doi:10.1038/nature17995

Local fitness landscape of the green fluorescent protein

Karen S. Sarkisyan[1,2,3,4,5]*, Dmitry A. Bolotin[1,3]*, Margarita V. Meer[4,5], Dinara R. Usmanova[4,5,6], Alexander S. Mishin[1,2], George V. Sharonov[1,7], Dmitry N. Ivankov[4,5,8], Nina G. Bozhanova[1], Mikhail S. Baranov[1,9], Onuralp Soylemez[4,5], Natalya S. Bogatyreva[4,5,8], Peter K. Vlasov[4,5], Evgeny S. Egorov[1], Maria D. Logacheva[9,10,11], Alexey S. Kondrashov[11,12], Dmitry M. Chudakov[1,3], Ekaterina V. Putintseva[1,3], Ilgar Z. Mamedov[1,3], Dan S. Tawfik[13], Konstantin A. Lukyanov[1,2] & Fyodor A. Kondrashov[4,5,14]

[*] authors contributed equally to the work

[1]Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry, Miklukho-Maklaya 16/10, 117997 Moscow, Russia.

[2]Nizhny Novgorod State Medical Academy, Minin Sq. 10/1, 603005 Nizhny Novgorod, Russia.

[3]Central European Institute of Technology, Masaryk University, Brno, 62500, Czech Republic.

[4]Bioinformatics and Genomics Programme, Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, 88 Dr. Aiguader, 08003 Barcelona, Spain.

[5]Universitat Pompeu Fabra (UPF), 08003 Barcelona, Spain.

[6]Moscow Institute of Physics and Technology, Institutskiy Pereulok 9, g.Dolgoprudny 141700, Russia.

[7]Faculty of Medicine, Moscow State University, Lomonosov Avenue 31/5, Moscow 119192, Russia.

[8]Laboratory of Protein Physics, Institute of Protein Research of the Russian Academy of Sciences, 4 Institutskaya Str., Pushchino, Moscow Region 142290, Russia.

[9]Pirogov Russian National Research Medical University, Ostrovitianov 1, Moscow 117997, Russia.

[10]A. A. Kharkevich Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow, 127051, Russia.

[11]Department of Bioinformatics and Bioengineering, Moscow State University, Moscow, 119234, Russia.

[12]Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, Michigan, 48109, USA.

[13]Department of Biological Chemistry, Weizmann Institute of Science, Rehovot 76100, Israel.

[14]Institució Catalana de Recerca i Estudis Avançats (ICREA), 23 Pg. Lluís Companys, 08010 Barcelona, Spain.

S1 Background

Despite considerable interest to the study of fitness landscapes, two obstacles have so far inhibited their in-depth experimental characterization. First, the impact of epistasis, whereby the effect of an amino acid change on protein function depends on the sequence context in which it occurs. If epistasis is common, the effect of multiple amino acid substitutions cannot be inferred from their individual effects and the fitness landscape can only be surveyed by considering combinations of mutations, with the number of possible genotypes becoming too large even for a modest number of mutations[1–5]. Second, the lack of high-throughput methods to assay function and to sequence intact DNA molecules coding an entire protein within the same experiment[2]. Several studies avoided these obstacles by focusing on small regions of the local fitness landscape, considering single mutations[3,6–10], a region of a gene[11–16], or functionally[17–19] or evolutionary[3,4,10,20–23] important sites. In contrast to studies that focused on fitness landscapes across large volumes of sequence space[24,25], we followed experimental studies[3,6,7,9,11,13–15,26,27] by selecting a single functional sequence that corresponds to a fitness peak on the landscape assayed in the laboratory, and explored the phenotypes conferred by sequences in the immediate vicinity of the original wildtype sequence (**Figure 1**).

We selected avGFP F64L sequence (Uniprot: P42212), or wildtype sequence here forth, and prepared its mutant library using error-prone PCR marking each mutant in the library by a unique nucleotide sequence, a molecular barcode. We cloned the library into a bacterial expression vector that contained a non-mutated reference red fluorescent protein and employed our sequencing strategy to sequence the entire length of the GFP with the barcodes (about 800 nt) using the Illumina MiSeq sequencing platform. We then performed fluorescence-activated cell sorting to sort the bacteria expressing our construct into eight populations with different brightness level of 510 nm emission. We sequenced the barcodes in each of the eight populations using Illumina HiSeq, making four replicas of the sorting and sequencing steps (**Extended Data Fig. 1**). The combination of the high throughput assaying of the fluorescence phenotype and our solution for sequencing of the entire gene presents a new approach to the study of fitness landscapes limited mainly by sequencing costs.

S2 Methods and Materials

S2.1 Construct design and expression

The length of the GFP coding sequence exceeds the current length limitations of next generation sequencing. Thus, we developed a strategy that allowed us to sequence the N-terminal and C-terminal halves of every mutant gene in the library independently while keeping the information about which N-terminal sequence corresponds to which C-terminal sequence. We marked every sequence with a unique 20 nucleotide sequence (molecular barcode) inserted after the coding sequence. The library preparation for sequencing included steps that produced either the C-terminal half of the gene or the N-terminal half fused to its barcode (see below).

We have chosen avGFP with a single mutation F64L that is known to enhance folding in *E. coli* at 37°C [ref 28], and optimized codon usage as the wildtype sequence, which we refer to as avGFP throughout the text. We made the avGFP gene by site-directed mutagenesis from EGFP sequence (pEGFP-C1 vector, Clontech) using self-assembly cloning[29]. We introduced an NheI restriction site before the first codon of avGFP, a SpeI restriction site in the middle of the avGFP gene without changing its amino acid sequence, and two stop codons at the end of the coding sequence. Following the stop codons we added a short GC-rich sequence (an annealing site for a primer), an AvrII restriction site and a NotI restriction site. NheI, SpeI and AvrII are the isocaudamer restriction enzymes (i.e. they leave identical sticky ends after digestion).

To assure accurate measurements of green fluorescence that are independent of the variance in the protein expression levels we added a reference non-mutated red fluorescent protein into the construct. We set the following requirements for the reference protein: fast chromophore maturation, monomeric state, lack of a green fluorescent state during chromophore synthesis and a large spectral distance from green fluorescence of avGFP. Based on these criteria we selected the mKate2 far-red fluorescent protein[30].

To choose the optimal design of the vector providing the best correspondence between expression level of the avGFP and mKate2 we tested several variants of the construct using only non-mutated avGFP. First, we tested a bicistronic-like architecture with coding sequences of avGFP and mKate2 separated by a short sequence containing a stop codon. In this case, the influence of transcription variability is eliminated, however the proteins are translated independently. Second, we tested a construct where avGFP and mKate2 were in the same coding sequence separated by a flexible Gly-Ser-rich linker GHGTGSTGSGSSAS, similar to the one used in tdTomato (Clontech). Finally, we tested an architecture with a rigid alpha-helical linker GSLAEAAAKEAAAKEAAAKAAAAS, which is known to eliminate interactions and Förster resonance energy transfer (FRET) between fluorescent proteins[31].

In the fusion constructs we placed mKate2 at the N-terminal part of the construct because such a design allows to express a fully functional mKate2 for any mutation in the C-terminally located avGFP. Furthermore, protein folding is influenced by the N-terminal part, thus a poorly folded avGFP mutant in the N-terminus may affect mKate2 folding and fluorescence, introducing a bias[32].

We measured the ratio of green to red fluorescence for cells expressing one of these three constructs. The bicistronic variant showed large scatter of this ratio signifying that in this case mKate2 brightness cannot be used as a measure of the GFP level of expression. Both versions of the fusion proteins resulted in a minimal level of heterogeneity and the construct with the rigid linker was selected.

To maximize the homogeneity of expression level and the green to red fluorescence ratio we tested different expression vectors: pQE-30 (Qiagen), pBAD (Life technologies) and pRFPCER [ref 33]. We also optimized expression and induction conditions. The best results were achieved when the construct was cloned into the pQE-30 expression vector. *E.coli* was grown on agar plates with LB medium overnight at 37°C followed by 24-hour incubation at 4°C. Then bacterial cells were transferred to liquid LB media and incubated for two hours at room temperature with 1 mM IPTG.

S2.2 Mutant library construction

We used error-prone PCR[34] to prepare a library of avGFP random mutants. Random mutagenesis was performed using an in-house analogue of the Diversity PCR Random Mutagenesis kit (Clontech, *50x mix: 0.2 мM dGTP, 0.2 мM dATP, 1 мM dCTP, 1 мM dTTP*). We have optimized the conditions of the PCR by varying the buffer composition and the number of cycles to obtain on average four mutations per coding region. One of the primers used for PCR contained a stretch of 20 random nucleotides, which introduced the molecular barcode downstream to the GFP coding sequence. The library was digested with NheI and NotI restriction enzymes and subcloned into the construct mKate2-rigid_linker-EGFP /pQE-30 instead of the EGFP gene. The final library diversity was estimated to be about 300,000 clones.

S2.3 Sequencing the library of mutants and DNA from sorted cells

At present, the Illumina MiSeq platform produces reads of up to 300 nucleotides, allowing to sequence up to 600 nucleotides by using a pair-end approach. To obtain the full avGFP sequence (750 bp + 50 bp of barcode and technical sequences) we sequenced the C-terminal half of the gene linked to the barcode and the N-terminal linked to the barcode independently and constructed the full-length sequence using the barcode information (**Extended Data Fig. 1**). We ligated the Illumina adapter by sticky-end ligation without using any amplification in sample preparation, allowing us to eliminate PCR errors.

The C-terminal half of the gene linked to the barcodes was obtained by digesting an aliquot of the plasmid library with SpeI and NotI. To link the N-terminal half of the gene with the corresponding barcodes of the C-terminal we digested the plasmid library with SpeI and AvrII, purified it through a gel, eliminating the short fragment corresponding to the C-terminal half of the gene, and then self-ligated it. Self-ligation was performed at low DNA concentration in the presence of both SpeI and AvrII to avoid intermolecular ligation. The resulting library was transformed into *E. coli*, grown on Petri dishes and purified again. The diversity of the resulting library of self-ligated plasmids was over 10,000,000 clones, exceeding the diversity of the original library by more than an order of magnitude. We then digested the library with NheI and NotI to obtain the N-terminal half of the gene linked with the barcodes.

Illumina TruSeq-like adapters were prepared by extension and digestion of two partially complementary oligonucleotides. Libraries with a low complexity reduce the quality of NGS sequencing, thus, we added between one and four random nucleotides after the TruSeq adapters to randomize the beginning of the reads and introduce shifts in the sequence. We ligated the adapters to the libraries by sticky-end ligation.

S2.4 Sorting

To functionally characterize the avGFP mutant genotypes we performed four replications of fluorescence activated cell sorting (FACS) of our library starting from an independent transformation of our plasmid library into *E. coli* XL1 Blue and passing the bacteria through the expression conditions described above. The sorting was performed using the FACS Aria III cell sorter (BD Biosciences). We set the gate for the red channel (excitation at 561 nm and detection at 670/14 nm), in order to exclude cells with abnormally low or high signal of mKate2 fluorescence. The cells with the red fluorescence in the selected range were sorted into eight populations based on the level of brightness in the green channel (excitation at 405 nm and detection at 510/30 nm). The green gates were selected approximately evenly across the logarithmic scale (**Supplementary Fig. 1**). The absolute values of gate borders were used when estimating the level of fluorescence of each genotype (**Supplementary Information S2**). For each replica, we sorted at least an order of magnitude more cells than the estimated complexity of our library (300,000 unique genotypes).

To avoid cell proliferation the sorted bacteria were kept on ice until the next step. A fixed number of cells with known barcodes (prepared separately, i.e. their barcodes were not present in the sorted library) were added to each of the sorted populations as the controls. Then we precipitated the cells at 7500 rcf and carefully removed most of the supernatant. The cell pellet was resuspended in the leftover of the supernatant (about 20μl) and subjected to several freeze-thaw cycles, incubated for 3 minutes at 95°C. 20 cycles of PCR were performed in order to amplify the barcode-containing fragment. We then used step-out PCR to flank the amplified regions with the Illumina TruSeq-like adapters and sequenced them with Illumina HiSeq.

To confirm that the sorting procedure does not bias the selection of genotypes from our library we have sequenced a sample of the unsorted bacteria. We found that the prevalence of genotypes in sorted and unsorted samples was similar, indicating that there is no significant effect of the folding of avGFP mutants on mKate2 fluorescence.

S2.5 Manual confirmation of stop-codon-containing genotypes and several cases of epistatic interactions

We performed site-directed mutagenesis to confirm some of the results obtained in our high-throughput approach. In particular, using the FACS machine, we measured the fluorescence levels of four genotypes carrying a nonsense mutation, and 14 genotypes that comprise five genotypes with multiple mutations for which we observed negative and positive epistasis in our data. (**Supplementary Table 1**).

S3 Raw data processing

S3.1 Assembly of genotypes for each barcode from MiSeq sequencing

We processed data from the MiSeq sequencing run to reconstruct full-length sequences of GFP and relate each GFP sequence to the corresponding barcode using the following pipeline:

1. The barcode was detected and extracted using a bitap algorithm from MiLib (https://github.com/milaboratory/milib, part of MiXCR software[35]).
2. Barcodes were clustered to rescue barcode sequences with minor errors. Briefly, two barcodes were clustered together if they differed by one deletion, insertion or substitution and had at least a three-fold difference in their abundances (number of reads with exactly the same barcode).
3. Both reads from each pair were aligned to wtGFP sequence using KAligner from MiLib.

4. Mutations and coverage of all positions of wtGFP were collected for each barcode cluster.
5. Clusters containing positions with the coverage less than 3 in the wtGFP sequence were removed.
6. To distinguish sequencing errors from real mutations every nucleotide was assigned a ratio value defined as the number of occurrences of a particular state in the corresponding position divided by coverage at that position.
7. Clusters containing positions with the ratio between 0.75 and 0.25 in the wtGFP sequence were removed.

S3.2 Estimating fluorescence for each genotype from HiSeq sequencing of sorted populations

To determine the genotype distribution across brightness populations we used Illumina HiSeq (single 50 bp reads). For each genotype, the barcode sequence was extracted using bitap algorithm from MiLib.

S3.3 Controls and normalization

A fixed number of cells with known barcodes were added to every population after sorting. These variants passed all sample preparation procedures together with the library being a control for each sample in each replica. When analysing the sequencing data we used these controls to translate the number of reads per barcode to the number of sorted cells. Barcodes with less than three cells across the population samples were removed.

S3.4 Data analysis for estimation of fluorescence levels

For some of the barcodes a bimodal distribution of cells across the fluorescence gate populations was observed. These distributions were not reproduced across experimental replicas, indicating that they represent an artifact of the experimental procedure rather than inherent genotype properties. We fitted each barcode distribution within each replica with two Gaussian distributions using actual values of logarithms of sorting gates boundaries. Thus, the resulting distributions parameters were expressed in actual brightness logarithm values. We filtered out the cases where the log-value of fluorescence of the major Gaussian component was below 0.65, or its sigma exceeded 0.4. When aggregating information from replicas we eliminated those barcodes for which less than three replicas were ±0.45 of the median value calculated across all replicas.

We used two groups of genotypes with known fluorescence levels to estimate false positive and false negative rates in our final dataset. The first group consisted of 2444 barcodes corresponding to the wildtype sequence; two of them conferred a low level of fluorescence, resulting in a false negative rate of 0.08%. The second group included 839 genotypes which incorporate mutations known to disrupt chromophore formation (when G67 or R96 are substituted) or prohibit the protein from the light absorption in the visible spectrum (Y66 mutated to non-aromatic amino acids). Two of these barcodes conferred high level of fluorescence implying a false positive rate of 0.24%. We found eleven instances of sequences with a nonsense mutation conferring a high level of fluorescence. We re-cloned four of these sequences, confirming that their expression produces cells with a measurably high level of fluorescence (**Supplementary Table 1**), suggesting the prevalence of ribosomal read-through of stop codons[36].

S3.5 Building of the final dataset

We excluded genotypes containing insertions, deletions or stop codons, and grouped barcodes with the same nucleotide or amino acid genotypes. The final dataset consisted of 56,086 unique nucleotide genotypes and 51,715 unique amino acid genotypes.

S4 Data analysis

S4.1 Single mutant analysis

Sequences with log-fluorescence < 3.0 have light intensity less than wild-type by $10^{3.72\text{-}3.0} \approx 5$ times (**Figure 2a**). 9.4% of genotypes had such low intensity, which we considered non-fluorescent. To estimate the fraction of deleterious fluorescent mutations we compared log-fluorescence value distributions of 2442 uniquely barcoded wildtype variants and 1114 sequences carrying a single missense mutation (**Figure 2a**). The distribution of single mutants is shifted considerably toward smaller values of log-fluorescence relative to the wildtype distribution (**Figure 2a**). Thus, on average single mutations decrease the fluorescence level slightly but the fluorescence levels of each individual mutation cannot be statistically differentiated from the wildtype. In other words, we cannot distinguish if a specific singleton is neutral or slightly deleterious. But we can estimate the upper limit for the fraction of neutral mutants in the distribution of single mutants. Truly neutral mutants must follow the same distribution as the wildtype. We assume that log-fluorescence values are drawn from the mixture distribution

$$p(F) = p_0 \cdot NeutralDistr + (1 - p_0) \cdot NonNeutralDistr\,,$$

where:

F - observed fluorescence
p_0 - probability that a singleton is neutral (fraction of neutral single mutations)
event A - observed singleton is neutral
event $B = \bar{A}$ - observed singleton is not neutral
$p_0 = P(A)$ - probability of singleton to be neutral (fraction of neutral single mutations)
δ - a range of fluorescence values
$p_\delta^A = P(F \in \delta | A)$ - probability that fluorescence of a neutral mutant falls in δ
$p_\delta^B = P(F \in \delta | B)$ - probability that fluorescence of a non-neutral mutant falls in δ
$p_\delta = P(F \in \delta)$ - probability that fluorescence of a single mutation falls in δ

We then calculate p_0:

$$p_\delta = P(F \in \delta) = P(A) \cdot P(F \in \delta | A) + P(B) \cdot P(F \in \delta | B) = p_0 \cdot p_\delta^A + (1 - p_0) \cdot p_\delta^B$$

$$p_0 = 1 - \frac{p_\delta^A - p_\delta}{p_\delta^A - p_\delta^B}$$

The only unknown parameter in the final equation is p_δ^B, however, it cannot be calculated because the distribution of fluorescence of non-neutral mutants is not known. Thus, we estimate the upper bound of p_0. Let $p_\delta^B = 0$, because it will give the maximum of p_0. Under this assumption the equation can be simplified to:

$$p_0 = \frac{p_\delta}{p_\delta^A}$$

To take into account the uncertainty in the calculation of p_δ^A and p_δ deriving from finite sample sizes, we draw values from their posterior probability distributions to build a distribution for p_0:

$$p_\delta \sim Beta(\frac{1}{2} + K_\delta, \frac{1}{2} + N_\delta - K_\delta)$$

$$p_\delta^A \sim Beta(\frac{1}{2} + K_\delta^A, \frac{1}{2} + N_\delta^A - K_\delta^A)$$

Here, N_δ is the total number of singletons, K_δ the number of singletons with fluorescence values in δ, N_δ^A the total number of different observed wildtype sequences and K_δ^A the number of wild-type observations with fluorescence values in δ. We chose δ to be in the following form $\delta = [F_0;\ +\infty]$. x was chosen to give minimal value of maximum likelihood estimation for p_0:

$$\hat{p}_0 = \frac{\hat{p}_\delta}{\hat{p}_\delta^A} = \frac{K_\delta / N_\delta}{K_\delta^A / N_\delta^A}$$

The calculated value for $F_0 = 3.875$. We then drew from the final distribution of p_0, sampling from distributions of p_δ^A and p_δ ($N = 10^5$), and the cumulative distribution of p_0. The upper 0.01 quantile corresponding to a p-value = 0.01 results in $p_0 \leq 22.6\%$.

S4.2 Measuring epistasis

A genotype with at least one mutation that eliminates fluorescence is expected to be non-fluorescent even without the action of negative epistasis. Similarly, a genotype consisting of mutations with individually negligible effects cannot reveal positive epistasis. Thus, our data can reveal negative and positive epistasis among mutations with weak and strong effects, respectively. We calculated epistasis as the deviation from non-additive effects of single mutations measured in logarithmic scale. Specifically, we determined the effects of each mutation *i* as Δ_i, the difference between its logarithmic level of fluorescence F_i and the log-fluorescence of the wildtype F_{wt} (eq. 1). We then calculated epistasis, *e*, as the difference between the effect of multiple mutations Δ_{mult} and the sum of the effects of each contributing mutation (eq. 2). If the sum of mutational effects, $-\sum_i (F_i - F_{wt})$ from eq. 1, exceeded the possible extrema of our system (calculated as mutational effects of the darkest and brightest mutants) we limited the estimate of the fluorescence to the extrema. The limitations were necessary to prevent the spurious detection of epistasis, which occurs when the sum of the individual mutant effect exceed physical limitations. For example, a genotype cannot have negative fluorescence, so the joint effect of several individually non-fluorescent mutations cannot cause the multiple genotype to show fluorescence lower than the minimal fluorescence observed in our data (to be non-fluorescent).

$$\Delta_i = F_i - F_{wt} \qquad \text{(eq. 1)}$$

$$e = \Delta_{mult} - Thr\left(\textstyle\sum_i \Delta_i\right), \qquad \text{(eq. 2)}$$

where $Thr(x) = \begin{cases} \Delta_{min}, & if\ x < \Delta_{min} \\ x, & if\ \Delta_{min} \leq x \leq \Delta_{max} \\ \Delta_{max}, & x > \Delta_{max} \end{cases}$

$\Delta_{min} = F_{min} - F_{wt}$, where F_{min} – minimum observed fluorescence,

$\Delta_{max} = F_{max} - F_{wt}$, where F_{max} – maximum observed fluorescence

S4.3 Multiple regression

In a non-epistatic model, we modeled log-values of fluorescence, F_j, as a linear combination of effects of individual single mutations: $F_j = \sum_i a_i \delta_{ij}$ for each j-th genotype, where $\delta_{ij} = 1$ if i-th mutation is present in genotype j, and zero otherwise, and a_i is the effect of i-th mutation on log-fluorescence to be fitted from the model. Mutations in the same codon were treated independently. The 51,715 genotypes consisted of different combinations of 1,817 single missense mutations.

In a truncation-selection-line landscape, individual mutations contribute weakly to fitness while the combination of several such mutations reduces fitness drastically once a certain threshold is reached. Therefore, in the non-linear epistatic model individual mutations were assumed to contribute linearly to a fitness potential, p, that can be viewed as some general property of the protein structure, such as protein stability ΔG: $p_j = \sum_i b_i \delta_{ij}$, where b_i is the supposed effect of i-th mutation on p. The log-value of fluorescence was assumed to be a sigmoid function of p centered at $p = 0$.

$$F_j \sim e^{-p_j}/(1 + e^{-p_j})$$

More specifically, p_j was defined from the following equation:

$$F_j = \left[\max(F_j) - \min(F_j) + 0.02\right] \times \left[e^{-p_j}/(1 + e^{-p_j})\right] + \min(F_j) - 0.01$$

To avoid asymptotic values of p, which would cause a division by 0, we added 0.01 to $\max(F_j)$ and subtracted 0.01 from $\min(F_j)$. We have chosen 0.01 as the magnitude of the epsilon value to be subtracted or added because it was the same order of magnitude as our error rate. The fraction of explained variance was the same for epsilon values of 0.001 (84.3%) and 0.0001 (84.2%) and 0.01 (84.7%).

S4.4 Model for estimation of the noise of our data

To estimate the level of noise in our data and its contribution to the calculated levels of epistasis, we fitted the following model using a maximum likelihood method on genotypes that are represented in our data by at least five different molecular barcodes (13018 barcodes, 523 different genotypes):

$$P(\beta_i) \sim (1 - g_0) * TruncatedNormal\left(\hat{\beta}_i, \sigma(\hat{\beta}_i)\right) + \ g_0 * \ Uniform$$

$$\sigma\left(\hat{\beta}_i\right) = c + d * Sigmoid(b * \hat{\beta}_i + a)$$

$$Sigmoid(x) = \frac{1}{1 + \ e^{-x}}$$

TruncatedNormal is a normal distribution truncated at minimal and maximal observed fluorescence values. *Uniform* is a uniform distribution between minimal and maximal observed fluorescence values, $\sigma(\hat{\beta}_i)$ was found to be different for various ranges of fluorescence values, and this dependence was approximated by a sigmoid-based function, with *a*, *b*, *c* and *d* are parameters of this function.

The model consists of two components. The Normal component of the model describes a normally distributed random noise produced by our method, while the Uniform component models possible outliers of unknown nature that we observed for some clonotypes that we used as controls (**Supplementary Information S3.4**). Parameter g_0 determines the ratio between these two components of the model, and after the fitting of the model its value was found to be 0.002, being congruent with our visual estimation based on the distribution of values that we measured for avGFP.

The best fit was provided by a = -5.90, b = 1.27, c = 0.17, d = -0.32 and g_0 = 0.0023.

We used the model to calculate the amount of false positive epistasis, that produced solely by noise and outliers. To determine this value for an N-mutation-clone we randomly sampled N single-mutants (to resemble actual distribution of fluorescence), calculated the expected fluorescence of the N-mutation-clone in the absence of epistasis as the sum of effects of single mutations and modeled measurements of fluorescence values using our model for all single-mutants and the N-mutation-clone. For such set of points we then calculated the level of epistasis (arising only from deviations from real fluorescence values due to noise and outliers). Our results show that the experimentally observed epistatic interactions are far more frequent than those produced by noise (**Supplementary Fig. 4**).

S4.5 Fraction of genotypes unexplained by a unidimensional model

To estimate the lower bound on the fraction of genotypes with unexplained brightness values we used an approach that is described in **Supplementary Information S4.1** These calculations take into account the noise of our measurements and low sample sizes.

Deviation ($\eta = |model\_estimate - observed|$) of the estimated fluorescence value from the observed was used as a main variable for the method (instead of the fluorescence itself, *F*). Distribution of deviations for the fluorescence values measured for the wildtype and dark genotypes, our control set, was used as a null distribution.

The target range δ was taken in the form of $\eta \in [0, \eta_0]$ – events with deviation less than a certain threshold η_0,

N_δ^A – total number of control measurements (wildtype and dark control genotypes),

K_δ^A – total number of controls with deviations less than η_0 (dark controls that had measured fluorescence less than $F_{min} + \eta_0$ and wildtype controls that had fluorescence greater than $F_{wt} - \eta_0$; where F_{min} is the minimal observed fluorescence and F_{wt} the median fluorescence of wildtype genotypes),

N_δ – total number of measured genotypes,

K_δ – number of genotypes where deviation of the model estimate was less than η_0.

The threshold deviation value η_0 was chosen to be 0.5 (less than 1% of events from distribution of independently measured control wild-type clones fall outside this range). For other threshold values employed method also provides statistically significant fractions of unexplained genotypes.

S4.6 Modeling brightness with artificial neural networks

To evaluate the complexity of the observed genotype-phenotype relationship we trained several networks of different complexities. Neural networks for fitting fluorescence values were optimized using the Caffe library[37]. To eliminate the effect of over-fitting we split our dataset into a training and a test set (90% and 10% of observed genotypes, respectively), and measured goodness of fit as an average square deviation (ASD) of predicted log-fluorescence from the observed log-fluorescence in the test set. All computational experiments were performed for five random partitions into the training and test sets. All optimizations were performed in three replicas to let the optimization algorithms to converge from several random initial weights and results with the best ASD for the train set were chosen.

The final architecture of the artificial neural network was chosen in such a way that it had a bottleneck of a size of a single neuron right after the input layer. Transfer function of this bottleneck neuron was linear with a threshold. Thus, the neural network was forced to find a single intermediate value (hidden value) calculated as a sum of weights of substitutions constituting genotype that can be transformed by a monotonic function into fluorescence. The monotonic function was modeled by a small neural subnetwork with weights (defining the form of the function) optimized along with the weights of individual mutations (**Extended Data Fig. 5e**).

Two neurons in the second hidden layer (**Extended Data Fig. 5e**) were enough to model nonlinearity required to transform value from the single layer into the log-fluorescence. Overall, a single value obtained by the addition of individual weights of single mutations constituting a multi-mutation genotype from our experiment was enough to estimate the fluorescence of the genotype, however, this value was transformed by a non-trivial threshold-like monotonic function to obtain the final log-fluorescence (**Figure 4**).

The resulting hidden value for individual singleton genotypes was compared with $\Delta\Delta G$ calculated with the Rosetta. Interestingly, these values are well correlated and connected by a simple linear function (see **Extended Data Fig. 5f**). We fitted this dependence with a linear function, and used its coefficients to draw the nonlinear transformation function learned by the neural network along with dependence of the median fluorescence on the estimated $\Delta\Delta G$ on the same plot.

Dependence of the Rosetta $\Delta\Delta G$ estimation and hidden value from neural network for the same singleton genotypes were fitted with a linear function: $HV = a * \Delta\Delta G + b$. This function was then used to plot a nonlinear transformation function learned by artificial neural network on the $\Delta\Delta G$ axis: $F(\Delta\Delta G) = NL(a * \Delta\Delta G + b)$

S4.7 Path accessibility on the fitness landscapes

We analysed the fraction of available paths on the fitness landscape in the sequence space two mutations away from avGFP. We considered all combinations of two fluorescent double mutants that were present in our data and that differed by four mutations from each other at four different sites. For each combination we constructed a graph (**Supplementary Fig. 3**), where each node represented a genotype and the edges connected the nodes that differed by one missense substitution. A total of 24 unique shortest paths are possible between each such pair of genotypes.

For each graph we calculated how many paths of length four, the shortest possible paths, exist between the two genotypes and how many of them were accessible for evolution. Following Maynard Smith [ref 2] we considered a path as accessible if all of the three intermediate genotypes were neutral or advantageous, that is conferred fluorescence at least as high as points at the ends of the path. In contrast, some authors consider paths as being accessible if all accessible steps increase fitness[3,38,39]. We grouped paths by the average Hamming distance of their nodes from the wildtype, and analysed each group separately taking into account noise of our measurements and sampling errors. For four groups of observed paths (with average hamming distance of 1.2, 1.6, 2.0, 2.4) we calculated the conservative estimate of the minimum proportion of inaccessible paths. We considered the upper bound on all accessible paths using an approach similar to that used to estimate the fraction of non-neutral single mutants (**Supplementary Information S4.1**).

For this analysis we only considered paths that started and ended with genotypes being fluorescent (log-fluorescence > 3.0). We used the difference between minimal brightness at the end-points of the path and the minimal brightness in the intermediate genotypes as a measure of fitness decrease on a path (Δ). Paths with a substantial decrease in fitness were considered inaccessible.

The method from **Supplementary Information S4.1** was applied with the following parameters.

Null distribution for Δ was sampled using the measurement error model (described in **Supplementary Information S4.4**) as follows:

1. $N_\delta^A = 10^6$ random double mutants with log-fluorescence greater than 3.0 were sampled to replicate real distribution of the fluorescence levels at the terminal genotypes of the paths.

2. Five fluorescence values were generated using the measurement model for each of the values sampled on the first step to model all brightness points along the path (two terminal and three intermediate genotypes). Thus, all differences between genotypes points on the same modelled path were solely introduced by measurement error.

3. Difference between minimal brightness in the terminal genotypes and the minimal fluorescence among the intermediate genotypes was calculated for each of modelled paths.

The target range δ was selected as $\Delta \in [\Delta_0, +\infty]$ - events with deviation value greater than Δ_0, which corresponds to -0.3 for which 1% of modelled paths from the null distribution fall below this value.

N_δ^A - total number of generated control values,

K_δ^A - total number of controls with Δ less than Δ_0 ,

N_δ - total number of paths with a specific average Hamming distance,

K_δ - number of paths where the decrease of log-fluorescence Δ was less than Δ_0,

The lower bound fraction of inaccessible paths was 0/596,228,074 (0%), 1,201,621/70,670,872 (1.7%), 13,324/202,308 (6.6%) and 6/174 (3.4%) for paths 1.2 (blue) 1.6 (orange), 2.0 (red), and 2.4 (purple), substitutions away from the wildtype, respectively (colour-coded with **Supplementary Fig. 3**).

S5 Evolutionary data analysis

S5.1 Multiple alignment and phylogeny

The multiple alignment of 95 GFP orthologues was constructed using three different methods MUSCLE[40], T-Coffee Expresso[41]and M-Coffee[42]. The phylogeny for the orthologues was taken from (http://phylot.biobyte.de/) and adjusted by MrBayes 3 [ref 43].

We mapped missense mutations conferring a non-fluorescent phenotype in avGFP to the multiple alignment identifying nine such instances: T62I, Q69R, G91A, I98F, I123F, S205F, L207R, E213K and H217D. Only the positions supported by all three alignments were taken into account and only amino acids found in 55 GFP orthologues emitting in the green spectrum. The ratios of fluorescent/non-fluorescent genotypes from **Supplementary Fig. 3b** are: 11/538, 79/3964, 119/2107, 83/703, 51/209, 27/48, 8/9, 2/3 for genotypes harboring 1, 2, 3, 4, 5, 6, 7 and 8 mutations, respectively.

S5.2 Fitness matrix model

Fitness matrix of a protein considers all fitness states one substitution away from a specific sequence[44] . The fitness matrix has *L* columns and 20 rows representing information about the local fitness landscape around a specific sequence, where *L* is the length of the protein and 20 is the number of all amino acids. A cell of the fitness matrix may have three different dynamic states, *C*, *A* or *B*. State *C*, the *current* state, represents the amino acid currently present at the site of protein sequence. State *A* is an amino acid state that is currently *available* for evolution, not inhibited by negative selection. State B is a *blocked* state, whereby such a state is at present strongly deleterious, but the same substitution may be available in a different genetic background.

An amino acid substitution corresponds to a reciprocal *C* to *A* switch in the fitness matrix at a site where the substitution occurred. The substitution may be epistatic such that as a result of its occurrence some blocked amino acid states became available ($B \rightarrow A$) and some available amino acids became blocked ($A \rightarrow B$). Thus, similar genotypes are expected to have a similar fitness matrix such that a single amino acid substitution would lead to just a few changes in the fitness matrix.

In terms of the fitness matrix model the probability of any substitution $A_{from} \rightarrow A_{to}$ is proportional to the probability for A_{to} to be in *A* state, P_A. Amino acid A_{to} was in state *A* or *B* between the branch where the substitution occurred and the closest node with A_{to} (not *C*, otherwise it would not be the closest node).The same is true for any potential convergent substitution. Let us name the rate of $A \rightarrow B$ switch as γ_1 and the rate of $B \rightarrow A$ switch as γ_2, where rate is the probability for a cell of the fitness matrix to switch its state on the same timeframe as a single amino acid substitution. The dynamics of probability for cell to be *A* can be described with a differential equation $dP_A/dt = -P_A \cdot \gamma_1 + (1 - P_A) \cdot \gamma_2$. If the state was *A* at t=0 then $P_A(t) = \frac{\gamma_2}{\gamma_1+\gamma_2} + \frac{\gamma_1}{\gamma_1+\gamma_2} \cdot e^{-(\gamma_1+\gamma_2)\cdot t}$. The obtained rate of convergent evolution r_k can be fitted with equation $P_A(t)$. Indeed, each potential convergent amino acid state was available immediately before the closest node. The protein distance multiplied by protein length was used as an approximation for *t*. Since r_k is defined as a normalized measure we fitted r_k in the following way: $r_k = c \cdot (\frac{\gamma_2}{\gamma_1+\gamma_2} + \frac{\gamma_1}{\gamma_1+\gamma_2} \cdot e^{-(\gamma_1+\gamma_2)\cdot(k \cdot 0.02 - 0.01)\cdot 199})$, where *c* is a constant. Here *L*=199, the number of sites for which the number of gaps in multiple alignment <10% that were used for r_k calculation. Only those k bins where $N_k \geq 100$ were used assuming that $N_k < 100$ do not have enough statistical power. We obtained $\gamma_1 = 0.009$ and $\gamma_2 = 0.005$.

S5.3 Convergent evolution and epistasis

In an epistatic fitness landscape, given a substitution $A_1 \rightarrow A_2$ in one sequence, the probability that $A_{from} \rightarrow A_{to}$ substitution at the same site in another sequence is convergent ($A_2 = A_{to}$) decreases with the evolutionary distance between the two sequences[25,44–46]. If two sequences are closely related then it is likely that the same substitutions would confer the same fitness in both sequences leading to a high rate of convergent evolution. Conversely, when the two sequences have diverged substantially, the rate of convergent evolution between them would be lower, because the same substitutions would confer different fitnesses in the two sequences. More generally, as two sequences diverge, some amino acid states that were neutral (available) in the ancestral sequence become deleterious (blocked) with some probability, and *vice versa* (see **Supplementary Information S5.2** for details). Thus, we calculated the rate of convergent evolution as a function of protein distance (1 minus the sequence identity) between nodes on the phylogenetic tree, which includes convergence towards extant state, and reconstructed ancestral states.

Rate of evolution was estimated as the number of substitutions divided by the number of sites, in which the substitutions could have occurred. We estimate the rate of convergent evolution in a similar manner by analyzing nonsynonymous substitutions on the phylogenetic tree of GFP. First, we counted all convergent substitutions. Given one reference substitution $A_{from} \rightarrow A_{to}$ another substitution $A_1 \rightarrow A_2$ was defined as convergent to the reference substitution if $A_{to} = A_2$. Second, we calculated the target set of convergent substitutions for the reference $A_{from} \rightarrow A_{to}$ substitution. The target set is the number of all substitutions $A_1 \rightarrow A_2$ that met the following three conditions, if $A_{from} \neq A_2$, if $A_{from} \rightarrow A_2$ could occur with one nucleotide substitution and if the node with the $A_1 \rightarrow A_2$ substitution is the closest node to the reference substitution of all nodes with the A_2 state. In essence, the target set is analogous to the number of sites in the typical measurement of the rate of evolution; in our case, it is the number of missed opportunities for a convergent substitution. The rate of convergent evolution is thus the number of convergent substitutions divided by the target set.

We reconstructed the phylogenetic tree for 95 GFP orthologues using a covarion model with MrBayes 3.2.5 [ref 43]. We then reconstructed the ancestral amino acid states for internal nodes using two different methods. Data shown in **Figure 5** was based on Bayesian reconstruction of ancestral amino acid states using a covarion model as implemented in MrBayes 3.2.5. Data obtained by maximum likelihood approach using marginal reconstruction in CODEML program of PAML 4 [ref 47] yielded the same results as Bayesian approach (**Supplementary Fig. 5**). For both methods of ancestral state reconstruction, for each site of the phylogenetic tree at each node counted the number of most probable amino acid states multiplying the count by the reconstructed probability. We considered data when choosing states with specific cut-off value of the reconstructed probability. The results were broadly congruent for data obtained with different cut-off values (**Supplementary Fig. 5**). In further analysis each substitution was used with a weight calculated as a multiplication of probabilities of reconstructed amino acids in nodes between which substitution occurred (for leafs the probability was equal to 1).

We calculated the number of convergent substitutions and the target sets using the reconstructed ancestral amino acid states, which includes convergence towards reconstructed ancestral amino acids. For each branch where an amino acid state appeared ($A_{from} \rightarrow A_{to}$ substitution occurred) we then found the closest non-descended branch where a substitution in the same site ($A_1 \rightarrow A_2$) has emerged (i.e. $A_{from} \neq A_2$) and calculated the protein distance between those two branches. We then formed a set of all such distances, $\mathcal{N} = \{n_1, n_2, \dots\}$ between nodes with two identical emergent amino acids, $A_{to} = A_2$, or distances between two convergent substitutions. We then formed a set of all distances $\mathcal{X} = \{x_1, x_2, \dots\}$ where $A_{from} \rightarrow A_2$ substitution could occur by one nucleotide substitution from any of the codons coding for A_{from} and A_2. We approximated the distribution of distances in sets $\mathcal{N}$ and $\mathcal{X}$ with a probability density function, defined in discrete 2% intervals, N_k and X_k, where k represents the protein distance. We used the Parzen window density estimator with a rectangular kernel function and bandwidth equals 4%, which was selected based on Silverman's rule[48] . X_k estimates the number of potential convergent substitutions for a specific distance k. N_k represents the number of convergent substitutions for a specific distance k. Therefore, the rate of convergent evolution $r_k = N_k/X_k$ The resulting r_k shown in **Figure 5**.

S6 Structural data analysis

S6.1 Structural analysis

For the structural analysis we used the GPF structure (pdb: 2WUR) with the highest available resolution[49], which included three substitutions to avGFP (I167T, Q80R and K238N). Buried and surface amino acid residues were obtained from [ref 50].

S6.2 ΔΔG prediction using Rosetta

The GFP chromophore geometry was optimized by the density functional method RB3LYP, using the 6-311+G** basis set, a restricted hybrid HF-DFT SCF calculation were performed using Pulay DIIS + Geometric Direct Minimization to get a set of ideal bond lengths and angles and incorporated as a non-canonical amino acid[50]. The crystal structure (pdb: 2wur) was renumbered and preminimized using Relax application[52] with the following flags: -flip_HNQ, -no_optH false, -nstruct 100, -ex1, -ex2, -use_input_sc. The output structure with the lowest total score was used for further calculations.

For calculating ΔΔG we used Rosetta's ddg monomer method[53], specifying one or two mutations at a time (all mutations were sorted to remove nonsense mutations, mutations in the chromophore, positions that are not resolved in the crystal structure and substituted positions (80 and 167)) and using the high resolution protocol flags: -ddg:weight_file soft_rep_design, -ddg:iterations 5, -ddg:dump_pdbs false, -ddg:local_opt_only false, -ddg:min_cst true, -ddg:suppress_checkpointing true, -in:file:fullatom, -ddg:mean false, -ddg:min true, -ddg:sc_min_only false, -ddg:ramp_repulsive true, -ddg:opt_radius 12.0, -score:fa_max_dis 9.0.

S7 Software

Java 8 program with the MiLib library was used for processing of raw sequencing reads to a list of final genotypes as well as distributions of barcodes across fluorescence gate populations. Further analysis were performed with Python 2.7 using Pandas (http://pandas.pydata.org), NumPy (www.numpy.org), SciPy (www.scipy.org) and Matplotlib in the Jupyter (https://jupyter.org) interactive environment using Anaconda (http://continuum.io) software package. The PyMOL Molecular Graphics System (Version 1.7.4 Schrödinger, LLC, https://www.pymol.org) was used for visualization of the protein structure. Optimizations of neural network weights were performed using a Python program utilizing Caffe deep learning framework[37]. All Rosetta runs were performed with weekly release 2015.12.57698.

S8 Data availability

Raw sequencing data were deposited to SRA under BioProject number PRJNA282342. Processed data sets are available at Figshare http://dx.doi.org/10.6084/m9.figshare.3102154.

S9 Supplementary Results

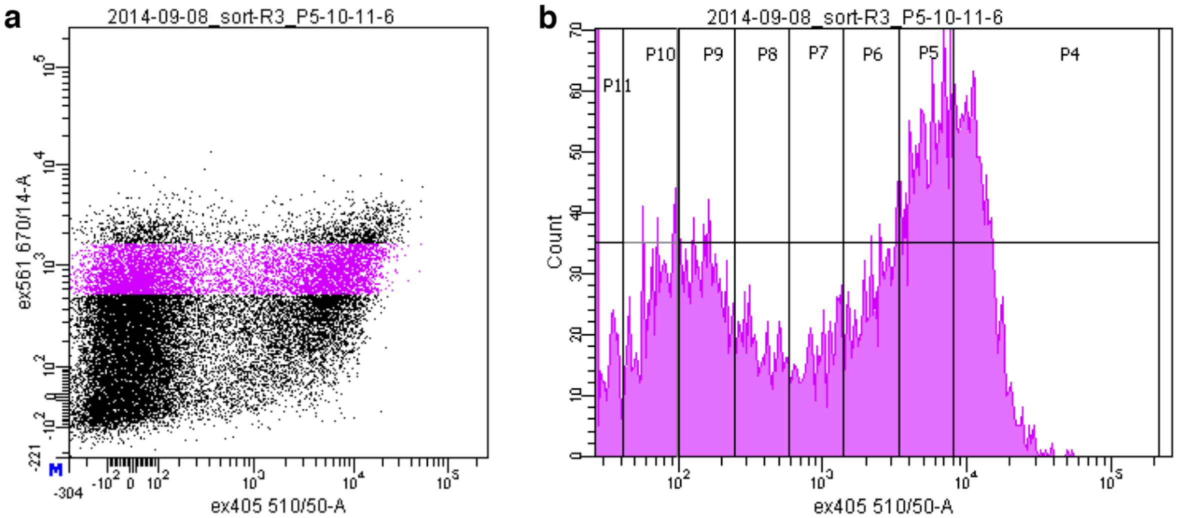
S9.1 Amino acid deletion analysis

We took the set of single amino acid-depleted GFP mutants from [ref 54] excluding instances when the cleavage of three nucleotides affected two amino acid sites. The resulting set consisted of deletion mutants conferring either high or low level of fluorescence (tolerated deletions, Table S1 and non-tolerated deletions, Table S2 from [ref 54], respectively). We related single mutations from our dataset to the tolerated and non-tolerated single amino acid deletions. Neutral missense mutations, those that did not affect fluorescence in our experiment, were not found in codons known to be essential for fluorescence when deleted (**Supplementary Fig. 2**).

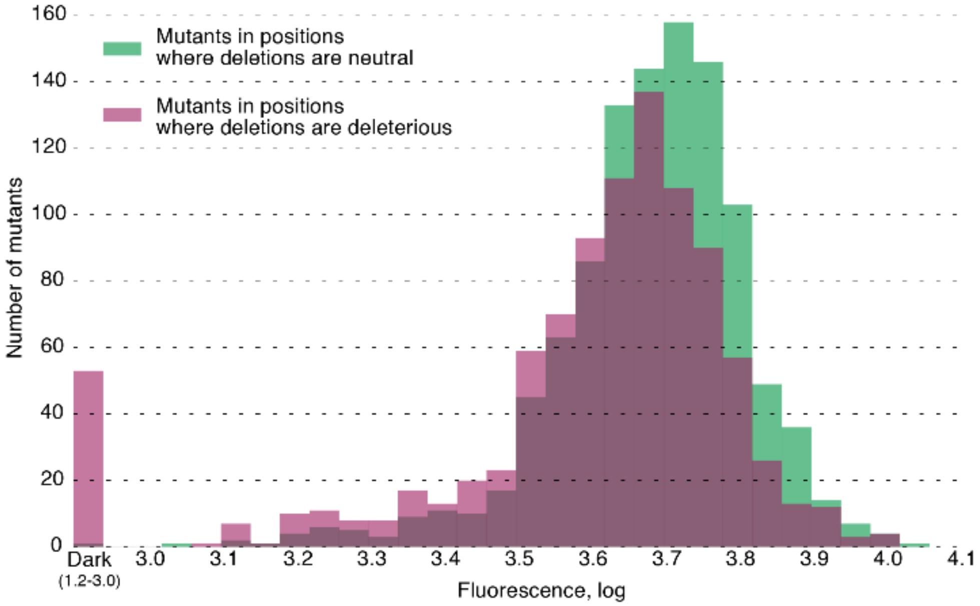
S9.2 Pathway analysis

One of the consequences of epistasis is that it can restrict the number of accessible paths between fit genotypes[3,4,2,19,26,55,56], where a path is deemed inaccessible if at least one intermediate genotype along the path confers a low fitness[3,4,2,10]. To investigate the influence of epistasis on path accessibility in the GFP local fitness landscape, we calculated the fraction of accessible paths between pairs of fluorescent genotypes[4] that each differ from the wildtype by two different mutations (**Supplementary Fig. 3**). We found no evidence that any of the paths that pass through the wildtype sequence were inaccessible (blue path in **Supplementary Fig. 3**). However, some paths that circumnavigated the wildtype sequence were inaccessible. We found that at least 1.7%, 6.6% and 3.4% were inaccessible for paths lying at the average Hamming distance 1.6, 2.0, and 2.4 from the wildtype, respectively (see **Supplementary Fig. 3**). We also analysed paths leading from avGFP to orthologous sequences of GFPs known to emit in the green spectrum. Up to 40% of genotypes in our data that contained only amino acid states found in avGFP orthologues, which are expected to confer a fluorescent phenotype[24], were non-fluorescent (**Supplementary Fig. 3**). The increase of the fraction of inaccessible paths with the distance from wildtype raises the question of whether or not the local shape of the avGFP fitness landscape is representative of the fitness landscapes on the scale of larger evolutionary distances.

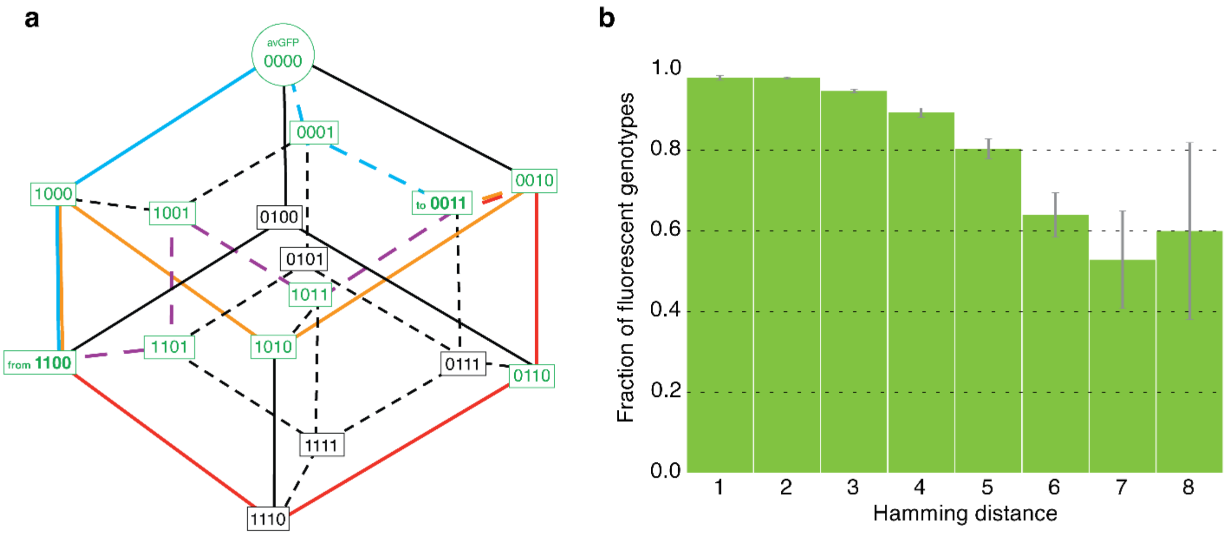
Supplementary Information Figures



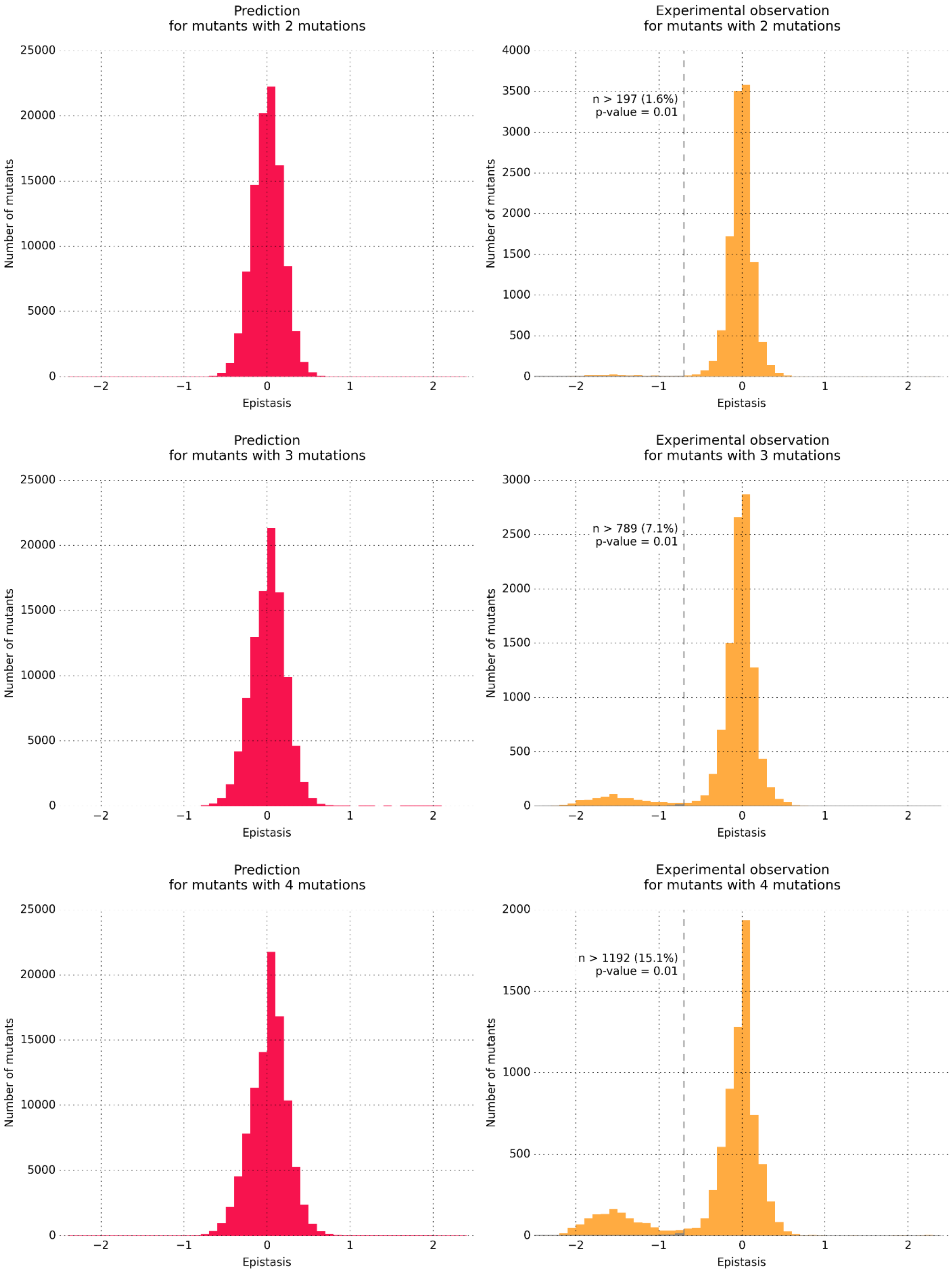
Supplementary Figure 1. **Distribution of gates in FACS**.

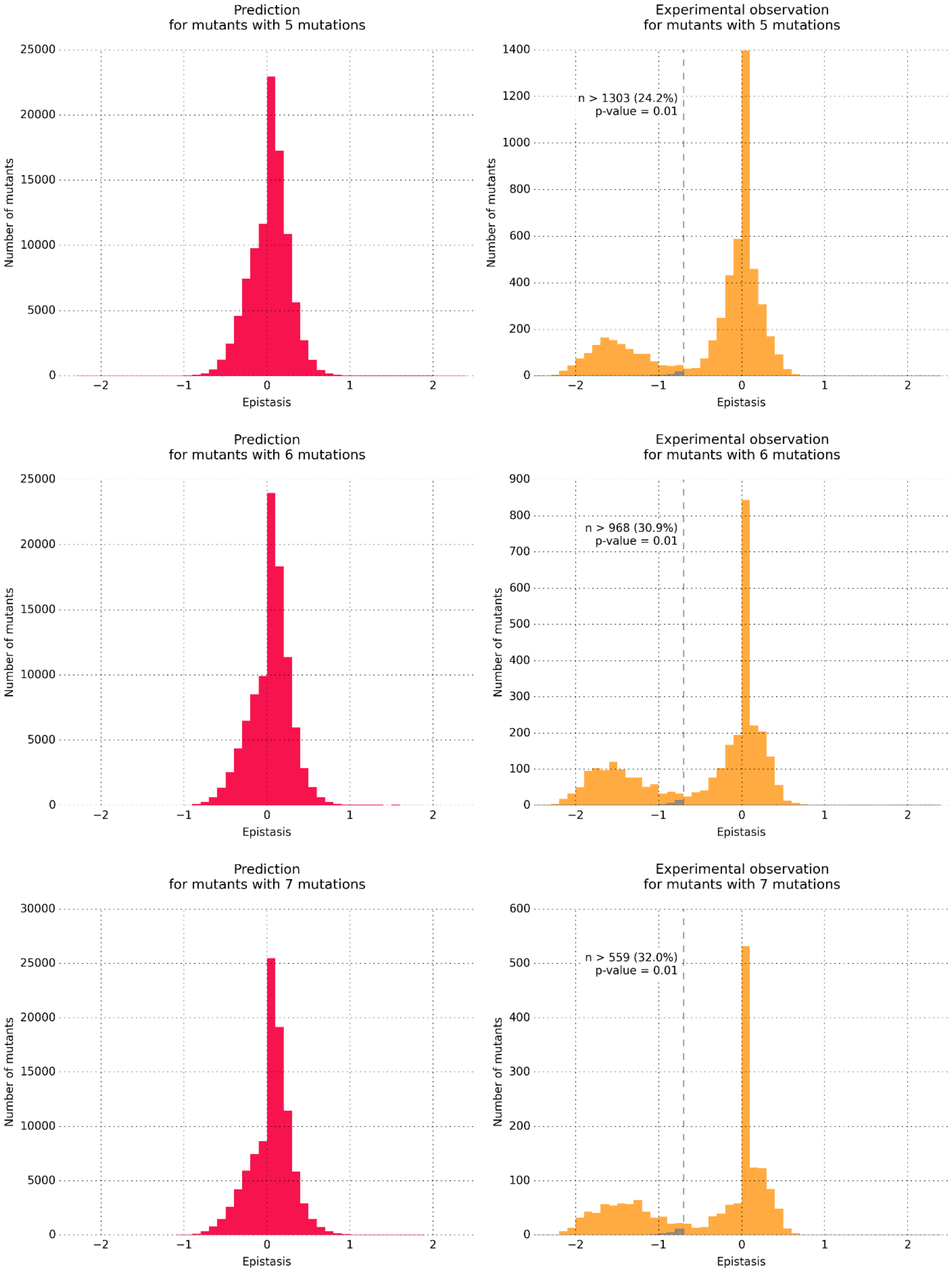


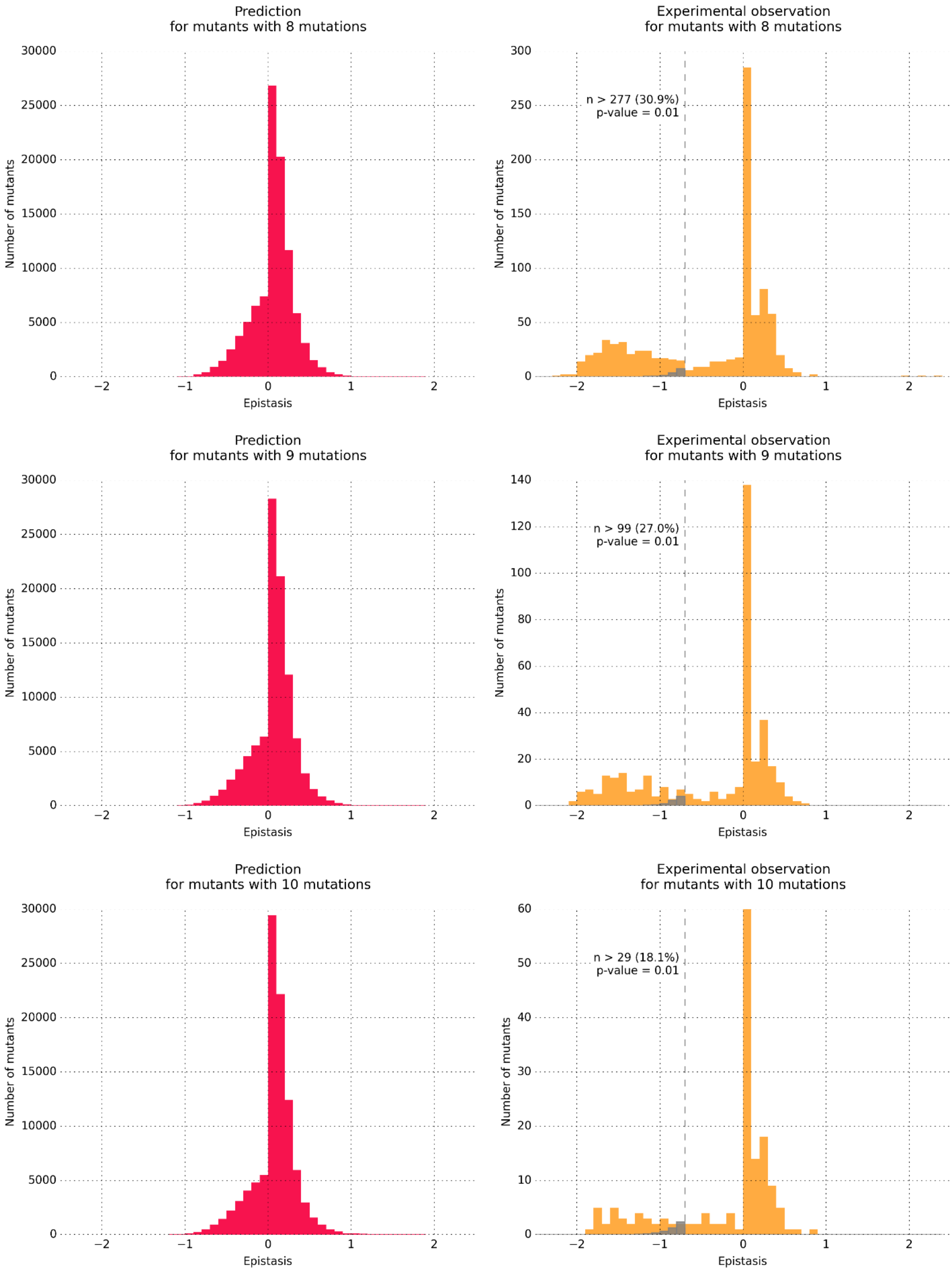
Supplementary Figure 2. Comparison of our data with the previously estimated effects on fluorescence of single amino acid deletions[68]. Distributions of the log-fluorescence observed in our study are given for amino acid sites where deletion was reported[68] to have no influence on fluorescence (green) and for sites where deletion led to loss of fluorescence (red).



Supplementary Figure 3. Path accessibility of the local fitness landscape. a, Shortest paths between two fluorescent genotypes, 1100 and 0011 (green boxes) separated by four mutations. Three intermediate genotypes are fluorescent (green boxes) at accessible paths (blue, orange and violet). Along inaccessible paths (red) at least one of the intermediate genotypes is non-fluorescent (black boxes). **b**, The fraction of non-fluorescent genotypes from all genotypes with mutations identical to amino acid states present in avGFP orthologues emitting in the green spectrum. Bars represent a binomial proportion confidence interval (confidence level 68%).


Prediction
for mutants with 2 mutations
Number of mutants
25000
20000
15000
10000
5000
0
−2
−1
0
1
2
Epistasis
Experimental observation
for mutants with 2 mutations
Number of mutants
4000
3500
3000
2500
2000
1500
1000
500
0
n > 197 (1.6%)
p-value = 0.01
−2
−1
0
1
2
Epistasis
Prediction
for mutants with 3 mutations
Number of mutants
25000
20000
15000
10000
5000
0
−2
−1
0
1
2
Epistasis
Experimental observation
for mutants with 3 mutations
Number of mutants
3000
2500
2000
1500
1000
500
0
n > 789 (7.1%)
p-value = 0.01
−2
−1
0
1
2
Epistasis
Prediction
for mutants with 4 mutations
Number of mutants
25000
20000
15000
10000
5000
0
−2
−1
0
1
2
Epistasis
Experimental observation
for mutants with 4 mutations
Number of mutants
2000
1500
1000
500
0
n > 1192 (15.1%)
p-value = 0.01
−2
−1
0
1
2
Epistasis


Prediction
for mutants with 5 mutations
25000
20000
15000
10000
5000
0
Number of mutants
−2
−1
0
1
2
Epistasis
Experimental observation
for mutants with 5 mutations
1400
1200
1000
800
600
400
200
0
Number of mutants
n > 1303 (24.2%)
p-value = 0.01
−2
−1
0
1
2
Epistasis
Prediction
for mutants with 6 mutations
25000
20000
15000
10000
5000
0
Number of mutants
−2
−1
0
1
2
Epistasis
Experimental observation
for mutants with 6 mutations
900
800
700
600
500
400
300
200
100
0
Number of mutants
n > 968 (30.9%)
p-value = 0.01
−2
−1
0
1
2
Epistasis
Prediction
for mutants with 7 mutations
30000
25000
20000
15000
10000
5000
0
Number of mutants
−2
−1
0
1
2
Epistasis
Experimental observation
for mutants with 7 mutations
600
500
400
300
200
100
0
Number of mutants
n > 559 (32.0%)
p-value = 0.01
−2
−1
0
1
2
Epistasis


Prediction for mutants with 8 mutations

Experimental observation for mutants with 8 mutations

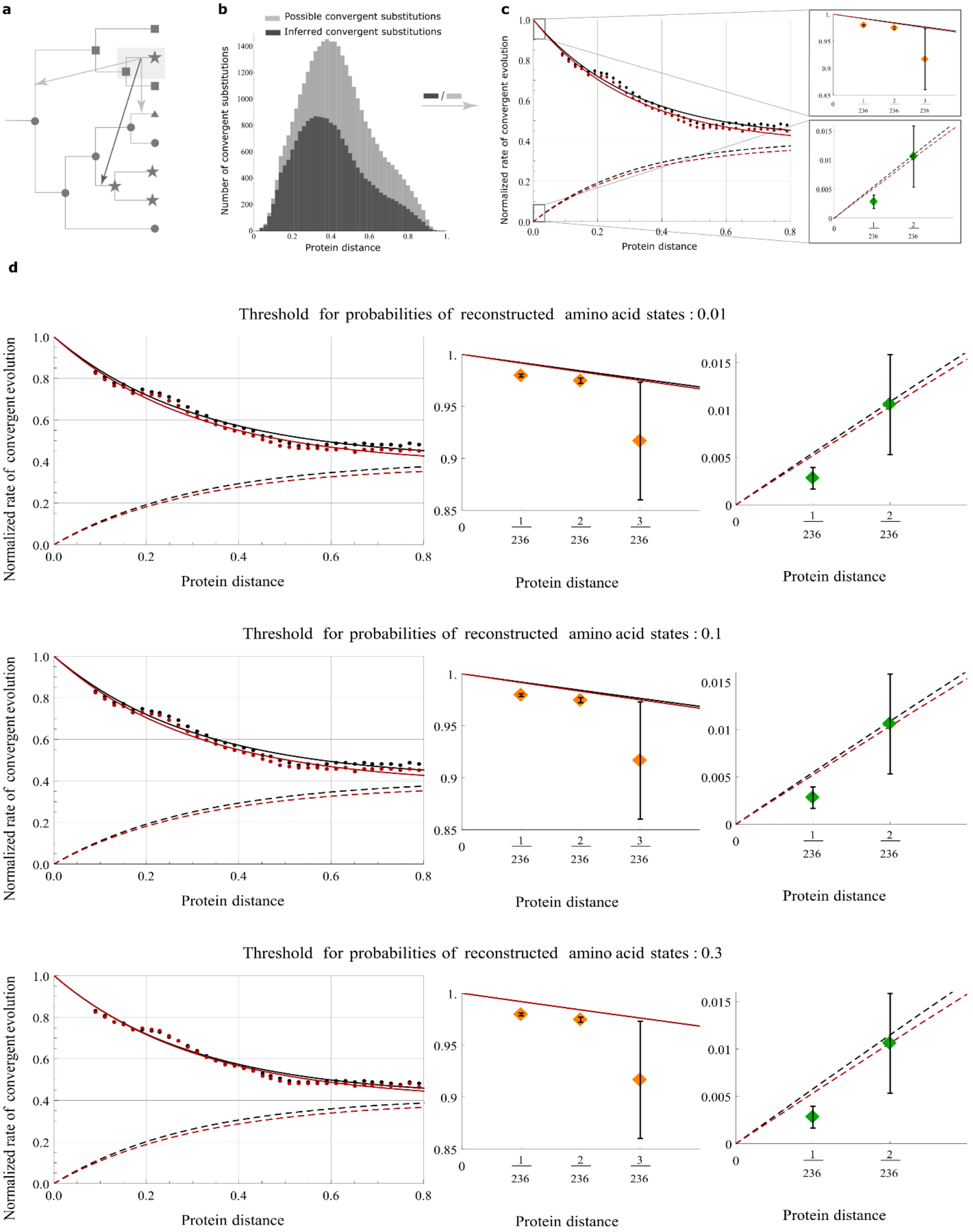
Prediction for mutants with 9 mutations

Experimental observation for mutants with 9 mutations

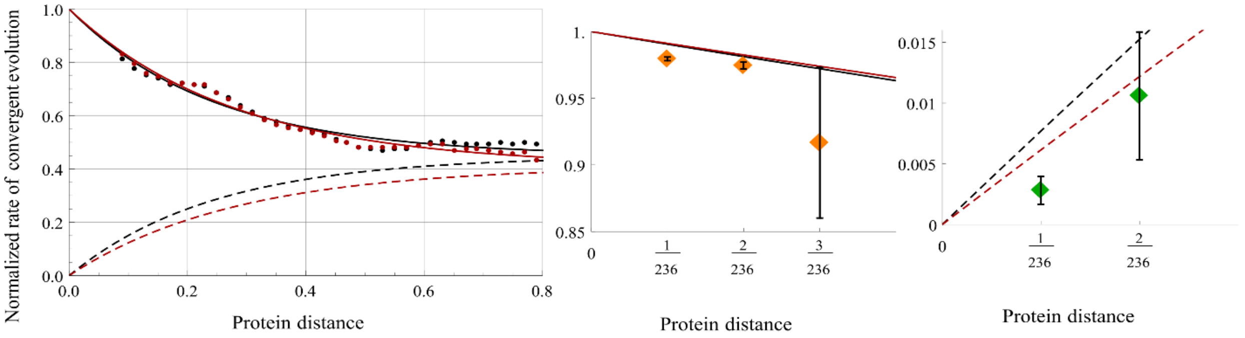
Prediction for mutants with 10 mutations

Experimental observation for mutants with 10 mutations

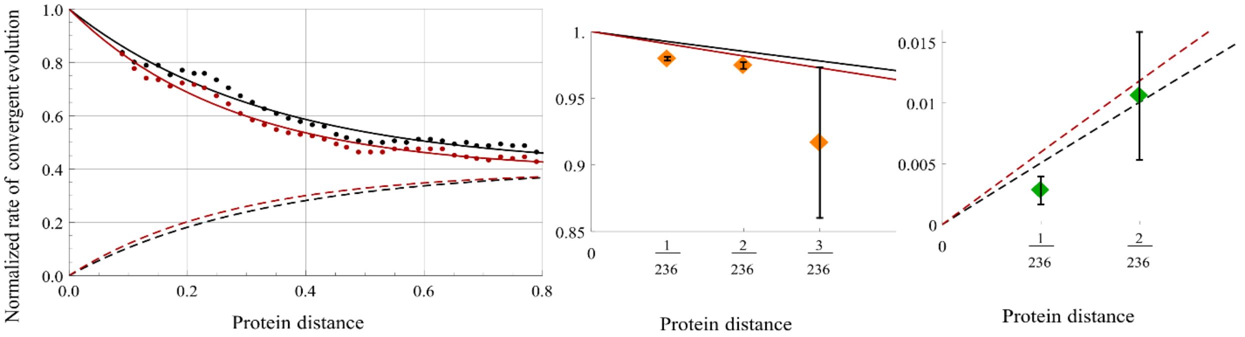
Supplementary Figure 4. The probability distribution of predicted false positive epistasis (red) and observed epistasis distribution (orange) for genotypes harboring different number of mutations away from the wildtype.


a
b
Possible convergent substitutions
Inferred convergent substitutions
Number of convergent substitutions
1400
1200
1000
800
600
400
200
0
0.2
0.4
0.6
0.8
1.
Protein distance
c
Normalized rate of convergent evolution
1.0
0.8
0.6
0.4
0.2
0.0
Protein distance
0.95
0.9
0.85
0.015
0.01
0.005
1/236
2/236
3/236
d
Threshold for probabilities of reconstructed amino acid states : 0.01
Threshold for probabilities of reconstructed amino acid states : 0.1
Threshold for probabilities of reconstructed amino acid states : 0.3
Normalized rate of convergent evolution
Protein distance

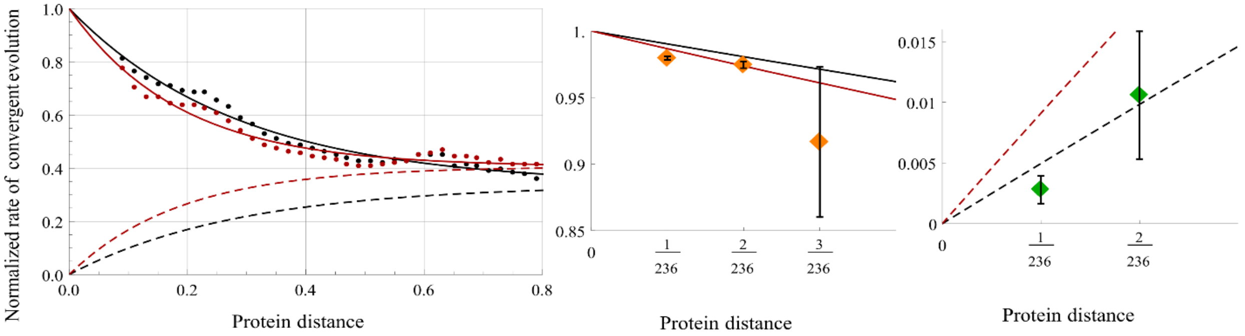
Threshold for probabilities of reconstructed amino acid states : 0.5



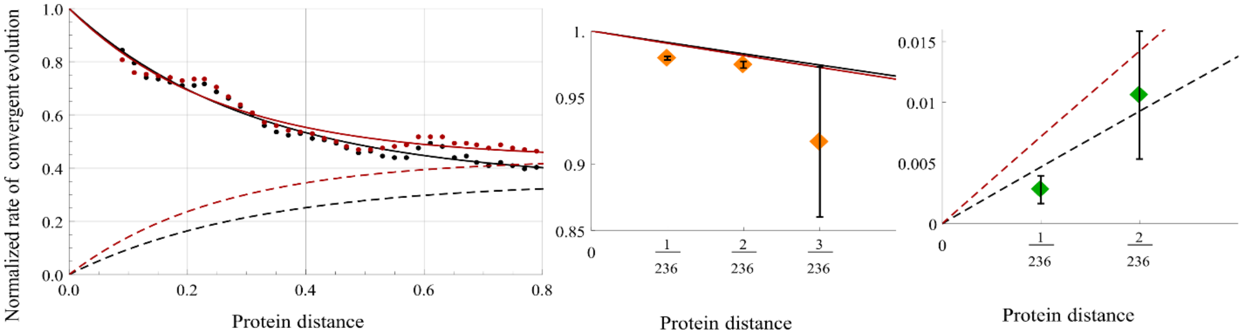
Threshold for probabilities of reconstructed amino acid states : 0.7



Threshold for probabilities of reconstructed amino acid states : 0.9



Threshold for probabilities of reconstructed amino acid states : 0.95



Supplementary Figure 5. a, For a given substitution on the phylogenetic tree (from ancestral state $A_{\blacksquare}$ to derived state $A_{\star}$) we calculate the sequence divergence to the nearest branch where a non-ancestral amino acid state (A_x) occurs (gray arrows). **b**, In 2% bins of divergence we count all such distances on the tree, counting only those distances where the derived ($A_{\star}$) and A_x amino acid states are equal (black arrow in **a**), representing instances of convergent substitutions. **c** The normalized rate of convergent evolution to terminal and reconstructed ancestral amino acid states for each distance bin (grey dots). The expected (line) and observed (green dots) probability that a single mutation remains fluorescent as the sequence accumulates other substitutions. The expected (broken line) and observed (orange dots) probability that a non-fluorescent mutation becomes fluorescent with sequence divergence. Results based on Bayesian (black lines) and maximum likelihood (red lines) approaches are shown. Bars represent a binomial proportion confidence interval (confidence level 68%). **d** Results of model predictions considering only ancestral amino acid states passing a probability threshold.

Supplementary Table 1. Confirmation of non-normalized fluorescence levels and epistasis for several genotypes.

Genotype	Green log fluorescence	Epistasis confirmed
wild type	4,3	
GFP deletion	2,3	
K166*	4,0	
K79*:Q80*	3,0	
K79*	3,4	
Q80*	3,4	
L44P	2,4	
C48R	3,8	
F165S	3,7	
V163A	4,4	
L141P	3,7	
F84S	3,8	
M88V	4,1	
K85T	3,8	
F83L	3,9	
F84S:M88V	2,3	Negative confirmed
F83L:K85T	2,3	Negative confirmed
L44P:F165S	2,4	Positive not confirmed
C48R:V163A	3,1	Positive not confirmed
L141P:F165S	2,3	Negative confirmed

Supplementary References

1. Wright, S. The roles of mutation, inbreeding, crossbreeding and selection in evolution. in *Proc. Sixth Int. Congr. Genet.* (ed. Jones, D. F.) **1,** 356–366 (Genetics Society of America, 1932).

2. Maynard Smith, J. Natural selection and the concept of a protein space. *Nature* **225,** 563–564 (1970).

3. De Visser, J. A. G. M. & Krug, J. Empirical fitness landscapes and the predictability of evolution. *Nat. Rev. Genet.* **15,** 480–90 (2014).

4. Kondrashov, D. A. & Kondrashov, F. A. Topological features of rugged fitness landscapes in sequence space. *Trends Genet.* **31,** 24–33 (2015).

5. Weinreich, D. M. & Knies, J. L. Fisher's geometric model of adaptation meets the functional synthesis: data on pairwise epistasis for fitness yields insights into the shape and size of phenotype space. *Evolution (N. Y).* **67,** 2957–2972 (2013).

6. Jacquier, H. *et al.* Capturing the mutational landscape of the beta-lactamase TEM-1. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 13067–13072 (2013).

7. Stiffler, M. A., Hekstra, D. R. & Ranganathan, R. Evolvability as a function of purifying selection in TEM-1 β-lactamase. *Cell* **160,** 882–892 (2015).

8. Firnberg, E., Labonte, J. W., Gray, J. J. & Ostermeier, M. A comprehensive, high-resolution map of a Gene's fitness landscape. *Mol. Biol. Evol.* **31,** 1581–1592 (2014).

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

10. De Visser, J. a. G. M., Cooper, T. F. & Elena, S. F. The causes of epistasis. *Proc. R. Soc. B Biol. Sci.* **278,** 3617–3624 (2011).

11. Olson, C. A., Wu, N. C. & Sun, R. A comprehensive biophysical description of pairwise epistasis throughout an entire protein domain. *Curr. Biol.* **24,** 2643–2651 (2014).

12. Bank, C., Hietpas, R. T., Jensen, J. D. & Bolon, D. N. A systematic survey of an intragenic epistatic landscape. *Mol. Biol. Evol.* **32,** 229–238 (2014).

13. Otwinowski, J. & Nemenman, I. Genotype to phenotype mapping and the fitness landscape of the E. coli lac promoter. *PLoS One* **8,** (2013).

14. Fowler, D. M. *et al.* High-resolution mapping of protein sequence-function relationships. *Nat. Methods* **7,** 741–746 (2010).

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

16. Reynolds, K. a., McLaughlin, R. N. & Ranganathan, R. Hot spots for allosteric regulation on protein surfaces. *Cell* **147,** 1564–1575 (2011).

17. Meini, M. R., Tomatis, P. E., Weinreich, D. M. & Vila, A. J. Quantitative description of a protein fitness landscape based on molecular features. *Mol. Biol. Evol.* **32**, 1774–1787 (2015).

18. Parera, M. & Martinez, M. A. Strong epistatic interactions within a single protein. *Mol. Biol. Evol.* **31,** 1546–1553 (2014).

19. Podgornaia, A. I. & Laub, M. T. Pervasive degeneracy and epistasis in a protein-protein interface. *Science (80-.).* **347,** 673–677 (2015).

20. Weinreich, D. M., Lan, Y., Wylie, C. S. & Heckendorn, R. B. Should evolutionary geneticists worry about higher-order epistasis? *Curr. Opin. Genet. Dev.* **23,** 700–707 (2013).

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

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

23. Poelwijk, F. J., Kiviet, D. J., Weinreich, D. M. & Tans, S. J. Empirical fitness landscapes reveal accessible evolutionary paths. *Nature* **445,** 383–386 (2007).

24. 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).

25. Povolotskaya, I. S. & Kondrashov, F. A. Sequence space and the ongoing expansion of the protein universe. *Nature* **465,** 922–926 (2010).

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

27. Hinkley, T. *et al.* A systems analysis of mutational effects in HIV-1 protease and reverse transcriptase. *Nat. Genet.* **43,** 487–489 (2011).

28. Cormack, B. P., Valdivia, R. H. & Falkow, S. FACS-optimized mutants of the green fluorescent protein (GFP). *Gene* **173,** 33–38 (1996).

29. Matsumoto, A. & Itoh, T. Q. Self-assembly cloning: a rapid construction method for recombinant molecules from multiple fragments. *Biotechniques* **51,** 55–66 (2011).

30. Shcherbo, D. *et al.* Far-red fluorescent tags for protein imaging in living tissues. *Biochem. J.* **418,** 567–574 (2009).

31. Arai, R., Ueda, H., Kitayama, A., Kamiya, N. & Nagamune, T. Design of the linkers which effectively separate domains of a bifunctional fusion protein. *Protein Eng.* **14,** 529–532 (2001).

32. Waldo, G. S., Standish, B. M., Berendzen, J. & Terwilliger, T. C. Rapid protein-folding assay using green fluorescent protein. *Nat. Biotechnol.* **17,** 691–695 (1999).

33. Osterman, I. A., Evfratov, S. A., Sergiev, P. V & Dontsova, O. A. Comparison of mRNA features affecting translation initiation and reinitiation. *Nucleic Acids Res.* **41,** 474–486 (2013).

34. Vartanian, J. P., Henry, M. & Wain-Hobson, S. Hypermutagenic PCR involving all four transitions and a sizeable proportion of transversions. *Nucleic Acids Res.* **24,** 2627–2631 (1996).

35. Bolotin, D. *et al.* MiXCR: a comprehensive software for adaptive immunity profiling. *Nat. Methods* **12**, 380–381(2015).

36. Poole, E. S., Brown, C. M. & Tate, W. P. The identity of the base following the stop codon determines the efficiency of in vivo translational termination in Escherichia coli. *EMBO J.* **14,** 151–158 (1995).

37. Jia, Y. *et al.* Caffe: convolutional architecture for fast feature embedding. *arXiv* **1408.5093,** (2014).

38. Lozovsky, E. R. *et al.* Stepwise acquisition of pyrimethamine resistance in the malaria parasite. *Proc. Natl. Acad. Sci. U. S. A.* **106,** 12025–12030 (2009).

39. Weinreich, D. M., Watson, R. A. & Chao, L. Perspective: Sign epistasis and genetic constraint on evolutionary trajectories. *Evolution* **59,** 1165–1174 (2005).

40. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32,** 1792–1797 (2004).

41. Armougom, F. *et al.* Expresso: automatic incorporation of structural information in multiple sequence alignments using 3D-Coffee. *Nucleic Acids Res.* **34,** W604–W608 (2006).

42. Wallace, I. M., O'Sullivan, O., Higgins, D. G. & Notredame, C. M-Coffee: combining multiple sequence alignment methods with T-Coffee. *Nucleic Acids Res.* **34,** 1692–1699 (2006).

43. Ronquist, F. & Huelsenbeck, J. P. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19,** 1572–1574 (2003).

44. Usmanova, D. R., Ferretti, L., Povolotskaya, I. S., Vlasov, P. K. & Kondrashov, F. A. A model of substitution trajectories in sequence space and long-term protein evolution. *Mol. Biol. Evol.* **32,** 542–554 (2015).

45. Naumenko, S. A., Kondrashov, A. S. & Bazykin, G. A. Fitness conferred by replaced amino acids declines with time. *Biol. Lett.* **8,** 825–828 (2012).

46. Rogozin, I. B., Thomson, K., Csürös, M., Carmel, L. & Koonin, E. V. Homoplasy in genome-wide analysis of rare amino acid replacements: the molecular-evolutionary basis for Vavilov's law of homologous series. *Biol. Direct* **3,** 7 (2008).

47. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24,** 1586–1591 (2007).

48. Silverman, B. W. *Density Estimation for Statistics and Data Analysis*. (Chapman & Hall/CRC, 1998).

49. Shinobu, A., Palm, G. J., Schierbeek, A. J. & Agmon, N. Visualizing proton antenna in a high-resolution green fluorescent protein structure. *J. Am. Chem. Soc.* **132,** 11093–11102 (2010).

50. Chudakov, D. M., Matz, M. V, Lukyanov, S. & Lukyanov, K. A. Fluorescent proteins and their applications in imaging living cells and tissues. *Physiol. Rev.* **90,** 1103–1163 (2010).

51. Renfrew, P. D., Choi, E. J., Bonneau, R. & Kuhlman, B. Incorporation of noncanonical amino acids into Rosetta and use in computational protein-peptide interface design. *PLoS One* **7,** e32637 (2012).

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

53. Kellogg, E. H., Leaver-Fay, A. & Baker, D. Role of conformational sampling in computing mutation-induced changes in protein structure and stability. *Proteins* **79,** 830–838 (2011).

54. Arpino, J. A. J., Reddington, S. C., Halliwell, L. M., Rizkallah, P. J. & Jones, D. D. Random single amino acid deletion sampling unveils structural tolerance and the benefits of helical registry shift on GFP folding and structure. *Structure* **22,** 889–898 (2014).

55. DePristo, M. A., Weinreich, D. M. & Hartl, D. L. Missense meanderings in sequence space: a biophysical view of protein evolution. *Nat. Rev. Genet.* **6,** 678–687 (2005).

56. Kondrashov, F. A. & Kondrashov, A. S. Multidimensional epistasis and the disadvantage of sex. *Proc. Natl. Acad. Sci. U. S. A.* **98,** 12089–12092 (2001).