowball, the accumulation of genetic incompatibilities in the course of speciation[43,57].

Evolutionary trajectories on the His3 fitness landscape

The presence of deleterious intermediate genotypes between highly fit ones leads to inaccessibility of some evolutionary trajectories between extant or ancestral sequences[6,7,31-33,58] **(Figure 5a)**. The sigmoid fitness function provides a mechanism for the restriction of available evolutionary trajectories in fitness space. Combinations of substitutions, all of which are neutral in some genetic backgrounds, can take a genotype beyond the fitness threshold through their additive effect on fitness potential, making some genotypes inaccessible for evolution (**Figure 5a**). Between any two fit genotypes, the fraction of intermediate genotypes that are unfit depends on the fitness potential of the two parental genotypes (**Figure 5b**). Evolution between two fit genotypes with high fitness potential can proceed unhindered because all intermediate genotypes also have high fitness potential and, consequently, high fitness. Conversely, when both fit genotypes are located close to the threshold, many of the intermediate genotypes between them have low fitness and many evolutionary paths between them are inaccessible (**Figure 5c**). Thus, the cliff-like threshold fitness function is the major determinant of the observation that not all paths between two fit genotypes are accessible to evolution (**Figure 5b**). We find that unfit intermediate genotypes are in genetic proximity with each other and are on a limited number of paths; the fraction of inaccessible paths is smaller than if the same number of unfit genotypes were distributed randomly in genotype space (**Figure 5d,e,f**).

The effect of synergistic epistasis dominates the His3 fitness landscape, affecting over 80% of amino acid substitutions from our library that occurred in His3 evolution. This synergistic epistasis may reflect the free energy of the protein[21,59,60], as evidenced by a correlation between the impact of substitutions on the free energy of His3p and fitness potential (**Extended Data Figure 8**). Similarly, instances of sign epistasis may also be explained by changes in protein stability; for example, in the 141T background C141S increased fitness and also had a positive effect on stability (**Figure 4c**). Consistent with protein stability contributing to the observed sign epistasis we find that sites that exhibited reciprocal sign epistasis are close together in the His3p structure **(Extended Data Figure 8)**. However, an additive contribution to free energy can lead only to a unidimensional fitness function[59], indicating that other non-additive mechanisms, such as catalytic activity or inter-subunit interactions, or a non-additive model of free energy, must be responsible for the multidimensionality of the His3 fitness landscape.

Conclusions

The concept of the fitness landscape introduced by Sewall Wright (Figures 1 and 2 in [6]) is an indispensable tool for understanding multiple biological phenomena[1-5]. Experimental high-throughput assays of random mutations have begun to unravel some local properties of fitness landscapes[4]. Here, we described a fitness landscape on a macroevolutionary scale by focusing on

amino acid states that have been put through the sieve of natural selection. We found that only 20% of substitutions that fixed in the evolution of His3 are universally neutral. For the remaining 80%, substitutions from His3 evolution had a profound influence on each other's effect on fitness, providing the first experimental confirmation that epistasis is a staple feature of molecular evolution[45]. Substitutions that occur in evolution have properties vastly different from those of random mutations, which are mostly deleterious[34]. Therefore, the way in which combinations of extant amino acid states affect fitness may also be different from that of combinations of random mutations. Unexpectedly, we found that the interaction of extant amino acid states was dominated by synergistic epistasis in a manner similar to that previously found for random mutations[11,14,15,17-21,27]. However, the accumulation of random mutations leads to low fitness much faster than the accumulation of extant amino acid states (compare Figure 2 from [27] and Figure 3b from [21] to **Figure 2a**).

A key difference of the fitness landscapes assayed by random mutations and through combinations of extant amino acid states is dimensionality. We found that up to 67% of sites with an extant amino acid state participate in multidimensional forms of epistatic interaction forming the basis of a multidimensional structure of the fitness landscape for most His3 segments. However, the presentation of the multidimensional sign epistasis was highly background-specific, affecting a relatively small number of genotypes. For example, the 5-dimensional fitness landscape of His3 segment 2 means that its genotype space cannot be described by fewer than five dimensions (such as the 5-dimensional sequences space in **Figure 1a**). Extrapolating from 20 amino acid long segments to a 220 amino acid long protein, the fitness landscape of the entire protein is expected to be a ~50 dimensional space. By contrast, large-scale fitness landscapes incorporating multiple mutations away from the wildtype sequence in a constant test environment have not displayed pervasive sign epistasis[19-21,27]. Multidimensional epistasis appears to be a more prevalent factor among substitutions that have been subject to positive selection[9,23-27,29-33,35-38], which may have been a factor in the fixation of some of the extant amino acid states in His3.

The evolutionary-relevant section of the His3 fitness landscape is best described as a fitness ridge, with the crest of the ridge defined by its high fitness potential. Evolution can proceed unhindered along the crest **(Figure 5c)**, however, pathway availability declines rapidly when evolution proceeds close to the edge of the fitness ridge. Occasionally, due to the influence of multidimensional sign epistasis a genotype will confer a fitness value not predicted by the fitness potential. The influence of sign epistasis on the His3 fitness ridge is that it is peppered with multidimensional holes and peaks along the landscape; however, their appearance on the landscape is, at present, unpredictable.

Acknowledgements

We thank Ben Lehner for feedback on the design of our study and Julia Domingo for technical assistance with preliminary analysis of our data, Jochen Hecht and the CRG Genomics Unit, and Svetlana Belogorodtdeva and Roman Belogorodtsev for technical assistance. The work was supported by HHMI International Early Career Scientist Program (55007424), the MINECO (BFU2012-31329 and BFU2015-68723-P), Spanish Ministry of Economy and Competitiveness Centro de ExcelenciaSevero Ochoa 2013-2017 grant (SEV-2012-0208), Secretaria d'Universitats i Recerca del Departament d'Economia i Coneixement de la Generalitat's AGAUR program (2014 SGR 0974), the CERCA Programme of the Generalitat de Catalunya, and the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013, ERC grant agreement 335980_EinME). L.B.C was supported by MINECO (BFU2015-68351-P) and AGAUR (2014 SGR 0974).

Methods

311 ***Data access***
312 The raw and processed data have been submitted to the NCBI Gene Expression Omnibus
313 (GEO;http://www.ncbi.nlm.nih.gov/geo/) under accession number GSE99990. A virtual machine
314 containing a running version of the data processing pipeline is available as a Docker image
315 https://hub.docker.com/r/gui11aume/epi/. The scripts to reproduce the figures are on Github at
316 https://github.com/Lcarey/HIS3InterspeciesEpistasis.
317
318 ***Study design***
319 The His3 gene was selected for three principle reasons, it is short, conditionally essential
320 and was not known to be involved in protein-protein interactions. Studying 20^{220} variants of His3
321 is impossible, thus, we have chosen an approach to survey the fitness landscape in a manner that
322 would elucidate the area most relevant to His3 evolution while managing the technical limitations
323 of our experimental design. We considered amino acid states found in extant species, focusing on
324 yeast species, which translated into a full combinatorial set of $\sim 10^{83}$ unique genotypes. Technically
325 it is feasible to measure fitness of up to 100,000 unique genotypes in a single growth experiment.
326 Therefore, we split the His3 gene into 12 independent segments such that the full combinatorial
327 set of extant amino acid states from 21 yeast species in each segment was 10,000 – 100,000
328 genotypes. We then considered the combinatorial library for each segment in an independent
329 growth experiment, which allowed us to study a tractable section of the sequence space while
330 considering trajectories across a vast part of the space connecting extant species (**Figure 1a**). We
331 constructed these combinations in 12 plasmid libraries and transformed them into a haploid His3
332 knockout *S. cerevisiae* strain. Growth rate (fitness) of yeast carrying different mutations in His3
333 was measured using serial batch culture in the absence of histidine.
334 We split the His3 gene sequence into segments in a manner agnostic to the structure of the
335 His3 protein (**Extended Data Figure 1a**). For technical reasons, a segment consisted of two
336 variable regions with a constant region between them (**Extended Data Figure 1b**). All growth
337 experiments were performed independently for each segment, with the exception of one
338 experiment on a limited group of genotypes from each segment which was done for the
339 normalization of fitness values across different segments (**Extended Data Figure 3**).
340 As a control, we measured the rate of growth of *S. cerevisiae* whose entire His3 gene
341 sequence came from another distant species. We found that the replacement of an entire gene
342 sequence of His3 leads to wild-type rates of growth of *S. cerevisiae* even when the His3 sequence
343 comes from very distant yeasts, as far as *S. pombe* (**Extended Data Figure 3**). Therefore, His3
344 appears to be an independent unit of the fitness landscape and is a good model for the study of
345 fitness landscapes of an isolated gene.
346
347 ***Measuring fitness***
348 *Plasmid construction*
349 The His3 open reading frame of *S. cerevisiae* was PCR amplified with its regulatory region
350 from 622 base pairs (bp) upstream of the open reading frame (ORF) to 237 bp downstream of the
351 ORF, using primers 126 and 127 (see **Supplementary Table 1**) from the wild-type prototroph
352 strain FY4. The PCR product was cloned into vector pRS416 using Gibson assembly (NEB,
353 E2611S). The His3 orthologues from other species were amplified from genomic DNA using
354 designed primers (**Supplementary Table 1**) and were cloned into the vector pRS416_his3,
355 replacing the ORF of *S. cerevisiae* by Gibson assembly (NEB, E2611S). Since the His3 orthologue
356 from *A. nidulans* contains an intron, the whole open reading frame was initially cloned into the
357 vector, and the intron was later removed by PCR-amplifying the whole plasmid without this
358 sequence, followed by recircularization.
359
360 *Genomic DNA extraction*
361 Genomic DNA from fungi (*Saccharomyces cerevisiae, Saccharomyces bayanus, Candida*
362 *glabrata, Saccharomyces castellii, Kluyveromyces lactis, Eremotheciumgossypii, Debaryomyces*

363 *hansenii, Lodderomycese longosporus, Aspergillus nidulans, Schizosaccharomyces pombe,*
364 *Candida guilliermondii, Saccharomyces kluyveri, Kluyveromyces waltii*) was extracted using
365 MasterPure™ Yeast DNA Purification Kit according to the manufacturer's instructions
366 (Epicentre, MPY80200).
367
368 *Mutant library construction*
369 Twelve independent mutant libraries, each for different regions of His3 **(Supplementary**
370 **Table 1)**, were generated based on the results of multiple alignment of His3 orthologues. The
371 alignment was built using the ClustalW alignment feature of the MEGA 6.0 software package[61]
372 and user-corrected.
373 Mutant libraries were constructed by fusion-PCR, leaving two variable regions separated
374 by a constant region. For each library, two contiguous fragments of His3 were amplified
375 independently, using 1 µg of *S. cerevisiae* (strain FY4) genomic DNA in separate Phusion
376 polymerase reaction mixes (Thermo Fisher Scientific, F530S) in GC buffer. For each PCR, one of
377 the primers was a degenerate oligonucleotide with a constant part at the 5' end required for the
378 fusion-PCR; the other primer was either 126 or 127. The cycling conditions for the PCR were
379 98°C for 30 s; 98°C for 20 s, 60 °C for 30s and 72°C for 1 min (25 cycles); and 72°C for 5 min.
380 The products were column-purified (QIAGEN, QIAquick PCR purification kit, 28104), eluted in
381 50 µl and mixed in equimolar proportion. The fusion-PCR was carried out by diluting 10 µl of the
382 mix to 25 µL of standard Phusion polymerase reaction mix in GC buffer. The cycling conditions
383 of the fusion-PCR were 98°C for 30 s; 98°C for 30 s, 60°C for 2 min and 72°C for 1 min (25
384 cycles); and 72°C for 5 min. The product of fusion was purified from agarose gel (Qiagen,
385 MinElute Gel Extraction Kit, 28604) and eluted in 10 µl of water. 10 µl of the product was used
386 as a template for additional 5 cycles of PCR reaction in Phusion polymerase reaction mix (Thermo
387 Fisher Scientific, F530S) in GC buffer, using primers 126 and 127. The cycling conditions were
388 as follows: 98°C for 30 s; 98°C for 20 s, 60°C for 30 s and 72°C for 1 min (5 cycles); and 72°C
389 for 5 min. The product was column-purified (QIAGEN, QIAquick PCR purification kit, 28104),
390 and used as an insert for Gibson assembly.
391 To create a library of His3 mutants, pRS416 plasmid was amplified using primers 128 and
392 129. The insert was cloned into the vector using Gibson assembly (NEB, E2611S). Ligated
393 products (200 - 300 ng/µL) were desalted by drop dialysis using 13 mm diameter, Type-VS
394 Millipore membrane (Merck Millipore, VSWP01300). 20 µL ElectroMAX DH10B competent
395 cells (Invitrogen, 18290015) were electroporated with 3 µL ligated products. 0.01% of the
396 electroporated bacteria were plated on ampicillin-containing medium in order to estimate the
397 complexity of the library; the remaining culture was grown overnight in 100 ml of liquid medium,
398 and the plasmid was extracted the next day. For each library, the maximum number of protein
399 sequences that can be generated was computed. Libraries were generated until to total complexity
400 reached at least 3 times this value.
401
402 *Yeast transformation and yeast library generation*
403 For each segment, yeast strain LBCY47 (*his3*:KanMX*leu2*Δ0 *met15*Δ0 *ura3*Δ0, derived
404 from BY4741) was transformed with 50 µg of pRS416_His3 mutant library using lithium acetate
405 transformation and plated onto glucose synthetic complete dropout plates lacking uracil. After 40
406 hours' growth at 30°C, approximately 0.5 million yeast colonies were scraped off the plates, mixed
407 together and washed 2 times with 100 ml of PBS.
408
409 *Bulk competition*
410 $4x10^9$cells were inoculated into 500 ml of glucose synthetic complete dropout medium
411 lacking uracil with 200 mg/L of G418, and grown at 30°C at 220 RPM for 6-8 h in order to
412 eliminate clones with low fitness irrespective of histidine biosynthesis. Cells were later pelleted
413 and washed with 50 ml of PBS. Approximately 10^{10} cells were inoculated into 1 L of synthetic
414 complete dropout medium lacking histidine, and grown at 30°C at 220 RPM for 168 h with 12 h

between bottlenecks: ~10^{10} cells were transferred into fresh medium ~10^8 cells from the culture were kept as sample for the given time point. Bulk competition for each library of mutants was done in two replicates to account for biological variability.

NGS library preparation

The relative abundance of yeast mutants was measured in 3 samples: 1) the initial population before selection was applied (t0), 2) the population after 12 h of growth in the selective medium (t1), and 3) the final population after 168 h of growth in the selective medium (t14). In order to extract plasmid DNA, $5x10^9$ cells from each sample were incubated in 300 µL of zymolyase buffer (1 M sorbitol, 0.1 M sodium acetate, 60mM EDTA (pH 7.0), 2 mg/ml zymolyase, 1% 2-Mercaptoethanol) at 37°C for 3 h. The plasmid DNA was purified from the obtained spheroplasts using QIAprep Spin Miniprep Kit (QIAGEN, 27104) according to the manufacturer's protocol. The obtained DNA was used as a template in a 25 µL of Q5 DNA polymerase reaction mix (NEB, M0491S), using staggered primers for demultiplexing in the following cycling conditions: 98°C for 30s; 98°C for 10s, 60°C for 30s and 72°C for 30s (18 cycles); and 72°C for 2 min. PCR products were purified using Agencourt AM Pure XP beads (Beckman Coulter, A63880), and eluted in 40 µL of TE buffer (pH 8.0). DNA extraction and PCR-amplification were repeated twice for every sample to account for the technical variability.
NGS libraries were prepared from 100 ng of the purified DNA amplicons using Ovation Rapid DR System (Nugen, 0319-32) according to manufacturer's instructions. Each library was visualized on a Bioanalyzer (Agilent Technologies) and quantified by qPCR with a Kapa Library Quantification Kit (Kapa Biosystems, KK4835). Twelve samples were pooled together (accounting for two biological replicates, two technical replicates and three time points) at the final concentration of 4 nM, and sequenced in the same lane. Samples were sequenced as 125-bp paired-end reads on a HiSeq2500 sequencer (Illumina) with v4 sequencing chemistry.

Yeast growth assay

Mutant strains were grown overnight in complete dropout medium lacking uracil. The cultures were diluted to 0.05 OD 600 nm, and grown for 5 h in the same medium. 2 µL of each culture were transferred into 96-well plates in 125 µL of complete dropout medium lacking histidine. Growth of the strains was monitored by measuring OD 600 nm every 60 min using Infinite M200 microplate reader (Tecan). Growth rate was defined as the slope of the logOD (time) line.

Growth rates of isolated strains

We isolated 197 strains from all segment libraries of extant amino acid combinations (9-26 strain per segment) and used Sanger sequencing to determine the sequence. For each strain we performed 6 repeats of growth assay and calculated the average growth rate. Fitness values from competition and growth rates are highly correlated (r=0.82 p=10^-48). Correlation was significant and greater than 0.6 for all segments except S9, where all selected genotypes appeared to be neutral (**Extended Data Figure 3**).

Initial data filtering

The individual sequences of the variants were recovered from pair-end reads with the following steps: The constant region between the two variable regions was identified by inexact matching allowing up to 20% errors using the Seeq library version 1.1.2 (https://github.com/ezorita/seeq). The reads are not oriented because the Illumina sequencing adapters were added by ligation, so the constant regions were searched on both reads. Forward and reverse reads were swapped when a match was found on the reverse read. This ensured that all of the sequences are in the same orientation. For multiplexing purposes, the sample identity was encoded in the left and right primers used to PCR-amplify the variants. To demultiplex the reads, we used inexact matching with the candidate primers, allowing up to 20% errors. This approach

was faster and less error-prone than using FLASH[62]. To merge the reads, the sequence of the reverse reads was reverse complemented and the constant region was searched by inexact matching allowing up to 20% errors. The position of the constant part in each read indicated how they must be stitched together. In the region of overlap, the consensus sequence was determined by picking the nucleotide with highest quality as indicated in the quality line of the fastq files. If 'N' persisted in the final sequence, the reads were discarded. The PCR primers were trimmed so that all the sequences of the same competition would start and end at the same location.

Reads that did not have the constant region, that could not be oriented or that could not be demultiplexed were discarded. The remaining errors in the reads were corrected by sequence clustering. We used starcode version 1.0 ref. [63] with default parameters and allowing up to two errors. The corrected reads were translated using the genetic code. Variants encoding the same proteins were not merged; they were kept separate for downstream analyses. A running Docker virtual machine with commented scripts to replay the whole the process is available for download at https://hub.docker.com/r/gui11aume/epi/.

DNA sequence variant frequency calculation and data filtering

The total number of reads for 12 segments, 3 time points and 4 replicas are shown in **Supplementary Table 1**. Genotypes frequencies are defined as the number of reads for a given genotype divided by the total number of reads in that replicate. Mean frequency was calculated over 4 replicas to be used in further analysis. However, to eliminate influence of outliers the median was taken instead of mean if absolute difference between mean and median was greater than the median value. Only genotypes present in both technical replicas of both biological replicas with at least ten reads (summed across all time points) in each of them were kept.

Noise estimation

The major factors causing noise in genotype frequency measurements are sampling errors, PCR amplification errors and genetic drift during the competition. For all of these factors, the amount of error depends on the genotype frequency. Therefore, we estimated measurement errors as the function of genotype frequency.

For a given segment, time point and a pair of biological or technical replicas for each genotype we calculated the mean frequency and the squared difference of frequencies from these two replicas. We sorted genotypes by mean frequency and grouped them such that each bin contains 5000 genotypes. We calculated the average frequency and the average squared difference in each bin. Additionally, squared error for frequency 0 was set equal to $\frac{1}{2}\cdot\left(\left(\frac{0.5}{N_i}\right)^2+\left(\frac{0.5}{N_j}\right)^2\right)$, where N_i and N_j are total read numbers in replicas i and j. Finally, by linear interpolation we obtained dependencies of squared differences as a function of frequency, $s_{ij}^2(f)$, where i and j are different replicas.

Using squared differences from pairwise comparison of replicas we can estimate variance of mean frequency over four replicas. Let numerate replicas 1, 2, 3, 4 where 1, 2 are technical replicas of the first biological repeat and 3, 4 are the technical replicas of the second biological repeat. Errors coming from the competition (e.g.: genetic drift) are shared for replicas 1, 2 and for replicas 3, 4. Let's call them Δf_{b_1} and Δf_{b_2} and their variances $\sigma_{b_1}^2$ and $\sigma_{b_2}^2$ respectively. Technical errors of sampling from the population and from PCR are unique for each replica. Let's call them $\Delta f_{t_i}, i = 1..4$ and their variances $\sigma_{t_i}^2, i = 1..4$ respectively. All variances are function of frequency and writing $\sigma_{X_i}^2$ we assume $\sigma_{X_i}^2(f)$.

In the introduced notations the mean frequency over 4 replicas is:

$$f = \frac{1}{4}\cdot(f_1+f_2+f_3+f_4) =$$

$$\frac{1}{4}\cdot\left(\left(f^*+\Delta f_{b_1}+\Delta f_{t_1}\right)+\left(f^*+\Delta f_{b_1}+\Delta f_{t_2}\right)+\left(f^*+\Delta f_{b_2}+\Delta f_{t_3}\right)+\left(f^*+\Delta f_{b_2}+\Delta f_{t_4}\right)\right) =$$

$$f^* + \frac{1}{2} \cdot \left(\Delta f_{b_1} + \Delta f_{b_2}\right) + \frac{1}{4} \cdot \left(\Delta f_{t_1} + \Delta f_{t_2} + \Delta f_{t_3} + \Delta f_{t_4}\right),$$

where f^* is the true frequency. Applying basic properties of variance the variance of mean frequency:

$$\sigma^2 = \frac{1}{4} \cdot \left(\sigma_{b_1}^2 + \sigma_{b_2}^2\right) + \frac{1}{16} \cdot \left(\sigma_{t_1}^2 + \sigma_{t_2}^2 + \sigma_{t_3}^2 + \sigma_{t_4}^2\right)$$

To estimate $\sigma_{b_1}^2, \sigma_{b_2}^2, \sigma_{t_1}^2, \sigma_{t_2}^2, \sigma_{t_3}^2, \sigma_{t_4}^2$ we used squared differences from pairwise comparison of replicas calculated above $s_{12}^2, s_{13}^2, s_{14}^2, s_{23}^2, s_{24}^2, s_{34}^2$:

$$\mathbb{E}[s_{12}^2] = \mathbb{E}[(\Delta f_{t_1} - \Delta f_{t_2})^2] = \sigma_{t_1}^2 + \sigma_{t_2}^2$$
$$\mathbb{E}[s_{13}^2] = \mathbb{E}[(\Delta f_{b_1} + \Delta f_{t_1} - \Delta f_{b_2} - \Delta f_{t_3})^2] = \sigma_{b_1}^2 + \sigma_{t_1}^2 + \sigma_{b_2}^2 + \sigma_{t_3}^2$$
$$\mathbb{E}[s_{14}^2] = \mathbb{E}[(\Delta f_{b_1} + \Delta f_{t_1} - \Delta f_{b_2} - \Delta f_{t_4})^2] = \sigma_{b_1}^2 + \sigma_{t_1}^2 + \sigma_{b_2}^2 + \sigma_{t_4}^2$$
$$\mathbb{E}[s_{23}^2] = \mathbb{E}[(\Delta f_{b_1} + \Delta f_{t_2} - \Delta f_{b_2} - \Delta f_{t_3})^2] = \sigma_{b_1}^2 + \sigma_{t_2}^2 + \sigma_{b_2}^2 + \sigma_{t_3}^2$$
$$\mathbb{E}[s_{24}^2] = \mathbb{E}[(\Delta f_{b_1} + \Delta f_{t_2} - \Delta f_{b_2} - \Delta f_{t_4})^2] = \sigma_{b_1}^2 + \sigma_{t_2}^2 + \sigma_{b_2}^2 + \sigma_{t_4}^2$$
$$\mathbb{E}[s_{34}^2] = \mathbb{E}[(\Delta f_{t_3} - \Delta f_{t_4})^2] = \sigma_{t_3}^2 + \sigma_{t_4}^2$$

Therefore, the variance of mean frequency f can be found as:

$$\sigma^2 = \frac{1}{16} \cdot \left((s_{13}^2 + s_{14}^2 + s_{23}^2 + s_{24}^2) - (s_{12}^2 + s_{34}^2)\right)$$

Recalling that variance and squared differences are a function of frequency:

$$\sigma^2(f) = \frac{1}{16} \cdot \left((s_{13}^2(f) + s_{14}^2(f) + s_{23}^2(f) + s_{24}^2(f)) - (s_{12}^2(f) + s_{34}^2(f))\right)$$

For each segment and time point we calculated the numerical function $\sigma^2(f)$. Then for each genotype having mean frequency f_x we estimated its variance as $\sigma^2(f_x)$

Merging amino acid genotypes

We merged nucleotide genotypes that corresponded to the same amino acid sequence and summed their frequencies and variances. We filtered out all genotypes *x* which had any of following patterns:

$f_x^{t_0} = 0, f_x^{t_1} = 0, f_x^{t_2} > 0$ or $f_x^{t_0} = 0, f_x^{t_1} > 0, f_x^{t_2} = 0$ or $f_x^{t_0} > 0, f_x^{t_1} = 0, f_x^{t_2} > 0$. Fraction of such genotypes were <0.5% for all segments except S9, for which it was 4.5%

For further analysis, this amino acid dataset was used except when specified.

Fitness estimation

Number of cells in a pool with particular genotype *x* after time interval *t* increases exponentially

$$n_x^t = n_x^0 \cdot Exp[s_x \cdot t],$$

where s_x is absolute fitness. Frequency of genotype *x* as well depends exponentially on absolute fitness with an additional multiplicative factor:

$$f_x^t = \frac{n_x^t}{N^t} = \frac{n_x^0 \cdot Exp[s_x \cdot t]}{N^t} = \frac{f_x^0 \cdot Exp[s_x \cdot t]}{N^t/N^0},$$

where N^t and N^0 are total cell numbers in a pool at time points 0 and *t*. Factor $\frac{1}{N^t/N^0}$ reflects the total growth of population, it changes with time but is the same for all genotypes. Therefore, we can rewrite genotype frequency at time *t* as:

$$f_x^t = f_x^0 \cdot Exp[(s_x - \overline{s^{0t}}) \cdot t],$$

where $\overline{s^{0t}} = \frac{1}{t} \cdot Log\left(\frac{N^t}{N^0}\right)$

In the measured dataset for each genotype *x* we have 3 measurements of frequency $f_x^{t_0}$, $f_x^{t_1}$, $f_x^{t_2}$ and their errors $\sigma^2(f_x^{t_0}), \sigma^2(f_x^{t_1}), \sigma^2(f_x^{t_2})$. To estimategenotype fitness we minimizedrelative squared errors of exponential fit as function of fitness s_x and initial frequency f_x^0:

$$(s_x, f_x^0) = \underset{s_x, f_x^0}{\operatorname{argmin}} \left(\frac{\left(f_x^{t_0} - f_x^0\right)^2}{\sigma^2\left(f_x^{t_0}\right)} + \frac{\left(f_x^{t_1} - f_x^0 \cdot Exp[(s_x - \overline{s^{01}}) \cdot t_1]\right)^2}{\sigma^2\left(f_x^1\right)} + \frac{\left(f_x^{t_2} - f_x^0 \cdot Exp[(s_x - \overline{s^{02}}) \cdot t_2]\right)^2}{\sigma^2\left(f_x^2\right)} \right) \quad (1)$$

This formula contains four parameters common for all genotypes from one segment: $\overline{s^{01}}, \overline{s^{02}}, t_1, t_2$. Further we will perform additional shifting and scaling of fitness values (see next section), therefore, without loss of generality we could set$\overline{s^{01}} = 0$ and $t_1 = 1$.Ideally,t_2/t_1 should equal 14, however, we noticed that this ratio does not hold for many segments and fitted $k = t_2/t_1$ from data instead of using value 14.

To find specific$\overline{s^{02}}$ and k for each segment we selected genotypes with high frequencies at t_0 ($t_0 > 25 \cdot 10^{-6}$) which corresponds to ~500-1000 reads per technical replicate. Each segment contains 10^3-10^4 genotypes that meet this criterion. We minimized eq. (1) for selected genotypes trying all possible combinations of $(\overline{s^{02}}, k)$ from a grid where $\overline{s^{02}} \epsilon [0,1.2]$ with step 0.01 and $k \epsilon [1,14]$ with step 0.1 and choose $(\overline{s^{02}}, k)$ which gives minimal (*).

Finally, given $(\overline{s^{02}}, k)$ for each segment we found s_x for each genotype. Errors for fitness values, s_x^{std}, were estimated as standard error of best-fit parameter.

For genotypes with frequencies pattern$f_x^{t_0} > 0, f_x^{t_1} = 0, f_x^{t_2} = 0$ fit of eq. (1) cannot be obtained. Therefore we defined upper boundary for their fitness value as $s_x^{boundary} = Log\left(\frac{1}{\max(N_1^{t_1}, N_2^{t_1}, N_3^{t_1}, N_4^{t_1})}\right)$, where $N_i^{t_1}, i = 1..4$ are total read numbers at time point t_1 in *i* replica.

Fitness rescaling

We scaled fitness such that lethal genotypes have fitness 0 and neutral genotypes have fitness 1. We assumed that genotypes with a stop codon or frame shift are lethal. Thus, for each segment we linearly rescaled the fitness distribution so that 95% of genotypes with nonsense mutations have a fitness of 0 and so that the local maximum of the fitness distribution of genotypes with extant amino acids is 1. The scaling around the local maximum led to the shift of fitness values of less than +/- 0.025 in each of the 12 segments compared to the measured wildtype strains and did not affect our results (for scale, we called a substitution non-neutral if its effect on fitness was > 0.4). All fitness values which became smaller than 0 were set to 0.

Control by comparison of synonymous sequences

We used nucleotide synonymous sequences as an internal control. The error rate for a measurement of fitness of an amino acid sequence depends on the number of synonymous sequences, *n*, that were used to estimate it. Therefore, we estimated the false discovery rates separately for categories with *n=1,..10* variants. For each amino acid genotype with more than *n* synonymous variants we merged random combination of *n* of its nucleotide genotypes and estimated fitness. We then calculated the difference between this fitness and the fitness of the corresponding amino acid sequence. We classified case as "false unfit" if difference was <-0.4 and as "false fit" if difference was >0.4. The fraction of such cases gives us false discovery rates for genotypes having *n* synonymous variants. To get total false discovery rates for each segment we averaged "false unfit" and "false fit" rates for different *n* with weights equal to the fractions of genotypes in amino acid dataset which have *n* synonymous variant (**Supplementary Table 1**).

Predicting fitness using deep learning

To predict the unidimensional fitness function based on additive contribution of extant amino acid states we used deep learning, a powerful machine learning technique, capable of constructing

virtually any function, even with a simple neural network architecture. To convert amino acid sequences into a binary feature matrix we used one-hot encoding in which each feature (column in the matrix) is the presence or absence of a particular amino acid state.

To optimize the accuracy/overfitting ratio, we tested over a hundred of different neural network architectures and parameters. As a starting point, selected several complicated architectures, which described our data but were prone to overfitting due to a large number of parameters. We then gradually reduced the number of layers and neurons to reduce the overfitting, while controlling for accuracy.

Our final architecture consists of three layers and four neurons in total (**Extended Data Figure 4**). Each of the neurons performs a linear transformation of the input and subsequently applies a non-linear activation function (a sigmoid) to the result. The output of the first layer, therefore, is a sigmoid of a linear transformation of the feature vector ($\mathbf{w}_1^T x$). The second layer decompresses the hidden nonlinear representation into 2 sigmoids, the combination of which is further linearly transformed with the only neuron of the third layer and wrapped into another sigmoid function:

$$F(x) = \sigma(w_{3,1}\sigma(w_{2,1}\sigma(\mathbf{w}_1^T x + b_1) + b_{2,1}) + w_{3,2}\sigma(w_{2,2}\sigma(\mathbf{w}_1^T x + b_1) + b_{2,2}) + b_3)$$

where $\sigma(t) = \frac{1}{(1+exp(-t))}$, $x -$ is the feature vector, $\boldsymbol{w}_1$ – the vector of weights, corresponding to the first layer, $w_{n,k}$- the weight corresponding to the n-th layer and the k-th neuron, b_n - bias corresponding to the n-th layer. This architecture is capable of fitting not only to sigmoid shapes, but other more complex functions.

In the first hidden layer, a single neuron was used to obtain the predicted fitness potential. Thus, each amino acid state was assigned a single unique weight. Multiplication of the feature and weight vectors gave the predicted fitness potential, which was transformed by a nonlinear phase shift function constructed by the four neurons of the neural network. The architecture simplicity avoids overfitting, which was further prevented by using 95:5 cross-validation. To ensure reproducibility, each of the models was constructed five independent times using random train-test splits.

Each model was trained for 500 epochs using mean squared error as the loss function. An unpublished adaptive learning rate method proposed by Geoff Hinton, RMSProp, was used as the optimizer. This algorithm is a version of a mini-batch stochastic gradient descent, utilizing the gradient magnitude of the recent gradients in order to normalize the current ones. All the weights were initialized using Xavier normal initializer[64].

Paths between pairs of fit genotypes

For analysis in Figure 5d, we first choose two fit "parental" genotypes, one randomly chosen genotype (eg: ABE) and the other parental genotype that is either *S. cerevisiae* wildtype genotype (inter-segmental) or another random fit genotype in the data (intra-segmental) (eg: abe). The two genotypes in this example are Hamming Distance 3 apart (HD=3). We next compute all (HD2-2) intermediate genotypes (eg: AbC, aBc, *et cetera*) and retain the subset that were experimentally measured. We represent the two parental genotypes and all measured intermediate genotypes as an undirected graph in which each genotype is a vertex. All genotypes one substitution apart are connected by an unweighted edge. The shortest possible path for a given pair of genotypes is of length HD. We find all shortest paths between the two parental genotypes using a breadth-first search. We next remove all vertices (genotypes) that are unfit, and recompute the number of shortest between the two parental genotypes. For example, in Fig5a, there are six paths of length three if you take into account all genotypes, but only three paths of length three if you take into account only fit genotypes.

Clustering of unfit genotypes in sequence space

For the analysis in Figure 5e,f, we first represent the two parental genotypes and all measured intermediate genotypes as an undirected graph in which each genotype is a vertex. All genotypes one substitution apart are connected by an unweighted edge. We can then compute the degree (# of genotypes of distance one) for each vertex (genotype). We do so using all measured genotypes or using only unfit genotypes or using the same number but randomly chosen genotypes. For the randomly chosen genotypes, the value is the average over 1000 runs.

Quantifying sign epistasis

For each substitution (eg: C -> S at position 141), we considered only those that exhibit a large fitness effect (abs. difference > 0.4) comprising a set of substitutions with large effects. For each substitution we divided the genetic backgrounds into two categories: those in which the substitution caused a > 0.4 increase in fitness, and those backgrounds in which the substitution caused > 0.4 decrease in fitness. A single substitution can cause a large increase in fitness in some backgrounds and a large decrease in others due to two possible reasons: sign epistasis or experimental error. To differentiate the two cases, we identified secondary substitutions that significantly alter the ratio of large increases to large decreases in fitness (Fisher's exact test, Bonferroni corrected p-value < 0.05). We only consider a site to be under sign epistasis if there is a second site that alters the frequency of sign epistasis in a statistically significant manner, i.e. more frequently than expected by chance alone.

Structural analysis

Structure prediction

An initial model was obtained with the I-TASSER server[60]. The list of top 10 PDB structural templates picked up by the I-TASSER included high-quality crystal structures of imidazoleglycerol-phosphate dehydratases from *Arabidopsis thaliana* and *Cryptococcus neoformans*. Coordinates of the top-scoring model (C-score=0.21, estimated TM-score = 0.74±0.11, estimated RMSD = 5.1±3.3Å) and the predicted normalized B-factor[66] were used for further analysis. The value of the model quality metric (TM-score >0.5) indicates a model of correct topology. The proteins structurally close to the final model (RMSD 0.6 - 1.7Å are PDB IDs 4MU0, 4GQU, 1RHY, 5DNL and 2AE8 from *Arabidopsis thaliana*, *Mycobacterium tuberculosis*, *Cryptococcus neoformans*, *Pyrococcusfuriosus* and *Staphylococcus aureus*.

We measured the distribution of distances (in angstroms) between pairs of residues that exhibit strong sign epistasis (**Supplementary Table 1**, ReallyPositivePair == TRUE), and compared it with the distribution of pairwise distances among residues for which we have sufficient data to be certain that a given pair does not exhibit sign epistasis (**Supplementary Table 1**, ReallyNegativePair == TRUE).

ΔΔG prediction

Cartesian_ddg application[67] from Rosetta version 2017.08.59291 was used for ΔΔG predictions. Top-scoring I-TASSER model was pre-minimized using the Relax[68] application in dual-space[69] with the flags: -relax:dualspace true; -ex1; -ex2; -use_input_sc; -flip_HNQ; -no_optH false; -relax:min_typelbfgs_armijo_nonmonotone; -nonideal. The best scoring model from 1000 structures was selected. The effect of up to 4 mutations (54,500 genotypes in total) was assessed in cartesian space with the Talaris_2014 score function, and the -fa_max_dis 9.0 flag. ΔΔG was estimated as a difference of mean score for 3 independent runs for every mutant and the wild-type score.

Figures

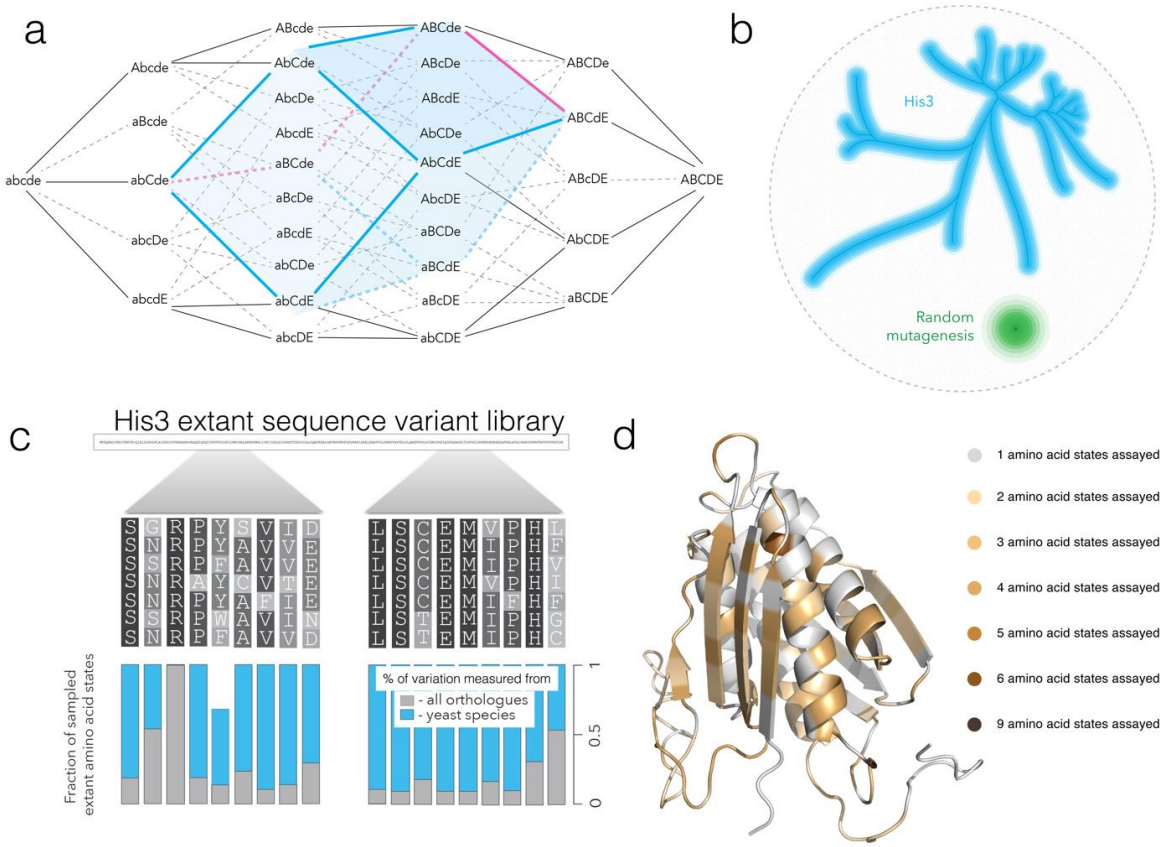


Figure 1. Combinatorial approach to the study of fitness landscapes. A fitness landscape is the representation of fitness for all possible genotypes composed of a specific set of loci. **a,** Following Figure 1 from Sewall Wright ref. [6] consider the genotype space consisting of 5 loci, each with two allele states (lower and uppercase letters). The entire genotype space is 5-dimensional consisting of 2^5 genotypes. Given two genotypes found in extant species (abCde and ABCdE in this example) surveying combinations of extant alleles substantially reduces the dimensionality, concomitantly reducing the number of genotypes to assay. The surveyed area (blue cube) considers all combinations of allele substitutions that have occurred in the course of evolution between the two sequences (red line), avoiding the sampling of combinations with less relevance to the evolutionary trajectory (black genotypes). **b,** Given the entire multidimensional genotype space (black circle) our approach considers an N-dimensional subspace consisting of the combinatorial set of amino acid states form extant species (blue line represents the yeast phylogeny and the surrounding blue space represents combinations of extant amino acids). By contrast, random mutagenesis studies consider only a local segment of the genotype space surrounding a specific genotype (green circle). **c**, A multiple alignment of orthologous sequences of His3 for segment 2 for which we incorporated almost all extant amino acid states from 21 yeast species (blue bars) and 10-100% extant states from a set of 396 orthologues (grey bars). **d**, The predicted structure of His3p with amino acid residues that were substituted in our library.

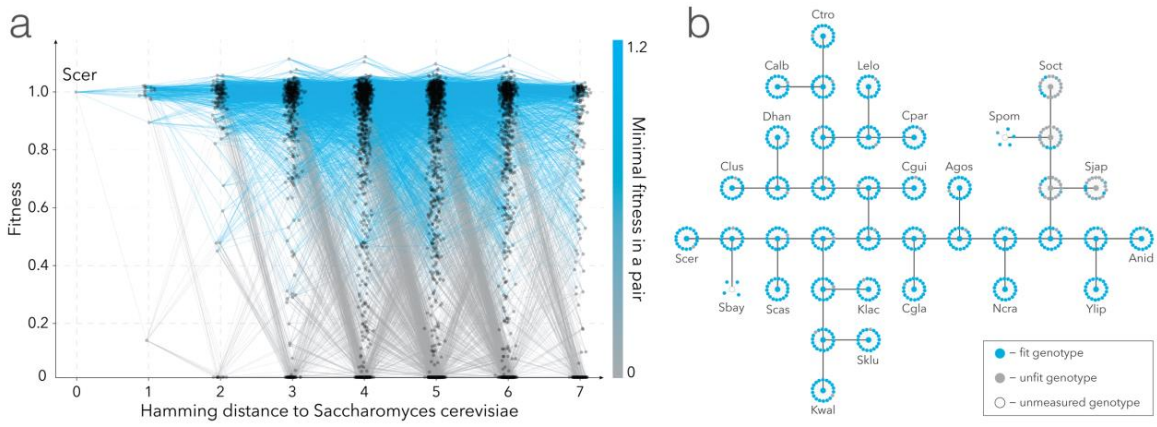


Figure 2. Visual representations of the fitness landscape. a, The fitness landscape for all assayed genotypes in segment 7. Nodes represent unique amino acid sequences with edges connecting those separated by a single amino acid substitution. Colour saturation represents the minimum fitness of the two connected nodes. **b,** On the yeast phylogeny (black lines), fitness of ancestral and extant nodes and genotypes one substitution away from these nodes are shown.

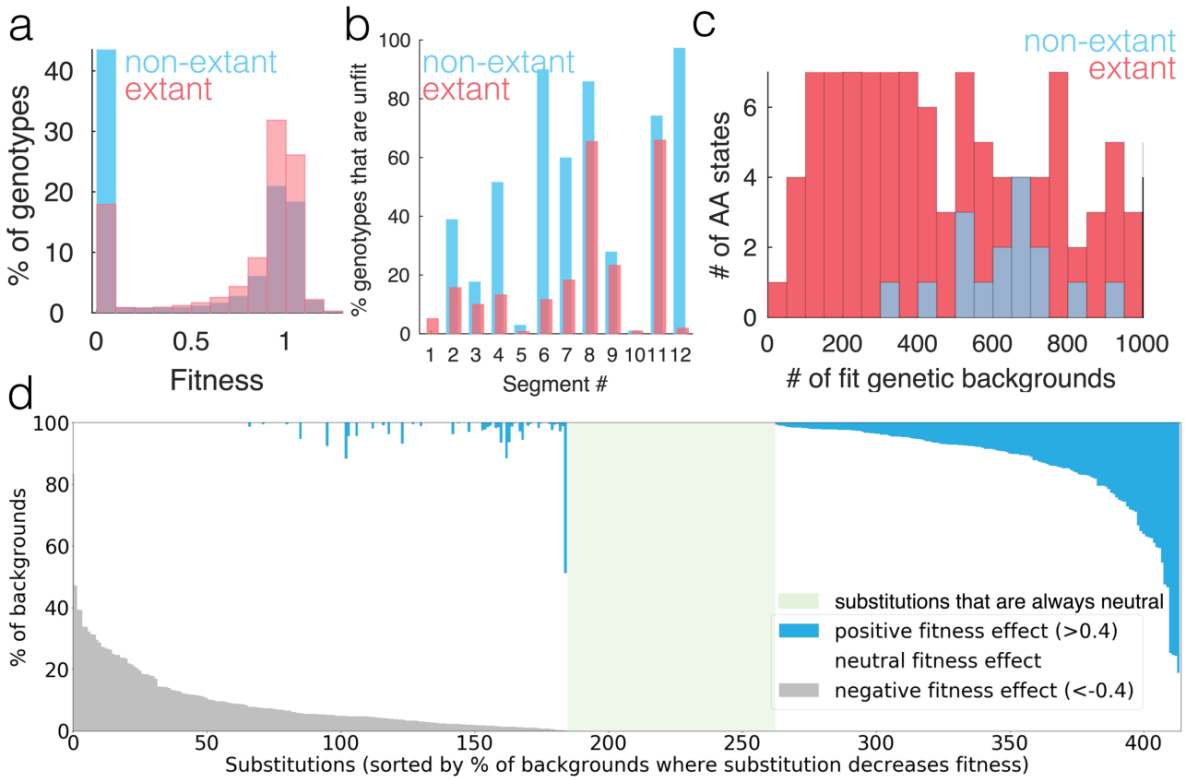


Figure 3. Fitness distributions. a, The distribution of fitness for genotypes composed of combination of extant amino acid states (red) and non-extant amino acid states (blue) at the same positions. **b**, The fraction of unfit genotypes per segment among genotypes consisting entirely from extant amino acid states (red) and those incorporating non-extant amino acid states (blue). **c**, The number of genotypes with high fitness that incorporates specific amino acid states. For each amino acid state, the number of genetic backgrounds that contain that amino acid state and are fit (fitness > 0) are shown. **d**, The percent of backgrounds in which a specific substitution is neutral (white), beneficial (blue) or deleterious (grey). The region marked in green shows substitutions that never have large effects (> 0.4) on fitness. Beneficial and deleterious effects are shown only if the frequency for a given substitution was higher than the false discovery rate.

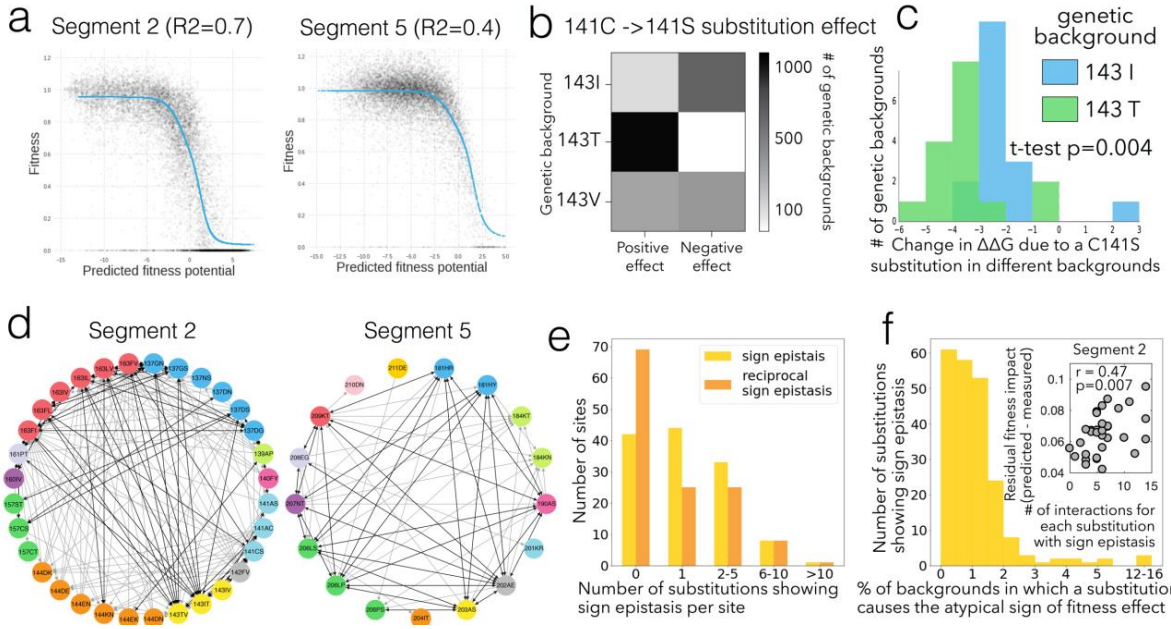


Figure 4. Epistasis and the His3 fitness landscape. a, Fitness as a function of predicted fitness potential (blue curve). **b,** Substitution C->S at site 141 in segment 2 more frequently has a positive effect on fitness in the background of T at site 143, a negative effect in the background of 143I and is equally likely to be strongly deleterious or strongly beneficial in the background of 143V. **c,** Predicted change in free energy upon a C141S substitution in all genetic backgrounds with either an I or T at 143 and that are fewer than four mutations away from *S. cerevisiae*. **d,** A network depiction of sign epistasis between amino acid substitutions. Colour coded sites with reciprocal sign epistasis (black lines) and unidirectional interactions (grey arrows) are shown. **e,** Distribution of the number of substitutions at each site under sign (yellow) and reciprocal sign (orange) epistasis. Sites with 0 interactions do not exhibit sign epistasis. **f,** The fraction of genotypes in which the substitution under sign epistasis has the less frequent effect on fitness. The inset shows that the ability of an additive model to accurately predict the fitness impact of a substitution decreases with increasing dimensionality of sign-epistasis.

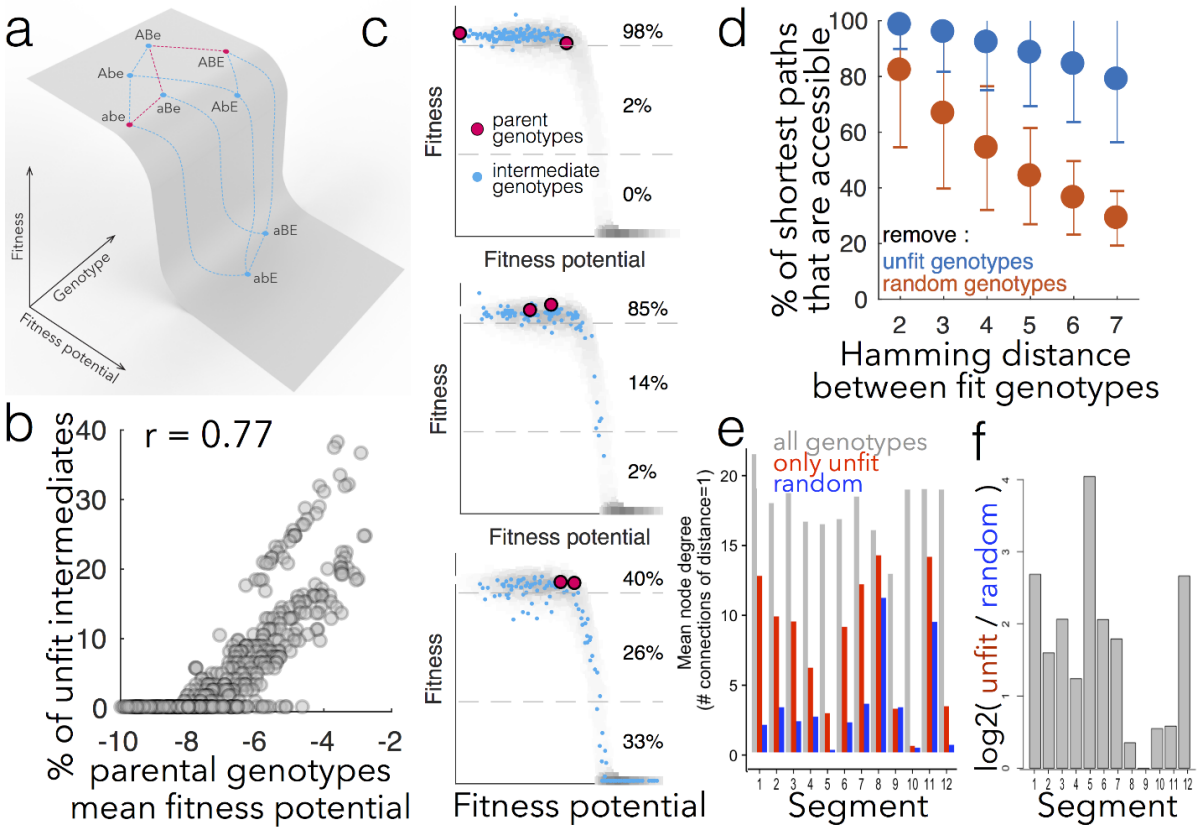
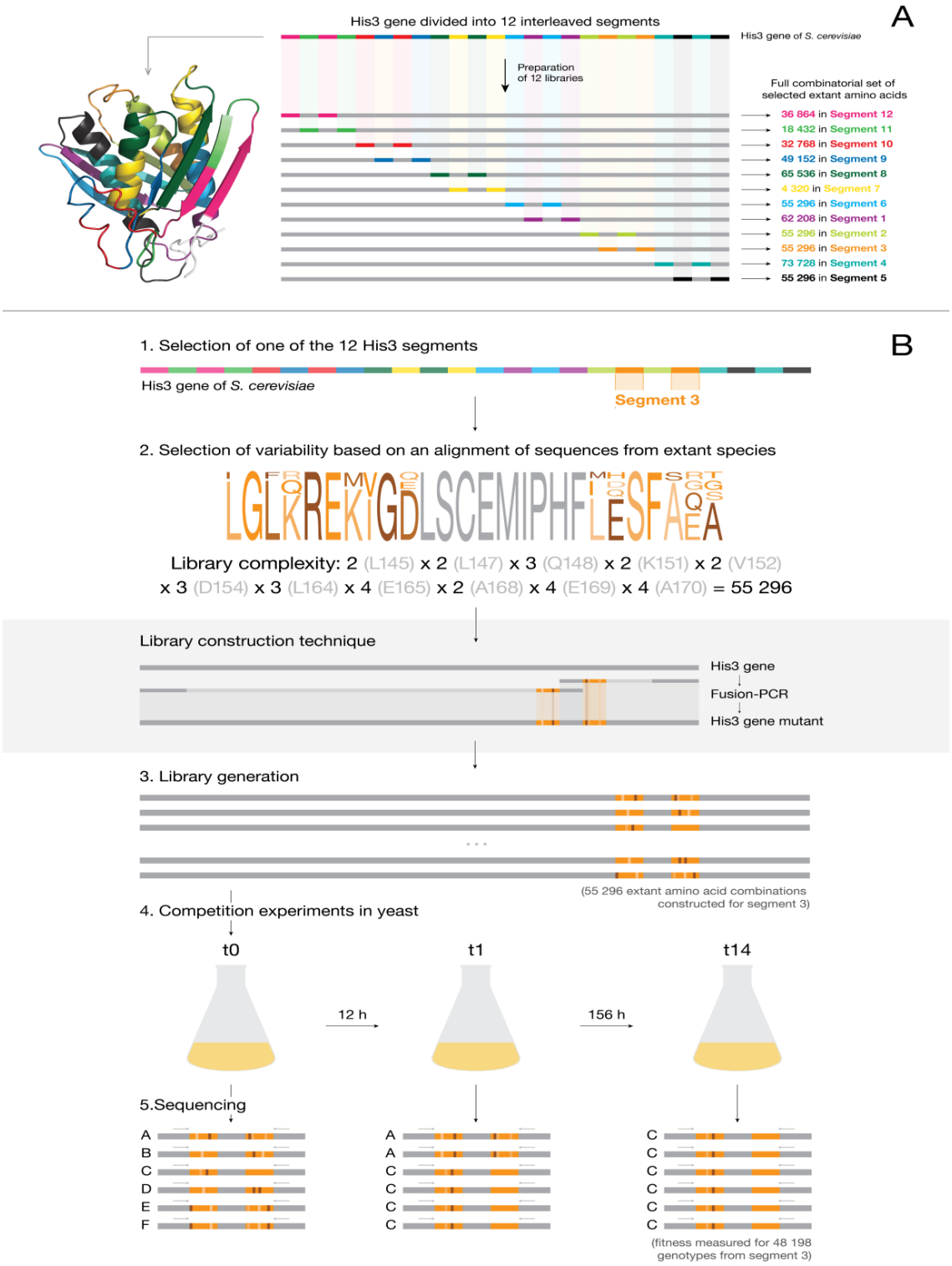


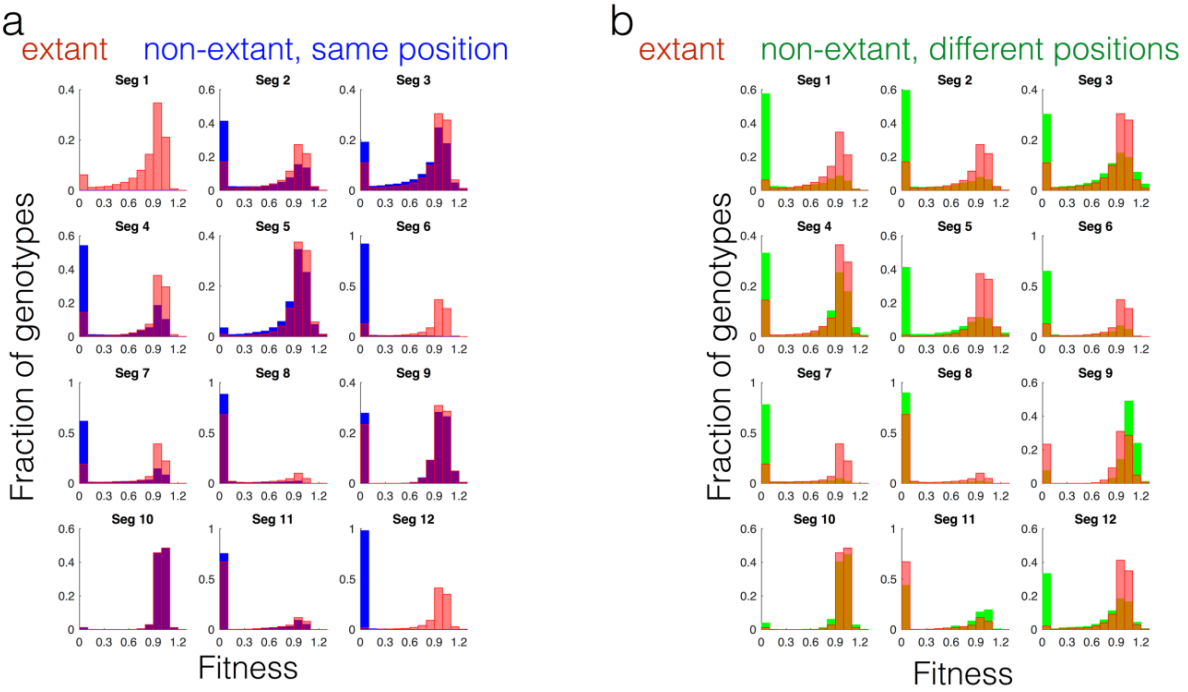
Figure 5. Analysis of evolutionary pathway accessibility. a, A threshold fitness potential function can lead to some paths being inaccessible between two genotypes of high fitness (abe, ABE) if the joint contribution of several alleles to the fitness potential (abE, aBE) leads to the fitness potential below the threshold. **b,** The fraction of unfit intermediate genotypes between two fit genotypes as a function of their average fitness potential. **c**, The grey area represents all genotypes in segment 7. When two fit genotypes (red dots) have high fitness potential, many paths between them will be accessible because many intermediate genotypes will also have high fitness potential and fitness (blue dots). **d**, The fraction of accessible shortest paths between two fit genotypes with unfit genotypes from data (blue) or the same number of randomly selected genotypes (orange), shown as a function of Hamming distance between two fit genotypes. Error bars are standard deviation. **e**, On a graph with genotypes connected by edges if they are one substitution apart we calculate the degree of connectivity (number of edges for each node) for all genotypes (grey), only unfit (fitness = 0) genotypes (red) and a graph with the same number of nodes as the graph with unfit genotypes but with nodes chosen at random (blue). **f**, For all segments except 9 the graph of only unfit genotypes is more interconnected than random graphs with the same number of genotypes.

Extended Data Figures

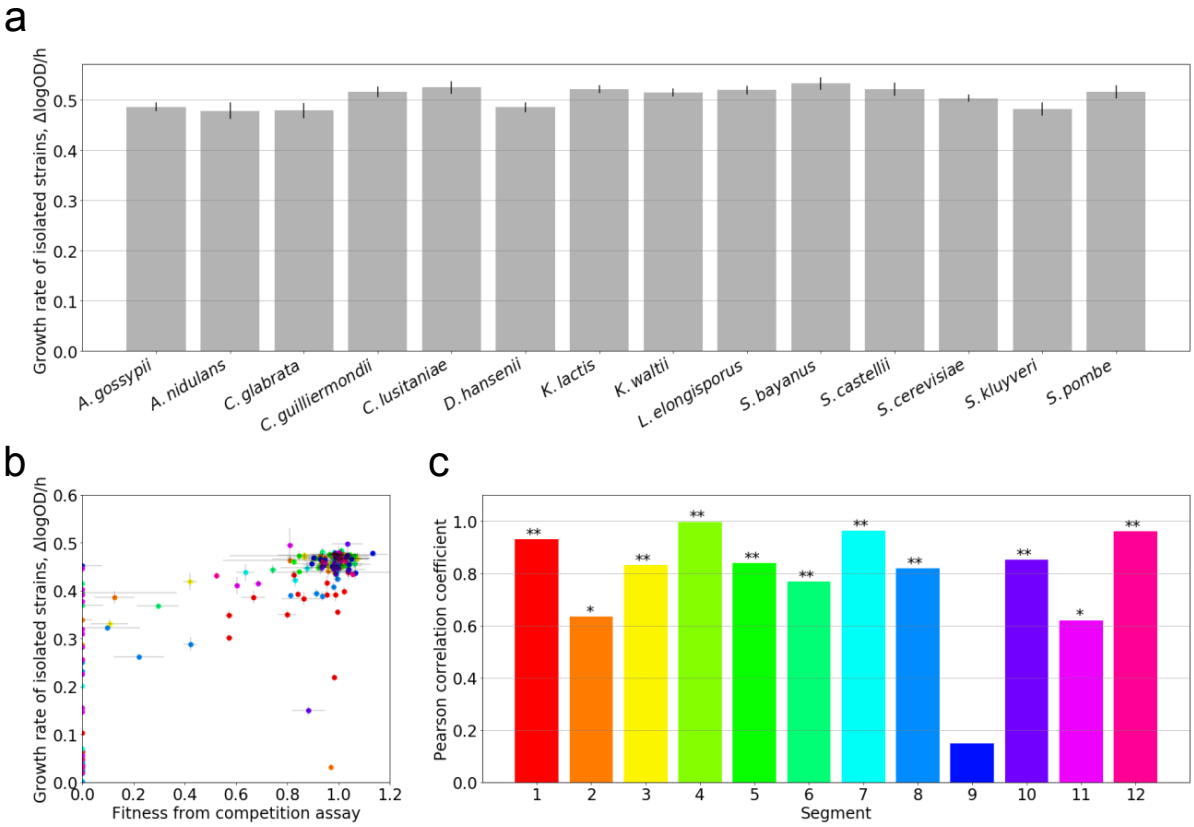


Extended Data Figure 1. Experimental design. a, The sequence of the His3 protein from *S. cerevisiae* was separated into 12 independent segments of similar lengths, such that the full combinatorial set of extant amino acid substitutions was less than 100,000 possible genotypes. These segments represented different combinations of structural elements of the His3 protein structure. **b**. For each of the 12 segments from His3, we selected extant amino acid states using a multiple alignment of His3 orthologues from 396 species, preferentially incorporating states from

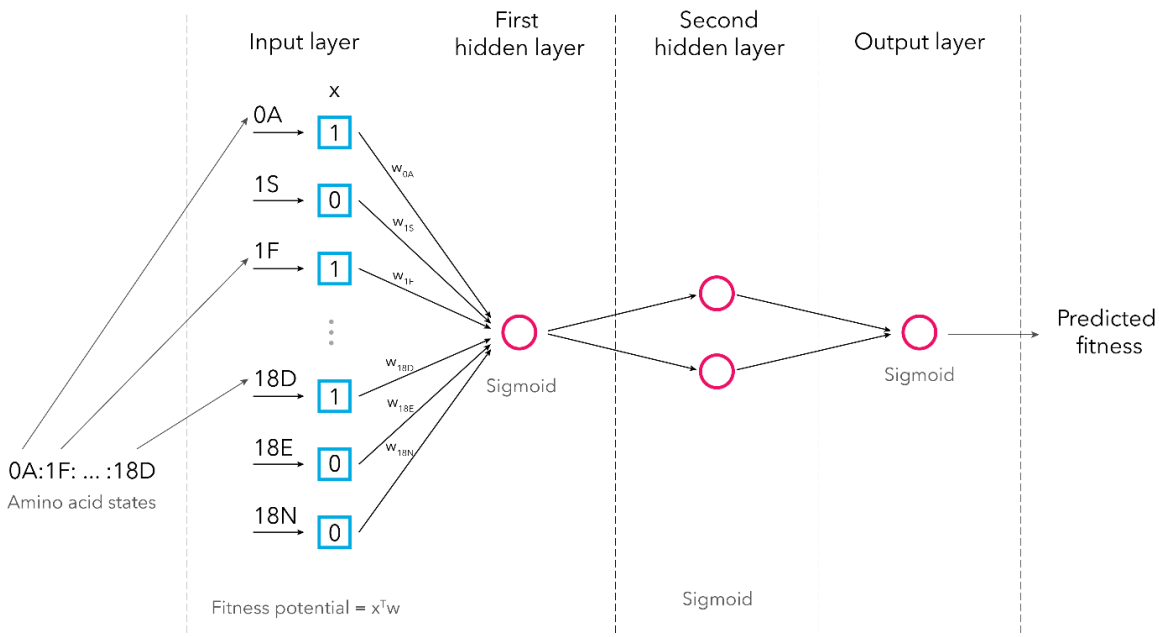
21 yeast species, the variability is shown in section3 as an example. Mutant degenerate codon libraries were constructed by fusion PCR of two synthesized variable halves of each segment. These high-complexity plasmid libraries were transformed into haploid His3 knockout *S. cerevisiae* strain. The growth rate of yeast carrying different extant amino acid state combinations in His3 gene was measured using serial batch culture in the absence of histidine with 12 hours between ~100-fold dilutions. To estimate the fitness of yeast mutants their relative abundance was measured at three points: in the initial population before selection (t0), in the population after 12 hours of growth in the selective medium (t1), and in the final population after 168 hours of growth in the selective medium (t14). To assess the fitness of individual mutants the segments from three populations were amplified and sequenced. The relative abundance of each sequence was used as a proxy for abundance of the associated yeast mutant, which in turn determines its fitness.



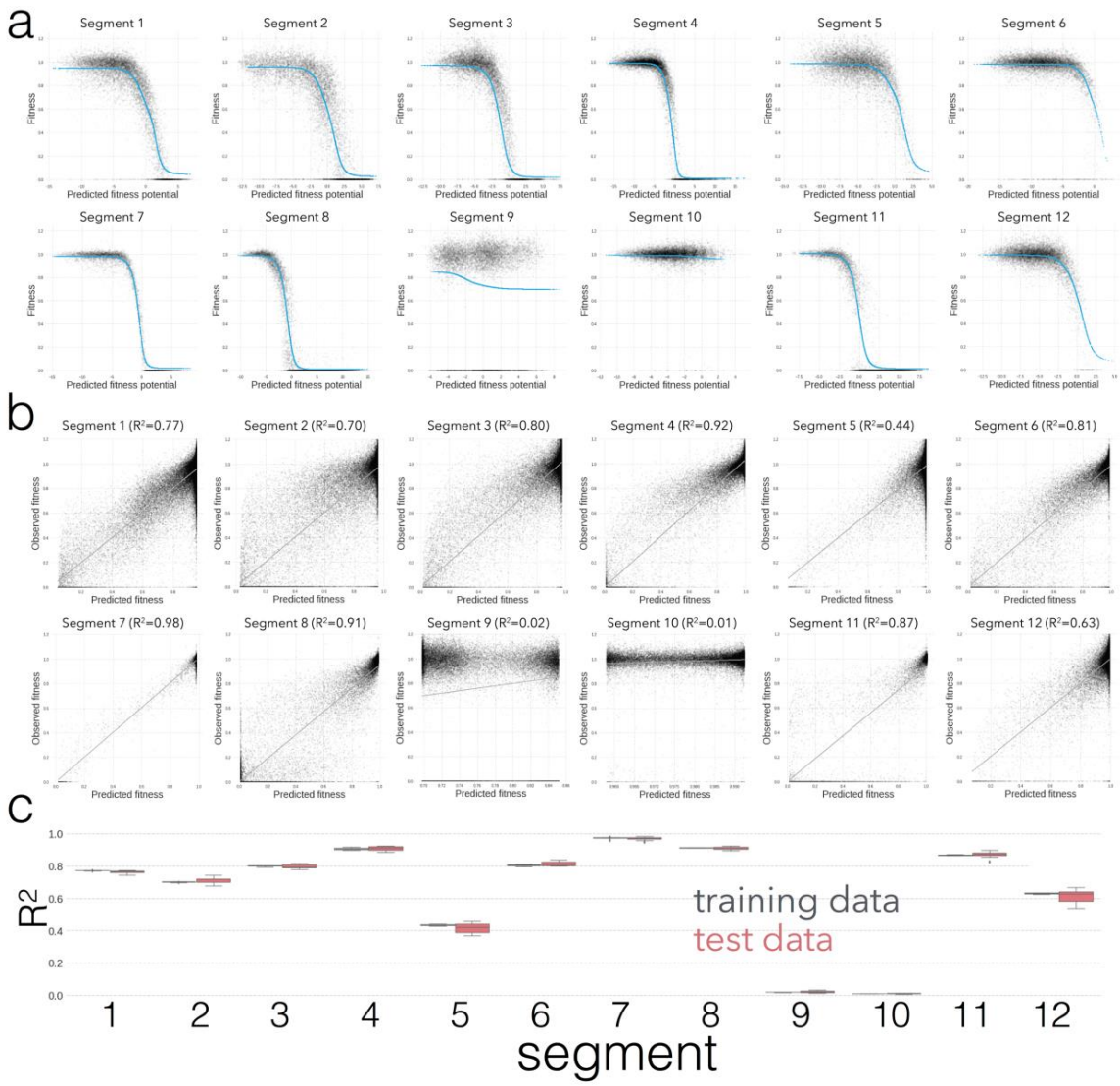
Extended Data Figure 2. Segment-specific fitness distributions for extant and non-extant amino acid states. a, The fitness distribution for each segment for genotypes consisting only of extant amino acid states (red) or that contain one or more non-extant amino acid states (blue)only at positions with a substitution in the extant library. **b,** The fitness distribution for each segment for genotypes consisting only of extant amino acid states (red) and genotypes with mutations at other positions in that segment (green).



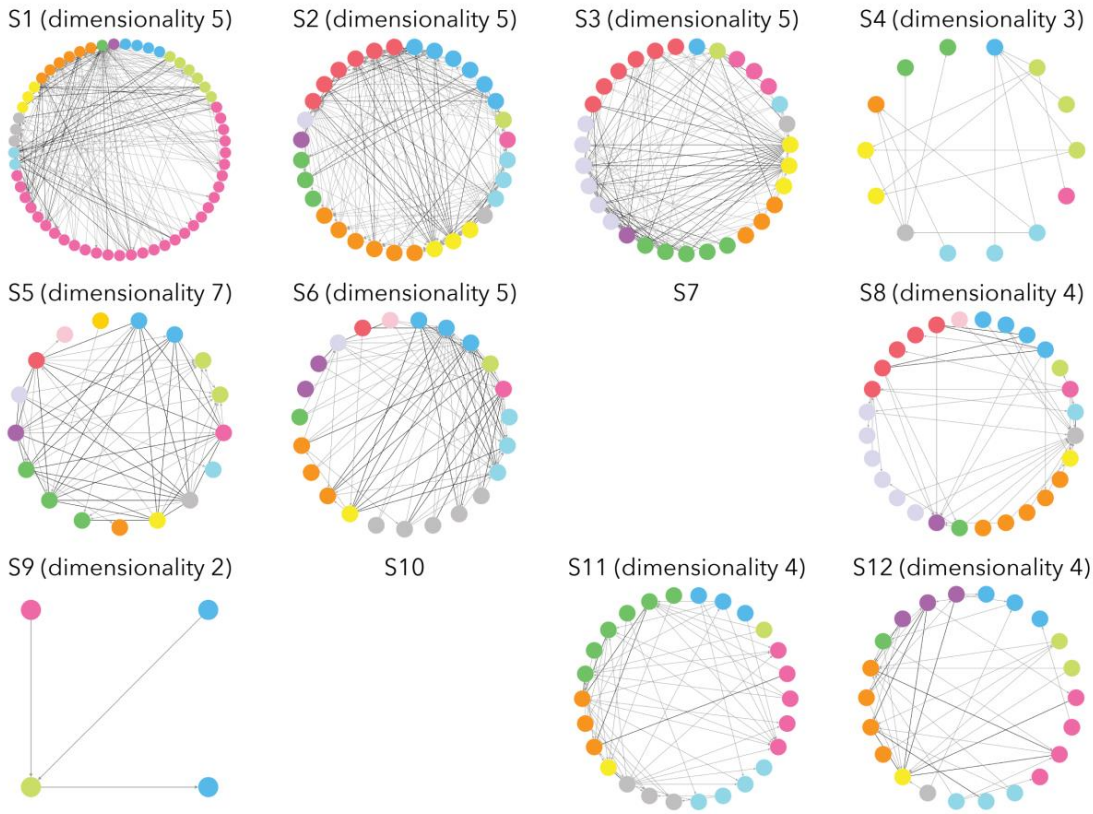
Extended Data Figure 3. Manual measurements of growth rates of isolated strains.a,His3 orthologues from different species complement a *Δhis3* deletion. Growth rates of fourteen *Δhis3* strains of *S*. *cerevisiae*, each of which contains a plasmid with the whole His3 orthologous gene from another yeast species. **b,** Comparison of fitness values from the pooled competition assay with growth rates of isolated strains as measured in 96 well plates in a plate reader. **c,** Pearson correlation coefficients between fitness values from competition and growth rates of isolated strains for each segment. Error bars for growth rates show s.e.m. from 6 replicas. Error bars for fitness from competition show standard error of best-fit parameter from fitness estimation, ** signifies p-value < 0.005, * signifies p-value <0.05.



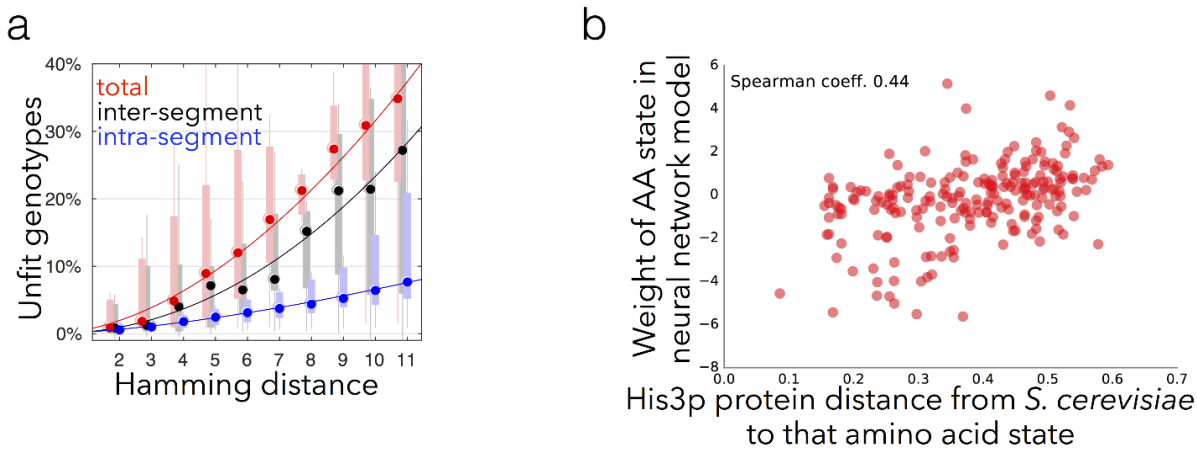
Extended Data Figure 4. Schematic representation of the deep learning approach. Each genotype was encoded as a binary vector (x). During training, each of the substitutions was assigned a weight (w_i), comprising a vector of weights (w). The multiplication of these two vectors is the fitness potential of the genotype. After going through three layers, each with a sigmoid activation function, the predicted genotype fitness was obtained.



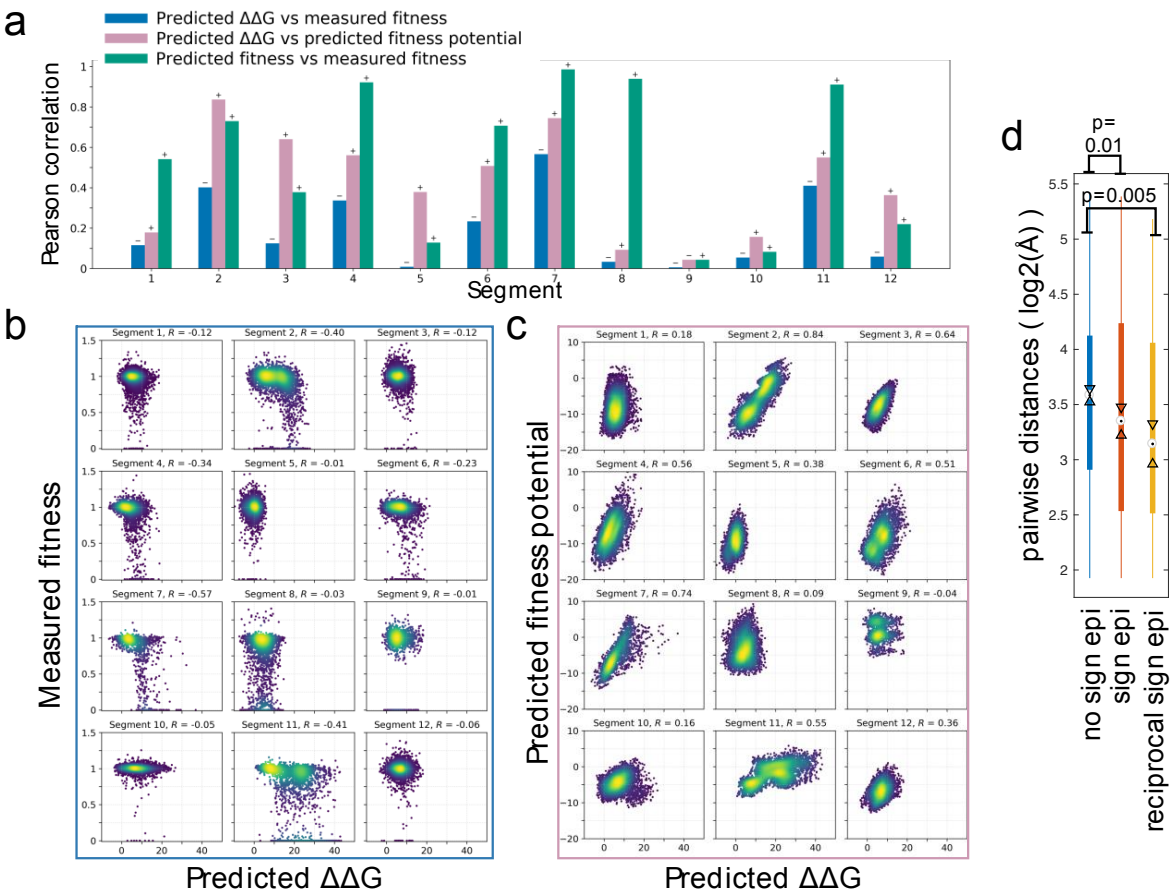
Extended Data Figure 5. The fitness potential of the His3 fitness landscape. a, Fitness potential predicted by the neural network as a function of the measured fitness for all 12 segments. **b**, The correlation between the fitness predicted by the fitness potential and the measured fitness. **c,** Training and test R^2 for each segment for 20-fold cross-validation.



Extended Data Figure 6. Sign epistasis dimensionality graphs for all twelve segments. Each node represents a substitution, with multiple substitutions at the same site having the same colour. Substitutions under reciprocal sign epistasis are indicated by black lines while grey arrows indicate unidirectional sign epistasis.



Extended Data Figure 7. Decoupling inter- and intra-segmental epistasis. a, The fraction of unfit genotypes between *S. cerevisiae* and any other genotype consisting of extant amino acid states with high (blue) or any (red) fitness, and genotypes in the latter but not the former category (black) as a function of the Hamming distance between the two boundary genotypes. Points indicate median, the bars and lines indicate 50% of the genotypes and genotypes 2.7 sigmas from the mean, respectively. **b,** The neural network model assigns higher weights to amino acid states that that first occur in His3 orthologues farther from *S cerevisiae*, indicating the presence of intrasegmental interactions.



Extended Data Figure 8. Protein stability and the fitness potential. a, A comparison of correlation coefficients between predicted and measured values across segments. **b,c**, correlations between the estimated impact of substitutions on free energy (ΔΔG), fitness potential and fitness. ΔΔG correlates better with fitness potential than with fitness. **d**, Pairs of sites that exhibit sign (connected by a light edge in Extended Data Figure 6) and those that exhibit reciprocal sign epistasis (connected by a dark edge in Extended Data Figure 6) are closer together in the His3p structure that randomly chosen non-connected pairs of positions that exhibit sign epistasis.

865

866 **References**

867

868 1. Dean, A. M. & Thornton, J. W. Mechanistic approaches to the study of evolution: the functional
869 synthesis. *Nature Rev. Genet.* **8**, 675–688 (2007).
870

871 2. Kogenaru, M., de Vos, M. G. J., & Tans, S. J. Revealing evolutionary pathways by fitness
872 landscape reconstruction. *Crit. Rev. Biochem. Mol. Biol.* **44**, 169–174 (2009).
873

874 3.Mackay, T. F. C. Epistasis and quantitative traits: using model organisms to study gene-gene
875 interactions. *Nature Rev. Genet.* **15**, 22–33 (2014).
876

877 4. de Visser, J. A. G. M. & Krug, J. Empirical fitness landscapes and the predictability of evolution.
878 *Nature Rev. Genet.* **15**, 480–490 (2014).
879

880 5. de Visser, J. A. G. M., Cooper, T. F., & Elena, S. F. The causes of epistasis. *Proc. Biol. Sci.*
881 **278**, 3617–3624 (2011).
882

883 6. Wright, S. The roles of mutation, inbreeding, crossbreeding and selection in evolution. *Proc.*
884 *Sixth Int. Congr. Genet.* **1**, 356–366 (1932).
885

886 7. Franke, J., Klözer, A., de Visser. J. A. & Krug, J. Evolutionary accessibility of mutational
887 pathways. *PLoS Comput. Biol.* **7**, e1002134 (2011).
888

889 8. Szendro, I. G., Schenk, M. F., Franke, J., Krug, J. & de Visser, J. A. G. M. Quantitative analyses
890 of empirical fitness landscapes. *J. Stat. Mech.* P01005 (2013).
891

892 9. Schenk, M. F., Szendro, I. G., Salverda, M. L. M., Krug, J., de Visser, J. A. G. M. Patterns of
893 epistasis between beneficial mutations in an antibiotic resistance gene. *Mol. Biol. Evol.* **30**, 1779–
894 1787 (2013).
895

896 10. Ferretti. L. et al. Measuring epistasis in fitness landscapes: The correlation of fitness effects of
897 mutations. *J. Theor. Biol.* **396**, 132–143 (2016).
898

899 11. de Visser, J. A., Hoekstra, R. F., & van den Ende, H. An experimental test for synergistic
900 epistasis and its application in Chlamydomonas. *Genetics* **145**, 815–819 (1997).
901

902 12. Fowler, D. M. et al. High-resolution mapping of protein sequence-function relationships.
903 *Nature Methods* **7**, 741–746 (2010).
904

905 13. Jacquier, H. et al. Capturing the mutational landscape of the beta-lactamase TEM-1. *Proc. Natl*
906 *Acad. Sci. USA* **110**, 13067–13072 (2013).
907

908 14. Melamed, D., Young, D. L., Gamble, C. E., Miller, C. R. & Fields, S. Deep mutational
909 scanning of an RRM domain of the Saccharomyces cerevisiae poly(A)-binding protein. *RNA* **19**,
910 1537–1551 (2013).
911

912 15. Roscoe, B. P., Thayer, K. M., Zeldovich, K. B., Fushman, D. & Bolon, D. N. Analyses of the
913 effects of all ubiquitin point mutants on yeast growth rate. *J. Mol. Biol.* **425**, 1363–1377 (2013).
914

915 16. Firnberg, E., Labonte, J. W., Gray, J. J. & Ostermeier, M. A comprehensive, high-resolution
916 map of a gene's fitness landscape. *Mol. Biol. Evol.* **31**, 1581–1592 (2014).
917
918 17. Olson, C. A., Wu, N. C. & Sun, R. A comprehensive biophysical description of pairwise
919 epistasis throughout an entire protein domain. *Curr. Biol.* **24**, 2643–2651 (2014).
920
921 18. Bank, C., Hietpas, R. T., Jensen, J. D. & Bolon, D. N. A systematic survey of an intragenic
922 epistatic landscape. *Mol. Biol. Evol.* **32**, 229–238 (2015).
923
924 19. Puchta, O., Cseke, B., Czaja, H., Tollervey, D., Sanguinetti, G., and Kudla, G. Network of
925 epistatic interactions within a yeast snoRNA. *Science*. **352**, 840–844 (2016).
926
927 20. Li, C. Qian, W. Maclean, C. J. & Zhang, J. The fitness landscape of a tRNA gene. *Science* **352**,
928 837–840 (2016).
929
930 21. Sarkisyan, K.S., et al. Local fitness landscape of the green fluorescent protein. *Nature* **533**,
931 397–401 (2016.)
932
933 22. Hietpas, R. T., Bank, C., Jensen, J. D. & Bolon, D. N. A. Shifting fitness landscapes in response
934 to altered environments. *Evolution* **67**, 3512–3522 (2013).
935
936 23. Parera, M. & Martinez, M. A. Strong epistatic interactions within a single protein. *Mol. Biol.*
937 *Evol.* **31**, 1546–1553 (2014).
938
939 24. Podgornaia, A. I. & Laub, M. T. Pervasive degeneracy and epistasis in a protein-protein
940 interface. *Science* **347**, 673–677 (2015).
941
942 25. Anderson, D. W. McKeown, A. N. & Thornton, J. W. Intermolecular epistasis shaped the
943 function and evolution of an ancient transcription factor and its DNA binding sites. *Elife* **4**, e07864
944 (2015).
945
946 26. Meini, M. R., Tomatis, P. E., Weinreich, D. M. & Vila, A. J. Quantitative description of a
947 protein fitness landscape based on molecular features. *Mol. Biol. Evol.* **32**, 1774–1787 (2015).
948
949 27. Bank, C., Matuszewski, S., Hietpas, R. T. & Jensen, J. D. On the (un)predictability of a large
950 intragenic fitness landscape. *Proc. Natl. Acad. Sci. USA*. 113, 14085–14090 (2016).
951
952 28. Poelwijk, F.J., Kiviet, D.J., Weinreich, D.M., & Tans, S.J. Empirical fitness landscapes reveal
953 accessible evolutionary paths. *Nature* **445**, 383–386(2007).
954
955 29. Gong, L. I. Suchard, M. A.& Bloom, J.D. Stability-mediated epistasis constrains the evolution
956 of an influenza protein. *Elife* **2**, e00631(2013).
957
958 30. Gong, L. I.& Bloom, J.D. Epistatically interacting substitutions are enriched during adaptive
959 protein evolution. *PLoS Genet*. **10**, e1004328(2014).
960
961 31. Starr, T. N. & Thornton, J. W. Epistasis in protein evolution. *Protein Sci*. **25**, 1204–1218
962 (2016).
963
964 32. Kumar, A. Natarajan, C. Moriyama, H. Witt, C.C. Weber, R.E. Fago, A.& Storz, J.F. Stability-
965 mediated epistasis restricts accessible mutational pathways in the functional evolution of avian
966 hemoglobin. *Mol. Biol. Evol.* **34**, 1240–1251(2017).

33. Tufts, D. M., *et al*. Epistasis constrains mutational pathways of hemoglobin adaptation in high-altitude pikas. *Mol. Biol. Evol.* **32**, 287–298 (2015).

34. Eyre-Walker A. & Keightley P. D. The distribution of fitness effects of new mutations. *Nat. Rev. Genet.* 8, 610–618 (2007).

35. Draghi, J. A. & Plotkin, J. B. Selection biases the prevalence and type of epistasis along adaptive trajectories. Evolution 67, 3120–3131 (2013).

36. Lunzer, M., Miller, S. P., Felsheim, R. & Dean, A. M. The biochemical architecture of an ancient adaptive landscape. Science 310, 499–501 (2005).

37. Palmer, A. C., Toprak, E., Baym, M., Kim, S., Veres, A., Bershtein, S. & Kishony, R. Delayed commitment to evolutionary fate in antibiotic resistance fitness landscapes. *Nat. Commun.* **6**, 7385 (2015).

38. Lalić, J. & Elena, S. F. Magnitude and sign epistasis among deleterious mutations in a positive-sense plant RNA virus. *Heredity* **109**, 71–77 (2012).

39. Keefe, A. D. & Szostak, J. W. Functional proteins from a random-sequence library. *Nature* **410**, 715–718 (2001).

40. du Plessis, L. Leventhal, G.E. & Bonhoeffer, S. How good are statistical models at approximating complex fitness landscapes? *Mol. Biol. Evol.* **33**, 2454–2468(2016).

41. Neme, R., Amador, C., Yildirim, B., McConnell, E. & Tautz, D. Random sequences are an abundant source of bioactive RNAs or peptides. *Nat. Ecol. Evol.* **1**, 0217 (2017).

42. Lum, P. Y. Edwards, S.& Wright, R. Molecular, functional and evolutionary characterization of the gene encoding HMG-CoA reductase in the fission yeast, *Schizosaccharomycespombe*. *Yeast* **12**, 1107–1124(1996).

43. Kondrashov, A.S. Sunyaev, S. & Kondrashov, F. A. Dobzhansky-Muller incompatibilities in protein evolution. *Proc. Natl. Acad. Sci. USA* **99**, 14878–14883 (2002).

44. Kvitek, D. J. & Sherlock, G. Reciprocal sign epistasis between frequently experimentally evolved adaptive mutations causes a rugged fitness landscape. *PLoS Genet.***7**, e1002056 (2011).

45. Breen, M. S., Kemena, C., Vlasov, P. K., Notredame, C. & Kondrashov, F. A. Epistasis as the primary factor in molecular evolution. *Nature* **490**, 535–538 (2012).

46. Gavrilets, S. Evolution and speciation on holey adaptive landscapes. *Trends Ecol. Evol.* **12**, 307–312(1997).

47. Salverda, M. L. M., Dellus, E., Gorter, F. A., Debets, A.J. M., van der Oost, J. Hoekstra, R. F. Tawfik, D. S. & de Visser, J.A.G. M. Initial mutations direct alternative pathways of protein evolution. *PLoS Genet.* **7**, e1001321(2011).

48. Weinreich, D. M., Delaney, N. F., de Pristo, M. A. & Hartl, D.L. Darwinian evolution can follow only very few mutational paths to fitter proteins. *Science* **312**, 111–114(2006).

1019 49. Weinreich, D. M. Lan, Y. Wylie, C. S.&Heckendorn, R. B. Should evolutionary geneticists
1020 worry about higher-order epistasis? *Curr. Opin. Genet. Dev.* **23**, 700–707(2013).
1021
1022 50. Poelwijk, F.J. Krishna, V. & Ranganathan, R. The Context-Dependence of Mutations: A
1023 Linkage of Formalisms. *PLoS Comput. Biol.* **12**, e1004771(2016).
1024
1025 51. Sailer, Z. R. & Harms, M. J. High-order epistasis shapes evolutionary trajectories.
1026 *PLoSComput. Biol.* **13**, e1005541 (2017).
1027
1028 52. Kondrashov, F. A. & Kondrashov, A. S. Multidimensional epistasis and the disadvantage of
1029 sex. *Proc. Natl. Acad. Sci. USA* **98**, 12089–12092 (2001).
1030
1031 53. Crow, J. F. Twenty-five years ago in Genetics: identical triplets. *Genetics* **130**, 395–398(1992).
1032
1033 54. Crow, J. F. & Kimura, M. Efficiency of truncation selection. *Proc. Natl. Acad. Sci. USA* **76**,
1034 396–399(1979).
1035
1036 55. Weinreich, D. M. Watson, R. A. & Chao, L. Perspective: Sign epistasis and genetic constraint
1037 on evolutionary trajectories. *Evolution* **59**, 1165–1174(2005).
1038
1039 56. Song, C. Gallos, L. K. Havlin, S. & Makse, H. A. How to calculate the fractal dimension of a
1040 complex network: the box covering algorithm. *J. Stat. Mech.* **2007**, P03006(2007).
1041
1042 57. Orr, H. A. The population genetics of speciation: the evolution of hybrid incompatibilities.
1043 *Genetics* **139**, 1805–1813 (1995).
1044
1045 58. Bridgham, J. T., Ortlund, E. A. & Thornton, J. W. An epistatic ratchet constrains the
1046 direction of glucocorticoid receptor evolution. *Nature* **461**, 515–519(2009).

1047 59. De Pristo, M. A. Weinreich, D. M. & Hartl, D. L. Missense meanderings in sequence space:
1048 a biophysical view of protein evolution. *Nature Rev. Genet.* **6**, 678–687 (2005).

1049 60. Bershtein, S., Segal, M., Bekerman, R., Tokuriki, N. & Tawfik, D. S. Robustness epistasis
1050 link shapes the fitness landscape of a randomly drifting protein. *Nature* **444**, 929–932 (2006).

1051 61. Tamura, K., Stecher, G., Peterson, D., Filipski, A. & Kumar, S. MEGA6: Molecular
1052 Evolutionary Genetics Analysis version 6.0. *Mol Biol Evol.* **30**, 2725–2729 (2013).

1053 62. Magoč, T. & Salzberg, S. L. FLASH: fast length adjustment of short reads to improve
1054 genome assemblies. *Bioinformatics* **27**, 2957–2963 (2011).

1055 63. Zorita, E., Cuscó, P. & Filion, G. J. Starcode: sequence clustering based on all-pairs search.
1056 *Bioinformatics* **31**, 1913–1919(2015).

1057 64. Glorot, X. & Bengio, Y. Understanding the difficulty of training deep feedforward neural
1058 networks. in Proceedings of the 13th International Conference on Artificial Intelligence and
1059 Statistics (2010).
1060
1061 65. Yang, J. & Zhang, Y. I-TASSER server: new development for protein structure and function
1062 predictions. *Nucleic Acids Res.* **43**, W174–W181(2015).

1063 66. Yang, J., Wang, Y. & Zhang, Y. ResQ: An Approach to Unified Estimation of B-Factor and
1064 Residue-Specific Error in Protein Structure Prediction. *J. Mol. Biol.* **428,** 693–701 (2016).
1065

1066 67. Park, H., Bradley, P., Greisen, P. Jr., Liu, Y., Mulligan, V. K., Kim, D.E., Baker, D. &
1067 DiMaio, F. Simultaneous optimization of biomolecular energy functions on features from small
1068 molecules and macromolecules. *J. Chem. Theory Comput.* **12**, 6201–6212(2016).

1069 68. Nivón, L. G., Moretti, R. & Baker, D. A Pareto-optimal refinement method for protein
1070 design scaffolds. *PLoS One* **8**, e59004(2013).

1071 69. Conway, P., Tyka, M. D., DiMaio, F., Konerding, D. E. & Baker, D. Relaxation of backbone
1072 bond geometry improves protein energy landscape modeling. *Protein Sci.* **23**, 47–55(2014).

1073