LETTER

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]

Fitness landscapes[1,2] depict how genotypes manifest at the phenotypic level and form the basis of our understanding of many areas of biology[2–7], yet their properties remain elusive. Previous studies have analysed specific genes, often using their function as a proxy for fitness[2,4], experimentally assessing the effect on function of single mutations and their combinations in a specific sequence[2,5,8–15] or in different sequences[2,3,5,16–18]. However, systematic high-throughput studies of the local fitness landscape of an entire protein have not yet been reported. Here we visualize an extensive region of the local fitness landscape of the green fluorescent protein from *Aequorea victoria* (avGFP) by measuring the native function (fluorescence) of tens of thousands of derivative genotypes of avGFP. We show that the fitness landscape of avGFP is narrow, with 3/4 of the derivatives with a single mutation showing reduced fluorescence and half of the derivatives with four mutations being completely non-fluorescent. The narrowness is enhanced by epistasis, which was detected in up to 30% of genotypes with multiple mutations and mostly occurred through the cumulative effect of slightly deleterious mutations causing a threshold-like decrease in protein stability and a concomitant loss of fluorescence. A model of orthologous sequence divergence spanning hundreds of millions of years predicted the extent of epistasis in our data, indicating congruence between the fitness landscape properties at the local and global scales. The characterization of the local fitness landscape of avGFP has important implications for several fields including molecular evolution, population genetics and protein design.

We assayed the local fitness landscape of avGFP by estimating the fluorescence levels of genotypes obtained by random mutagenesis of the avGFP sequence (Fig. 1). We used fluorescence-activated cell sorting (Supplementary Fig. 1) and sequenced the entire GFP coding region to assay the fluorescence of many thousands of genotypes created by random mutagenesis of the wild-type sequence (Supplementary Information 2 and Extended Data Fig. 1). We applied several strategies to minimize the error of our estimate of fluorescence (Supplementary Information 3.4 and 4.4), which was estimated from thousands of independent measurements of the wild-type sequence (false negative error rate = 0.08%) and genotypes incorporating mutations known to eliminate fluorescence (false positive error rate = 0.24%). Our final data set included 56,086 unique nucleotide sequences coding for 51,715 different protein sequences. Our procedure introduced an average of 3.7 mutations per gene sequence, and most assayed genotypes contained several, up to 15, missense mutations. Still, since the total number of possible sequences grows exponentially with the number of mutations, the fraction of sampled sequences was tiny for sequences containing more than two mutations (Extended Data Table 1). We used these data to survey the local fitness landscape of GFP, analysing the effect of single, double and multiple mutations.

The distribution of fitness effects of individual missense mutations was assayed by comparing the distribution of fluorescence of wild-type avGFP amino acid sequences, tagged by different molecular barcodes, and the distribution of fluorescence of sequences carrying a single mutation (Supplementary Information 4.1). We found that at least 75% of mutations had a deleterious effect on fluorescence, including 9.4% of single mutations conferring a more than fivefold decrease in fluorescence, but for many mutations the effect was small (Fig. 2a). Accordingly, genotypes with multiple missense mutations were more likely to have low fluorescence, and most genotypes carrying five or more missense mutations were non-fluorescent (Extended Data Fig. 2). Mutations with a strong effect on fluorescence preferably resided at sites that coded for amino acid residues oriented internally towards the chromophore (Fig. 2b, c), which is consistent with data on other proteins on the preference of deleterious mutations to target buried residues[9,11–13]. The effect of mutations on fluorescence was positively correlated with site conservation (Extended Data Fig. 3a, Spearman's rank correlation coefficient = 0.40, $P = 1.44 \times 10^{-10}$), and mutations with a deleterious impact were less likely to be found in orthologous sequences (Extended Data Fig. 3b) and more likely to be found in sites with a deleterious deletion (Supplementary Fig. 2). Still, ~10% of mutant states conferring a non-fluorescent phenotype were nevertheless fixed in long-term evolution (Extended Data Fig. 3b), and a substantial fraction of genotypes containing only mutations leading to amino acid states from GFP orthologues was non-fluorescent (Supplementary Fig. 3), indicating that epistasis affects the avGFP fitness landscape[16].

Interaction of deleterious mutations can manifest in positive epistasis, when the joint effect of mutations is weaker than their independent contribution, or negative epistasis when the joint effect is stronger (Fig. 3a). Light intensity is perceived in the logarithmic scale by living beings, including jellyfish[19]. Thus, we defined epistasis e as the deviation from additivity of effects of single mutations on the logarithmic

[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.
*These authors contributed equally to this work.

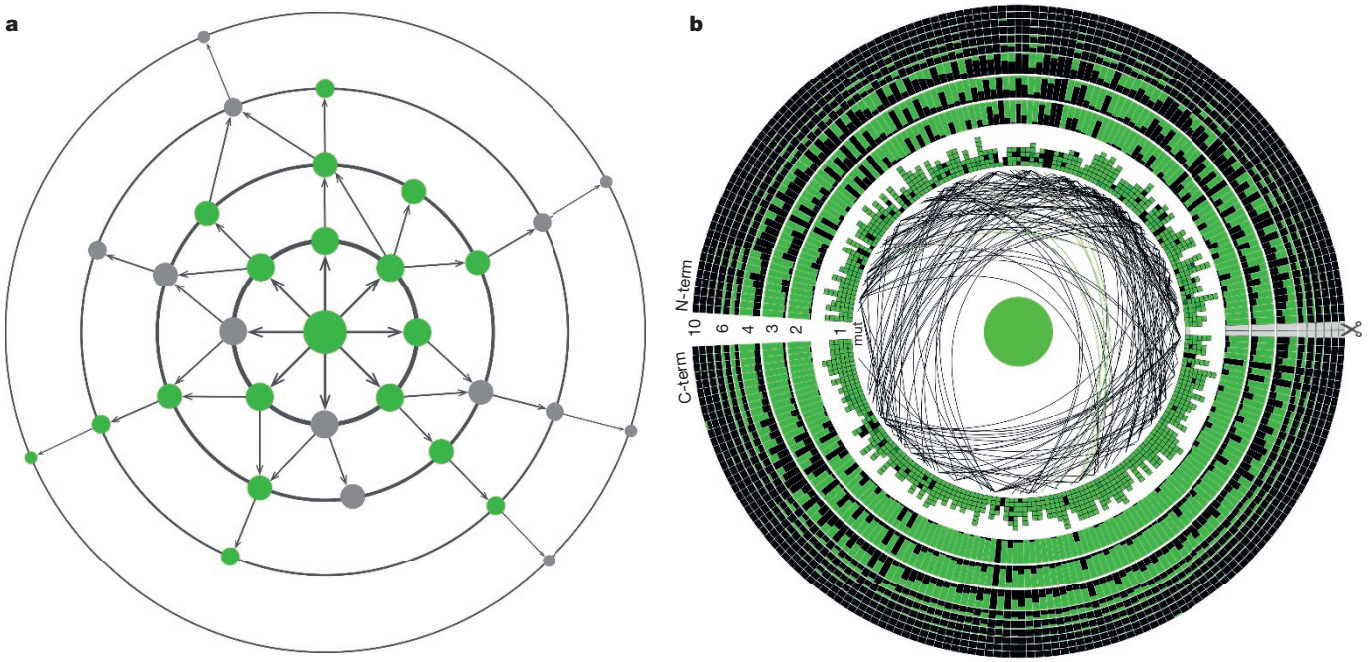


Figure 1 | Exploring the local fitness landscape. a, Wild-type avGFP (centre) and most single mutants (innermost circle) fluoresce green. Genotypes with multiple mutations may exhibit negative epistasis, with combinations of neutral mutations creating non-fluorescent phenotypes (grey), or positive epistasis, in which a mutation in a non-fluorescent genotype restores fluorescence. **b**, The GFP sequence arranged in a circle, each column representing one amino acid site. In the first circle, the colour intensity of the squares indicates the brightness of a single mutation at the corresponding site relative to the wild type, shown in the centre. Sites with positive and negative epistatic interactions between pairs of mutations are connected by green and black lines, respectively. In circles further away from the centre, representing genotypes with multiple mutations, the fraction of the column coloured green (black) represents the fraction of genotypes corresponding to high (low) fluorescence among all assayed genotypes with a mutation at that site. Scissors indicate the restriction site.

scale. We compared the decrement of the log-fluorescence of a multiple mutant F_{mult} to the sum of decrements of individual mutants, such that $e = (F_{\mathrm{mult}} - F_{\mathrm{WT}}) - \sum_i (F_i - F_{\mathrm{WT}})$, where F_{WT} and F_i are the log-fluorescence values conferred by wild-type avGFP and avGFP with the i-th single missense mutation, respectively. We restricted the expected fluorescence of the multiple mutant $\sum_i (F_i - F_{\mathrm{WT}})$ to the observed maximum or minimum levels (Supplementary Information 4.2). This eliminated spurious detection of epistasis, which could otherwise occur, for example, when a non-fluorescent double mutant consists of two mutations, both of which individually confer a non-fluorescent genotype. We defined strong epistasis as $|e| > 0.7$, or as cases in which the observed fluorescence differed from the expected by at least five-fold, with a false discovery rate of $<1\%$ (Supplementary Fig. 4).

Negative epistasis affected up to 30% of all genotypes, depending on the number of mutations (Fig. 3b, c), which resulted in a larger than expected fraction of non-fluorescent genotypes (Fig. 3c). Genotypes

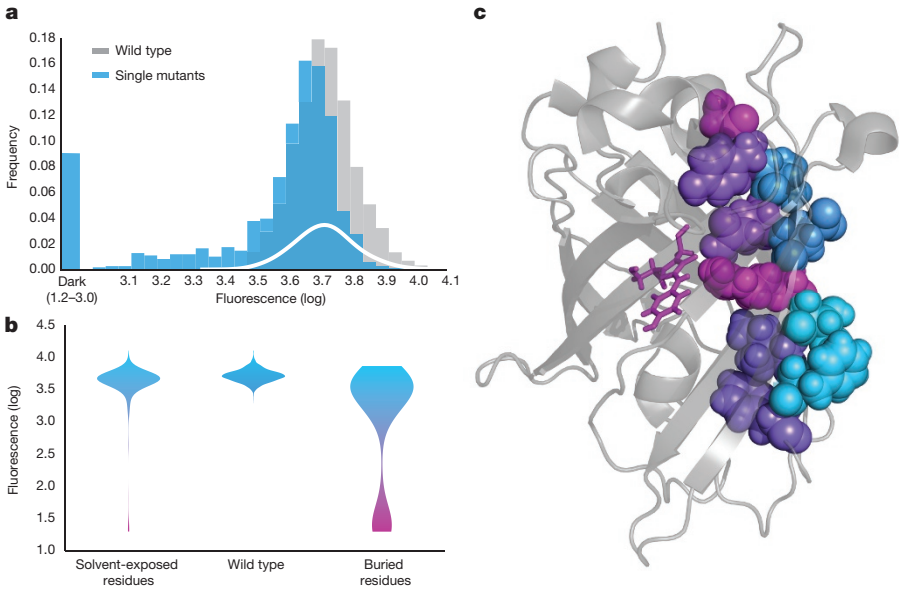


Figure 2 | The effect of single mutations on avGFP. a, The distributions of independently measured fluorescence for 2,442 wild-type sequences (grey), 1,114 single mutants (blue), and the estimated fraction of neutral mutations (white). **b**, **c**, Single missense mutations strongly decreasing fluorescence (violet) tended to occur at sites with internally oriented residues (**b**), shown on a selected β-strand of the GFP structure (**c**).

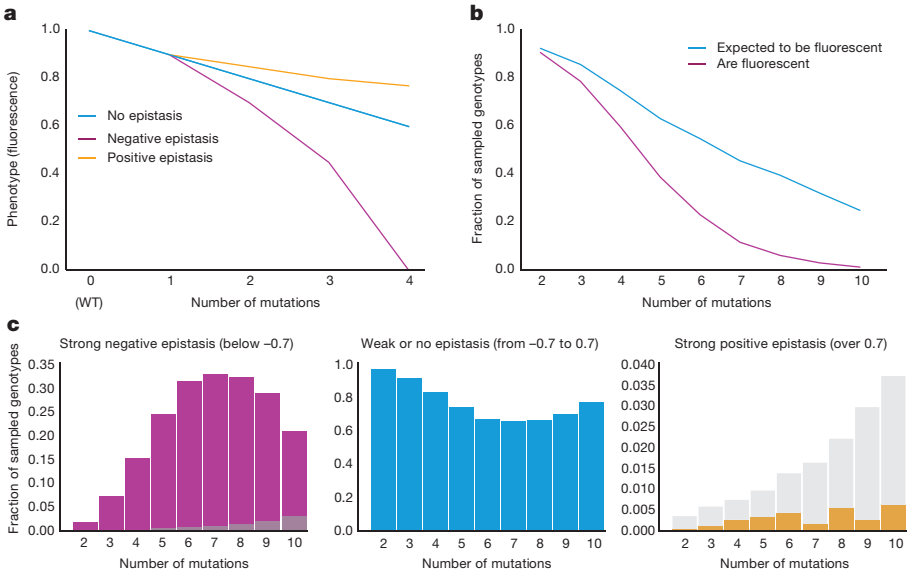


Figure 3 | Prevalence of epistasis in the local fitness landscape of avGFP. **a**, A hypothetical representation of negative and positive epistasis as a function of the number of single mutations from avGFP. WT, wild type. **b**, The fraction of observed non-fluorescent genotypes (red) and the expected fraction of non-fluorescent genotypes calculated as the sum of the log-fluorescence effects of individual mutations (blue). **c**, The distributions of epistasis for negative and positive epistasis of different strength, with the expected false discovery rate shown in grey.

carrying more than seven mutations showed a decrease in the prevalence of negative epistasis because many genotypes carrying multiple mutations were expected to lose fluorescence even without epistasis (Fig. 3b). Positive epistasis was rare in avGFP, on the order of accuracy of our method. We sampled ~2% of all possible pairs of mutations (Extended Data Table 1), assaying 30% of pairs of amino acid sites (16,898 out of 55,696, Extended Data Fig. 4a). Epistatic pairs of sites were located across the avGFP sequence (Extended Data Fig. 4a), mostly beyond the range of direct physical interaction of amino acid residues (Extended Data Fig. 4b) but marginally closer together than random (Extended Data Fig. 4c, $P < 0.0004$, Mann–Whitney U-test). Epistasis was found among 96% of mutations with weak effect (Extended Data Fig. 4d), suggesting that their joint effect brings the protein over some stability margin[8,20]. Finally, epistasis was more common between pairs of sites in which both residues are internally oriented (Extended Data Fig. 4e). Taken together, these data indicate that epistasis was more common at functionally important sites.

In a unidimensional landscape, fitness is a monotonic function of an intermediate variable known as fitness potential[21,22], which is the sum of effects of individual mutations. We used multiple regression considering a non-epistatic fitness function in which log-fluorescence, F, is equal to the linear predictor, the fitness potential, p, such that $F = f(p) = p$. This simplest, non-epistatic model explained only 70% of the initial sample variance ($\sigma^2 = 1.12$ and $\sigma^2 = 0.34$ before and after the application of the model, respectively). Using the variance of the 2,442 wild-type fluorescence measurements, we estimated that ~1% of the initial sample variance can be attributed to noise ($\sigma^2 = 0.0097$), indicating that the remaining 29% of sample variance cannot be explained without epistasis.

The simplest form of an epistatic fitness function is when fitness is a monotonic nonlinear function of p[21,22]. The lack of genotypes with intermediate fluorescence (Extended Data Fig. 5a) suggests that the avGFP fitneses landscape can be described by a truncation-like fitness function[23]. We therefore modelled F as a sigmoid function of p, which explained 85% of the initial sample variance ($\sigma^2 = 0.17$). A more complex sigmoid-shaped fitness function refined with a neural network approach (Supplementary Information 4.6) explained 93.5% of the initial sample variance ($\sigma^2 = 0.065$, Extended Data Fig. 5), confirming that the fitness landscape can mostly be represented by a unidimensional threshold function (Fig. 4), which can arise from the joint contribution of mutations to protein stability[8,13,14,20,24]. The average fluorescence of single mutants of avGFP as a function of the predicted protein destabilization, $\Delta\Delta G$, reveals a threshold around 7–9 kcal mol^{-1} (Fig. 4). Notably, the hidden value found by the artificial neural network for single mutants correlated to the predicted $\Delta\Delta G$ (Fig. 4 and Extended Data Fig. 5f), confirming a probable influence of protein stability on the nature of epistasis in avGFP. The threshold fitness function does a remarkably good job in approximating the entire fitness landscape, explaining ~95% of all variance. However, when taking into account the error rate of our data set, we estimate that at least 0.3% of genotypes cannot be explained by the threshold fitness function (Supplementary Information 4.5 and Extended Data Fig. 5d), representing instances of multidimensional epistasis[2,5,7].

We compared the local avGFP fitness peak to the global GFP fitness landscape using sequences of GFP orthologues. Negative, threshold-like epistasis, leading to truncation selection[23] against slightly

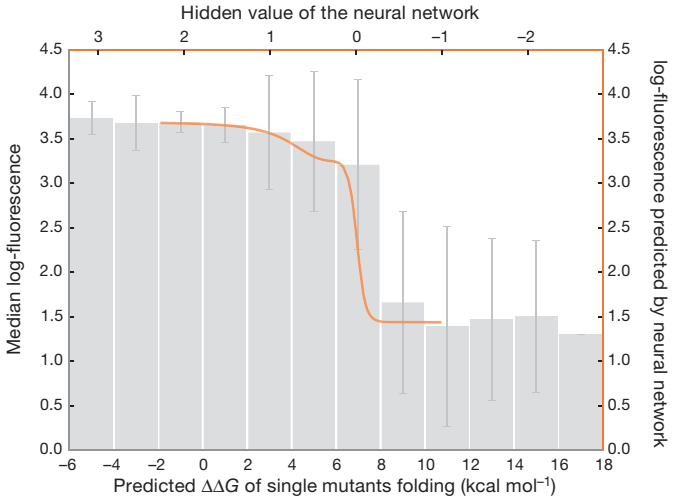


Figure 4 | Modelling genotype to phenotype relationship. Median fluorescence of GFP with single mutations as a function of their effect on predicted folding energy ($\Delta\Delta G$), overlaid with the independently obtained sigmoid-like fitness function predicted by the neural network (orange line). Error bars denote s.d.

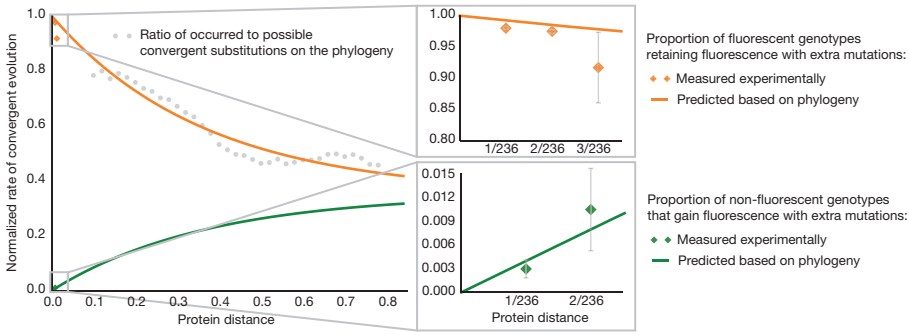


Figure 5 | The fitness matrix model of GFP long-term evolution. The normalized rate of convergent evolution to terminal and reconstructed ancestral amino acid states for each distance bin (grey dots). The expected (orange line) and observed in experimental data (orange dots) probability that a single mutation remains fluorescent as the sequence accumulates other substitutions. The expected (green line) and observed (green dots) probability that a non-fluorescent mutation becomes fluorescent with sequence divergence. Bars represent a binomial proportion confidence interval (confidence level 68%).

deleterious mutations may prevent their accumulation in evolution[25,26]. Thus, we compared the fraction of neutral single mutations to the rate of nonsynonymous and synonymous evolution (d*N*/d*S*), a proxy for the average strength of selection. The average d*N*/d*S* across a broad phylogenetic range was 0.35 ± 0.1 (mean $\pm$ s.d.), and 0.17 when avGFP was compared to the orthologue from the closest fluorescent relative *Aequorea macrodactyla*. These measurements are similar to the estimated proportion of neutral mutations in avGFP (0.23), suggesting that the proportion of neutral mutations is similar across distant fitness peaks. The rate at which the phenotypic effect of amino acid substitutions changes across evolution, which is reflected in the changing rate of convergent evolution across the phylogenetic tree[27,28], can be used to model the prevalence of epistasis with the fitness matrix model[28]. This model approximates the prevalence of epistasis as the proportion of amino acid mutations that markedly change their effect on fitness after the occurrence of substitutions at other sites (Supplementary Information 5). Applying the fitness matrix model to the GFP multiple alignment, we predicted the proportion of mutations that change their effects on fluorescence when found in a different genetic background, revealing prevalence of positive and negative epistasis, which concur with our experimental observations (Fig. 5 and Supplementary Fig. 5). The congruence of the data from an evolutionary trajectory spanning hundreds of millions of years with experimental data are remarkable, suggesting similarity in the local and global structures of the fitness landscape shaped by strong epistatic interactions.

Our study provides complementary results of the analysis of single and double mutations to several previous studies[9,11–14], and a novel depiction of a large segment of the fitness landscape of a single protein. The proportion of neutral single mutations in our data was similar to that observed when fitness was assayed directly or through competition experiments[10,24,29] but substantially lower than that observed in functional studies[4,10,11,13,17]. Furthermore, the propensity of multiple mutations to have a stronger negative effect on fitness than the sum of individual mutation effects has been observed[9,10,12,14]. However, because our analysis considered genotypes carrying multiple mutations, we infer a wider picture of the local fitness landscape (Supplementary Video 1). The avGFP fitness peak is narrow and defined by negative epistasis, best described by truncation selection in which fluorescence is eliminated if the joint effect of mutations exceeds a threshold of an intermediate property, possibly protein stability[8,20]. Such a landscape increases the efficacy of selection against slightly deleterious mutations[23], preventing their accumulation in evolution[25,26]. Simultaneously, the fitness landscape is approximately non-epistatic near the fitness peak. If other proteins have a similar fitness landscape it would support the nearly neutral theory of evolution[30], explaining the selective forces and evolutionary dynamics of mutations with negligible individual effects on fitness.

The broad congruence of our data with the prevalence of epistasis from long-term evolution suggests that the shape of the local fitness landscape can be extrapolated to a larger scale. Yet, epistasis between sites coding for residues with a direct interaction in protein structure was rare, contrasting with observation of such instances in long-term evolution[16] and a mutation assay of the RNA recognition motif (RRM) domain[12]. Thus, the local fitness landscape spanning a few mutations from a single fitness peak may be approximated by a unidimensional threshold fitness potential function; however, this simple fitness function may not be appropriate to describe fitness landscapes that incorporate fitness ridges connecting sequences of more divergent orthologues[27]. The nature of global fitness landscapes, especially the interaction between local and global scales, remains to be explored.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 28 April 2015; accepted 7 April 2016.
Published online 11 May 2016.

1. Wright, S. The roles of mutation, inbreeding, crossbreeding and selection in evolution. *Proc. Sixth Int. Congr. Genet.* **1,** 356–366 (1932).
2. de Visser, J. A. G. M. & Krug, J. Empirical fitness landscapes and the predictability of evolution. *Nature Rev. Genet.* **15,** 480–490 (2014).
3. Dean, A. M. & Thornton, J. W. Mechanistic approaches to the study of evolution: the functional synthesis. *Nature Rev. Genet.* **8,** 675–688 (2007).
4. Soskine, M. & Tawfik, D. S. Mutational effects and the evolution of new protein functions. *Nature Rev. Genet.* **11,** 572–582 (2010).
5. 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).
6. Mackay, T. F. C. Epistasis and quantitative traits: using model organisms to study gene-gene interactions. *Nature Rev. Genet.* **15,** 22–33 (2014).
7. Taylor, M. B. & Ehrenreich, I. M. Higher-order genetic interactions and their contribution to complex traits. *Trends Genet.* **31,** 34–40 (2015).
8. 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).
9. Fowler, D. M. *et al.* High-resolution mapping of protein sequence-function relationships. *Nature Methods* **7,** 741–746 (2010).
10. 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).
11. Jacquier, H. *et al.* Capturing the mutational landscape of the beta-lactamase TEM-1. *Proc. Natl Acad. Sci. USA* **110,** 13067–13072 (2013).
12. 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).
13. 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).
14. 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 (2015).
15. 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).

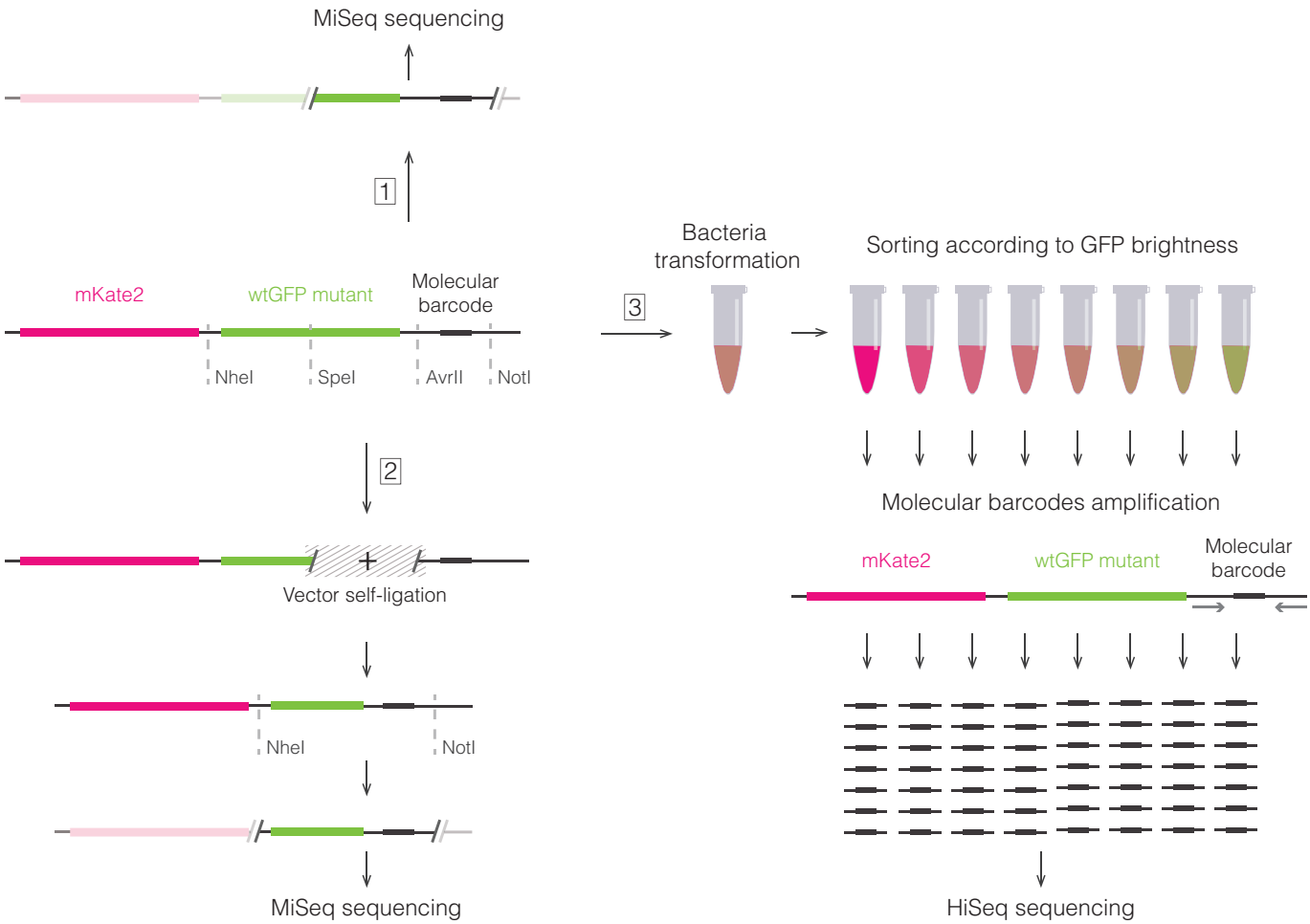
16. Kondrashov, A. S., Sunyaev, S. & Kondrashov, F. A. Dobzhansky–Muller incompatibilities in protein evolution. *Proc. Natl Acad. Sci. USA* **99,** 14878–14883 (2002).
17. Firnberg, E., Labonte, J. W., Gray, J. J. & Ostermeier, M. A comprehensive, high-resolution map of a gene's fitness landscape. *Mol. Biol. Evol.* **31,** 1581–1592 (2014).
18. Parera, M. & Martinez, M. A. Strong epistatic interactions within a single protein. *Mol. Biol. Evol.* **31,** 1546–1553 (2014).
19. Coates, M. M., Garm, A., Theobald, J. C., Thompson, S. H. & Nilsson, D. E. The spectral sensitivity of the lens eyes of a box jellyfish, *Tripedalia cystophora* (Conant). *J. Exp. Biol.* **209,** 3758–3765 (2006).
20. DePristo, M. A., Weinreich, D. M. & Hartl, D. L. Missense meanderings in sequence space: a biophysical view of protein evolution. *Nature Rev. Genet.* **6,** 678–687 (2005).
21. Milkman, R. Selection differentials and selection coefficients. *Genetics* **88,** 391–403 (1978).
22. Kimura, M. & Crow, J. F. Effect of overall phenotypic selection on genetic change at individual loci. *Proc. Natl Acad. Sci. USA* **75,** 6168–6171 (1978).
23. Crow, J. F. & Kimura, M. Efficiency of truncation selection. *Proc. Natl Acad. Sci. USA* **76,** 396–399 (1979).
24. Rockah-Shmuel, L., Tóth-Petróczy, Á. & Tawfik, D. S. Systematic mapping of protein mutational space by prolonged drift reveals the deleterious effects of seemingly neutral mutations. *PLOS Comput. Biol.* **11,** e1004421 (2015).
25. Li, W. H. Models of nearly neutral mutations with particular implications for nonrandom usage of synonymous codons. *J. Mol. Evol.* **24,** 337–345 (1987).
26. Akashi, H. Inferring weak selection from patterns of polymorphism and divergence at 'silent' sites in *Drosophila* DNA. *Genetics* **139,** 1067–1076 (1995).
27. Povolotskaya, I. S. & Kondrashov, F. A. Sequence space and the ongoing expansion of the protein universe. *Nature* **465,** 922–926 (2010).
28. 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).
29. Eyre-Walker, A. & Keightley, P. D. The distribution of fitness effects of new mutations. *Nature Rev. Genet.* **8,** 610–618 (2007).
30. Ohta, T. Slightly deleterious mutant substitutions in evolution. *Nature* **246,** 96–98 (1973).

Supplementary Information is available in the online version of the paper.

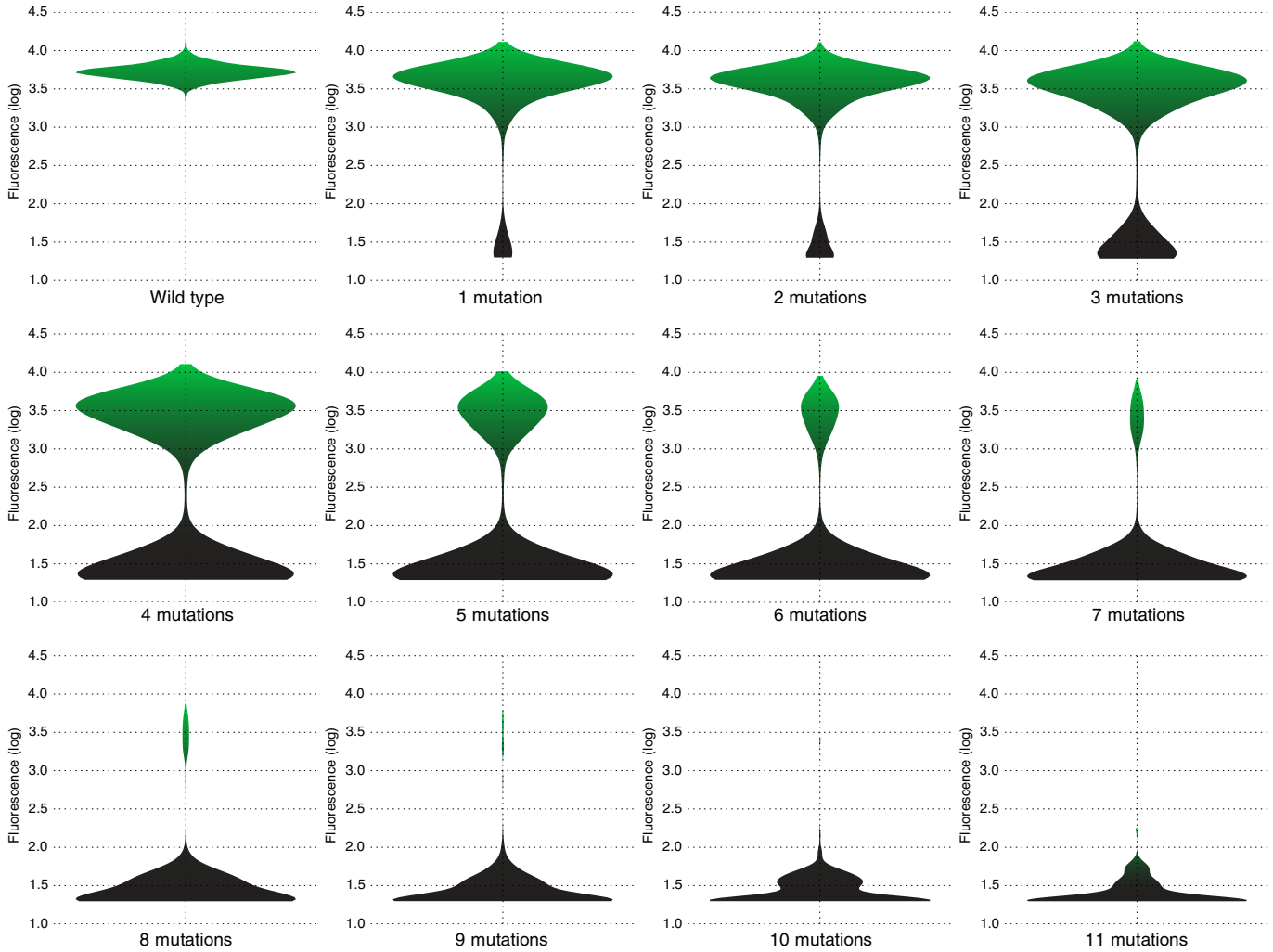
Acknowledgements We thank Y. Kulikova and G. Filion for discussion on statistical analysis and I. Osterman, R. Moretti and J. Meiler for technical assistance and M. Friesen for a critical reading of the manuscript. We thank H. Himmelbauer, CRG Genomic Unit and the Russian Science Foundation project (14-50-00150) for sequencing. Experiments were partially carried out using the equipment provided by the IBCH core facility (CKP IBCH). The work was supported by HHMI International Early Career Scientist Program (55007424), the EMBO Young Investigator Programme, MINECO (BFU2012-31329), Spanish Ministry of Economy and Competitiveness Centro de Excelencia Severo 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), Russian Science Foundation (14-25-00129) and the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013, ERC grant agreement, 335980_EinME).

Author Contributions K.S.S. and M.V.M. conceived the idea for the experiment; K.S.S., D.A.B., M.V.M., A.S.M., G.V.S., M.D.L., D.M.C., E.V.P., I.Z.M., D.S.T., K.A.L. and F.A.K. participated in experimental design; K.S.S., D.A.B., M.V.M., G.V.S., E.V.P., E.S.E. and M.D.L. performed the experiments; K.S.S., D.A.B., M.V.M., D.R.U., A.S.M., D.N.I., N.G.B., M.S.B., O.S., N.S.B., P.K.V., A.S.K. and F.A.K. performed data analysis; K.S.S., D.A.B., M.V.M., D.R.U., D.N.I. and F.A.K. wrote the paper.

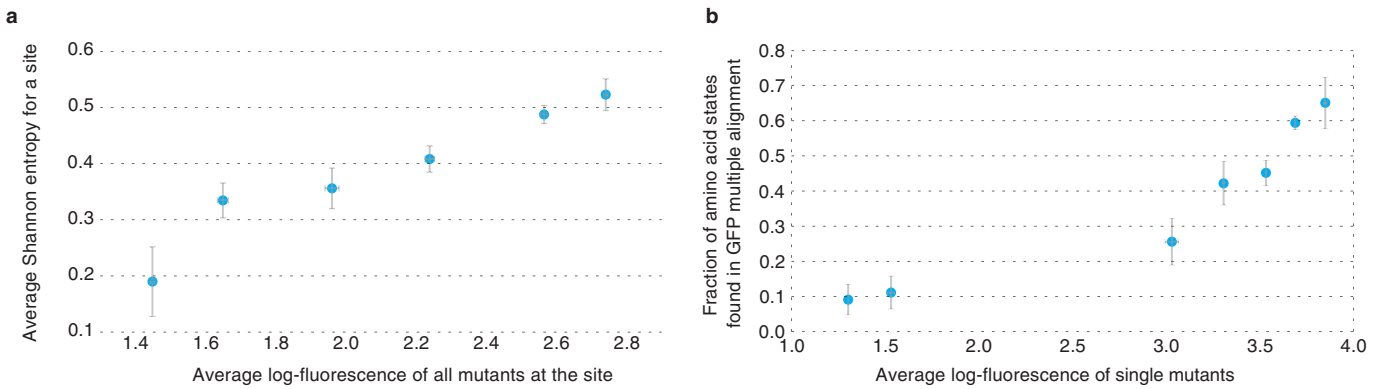
Author Information Raw sequencing data were deposited in the Sequence Read Archive (SRA) under BioProject number PRJNA282342. Processed data sets are available at Figshare http://dx.doi.org/10.6084/m9.figshare.3102154. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to F.A.K. (fyodor.kondrashov@crg.es).



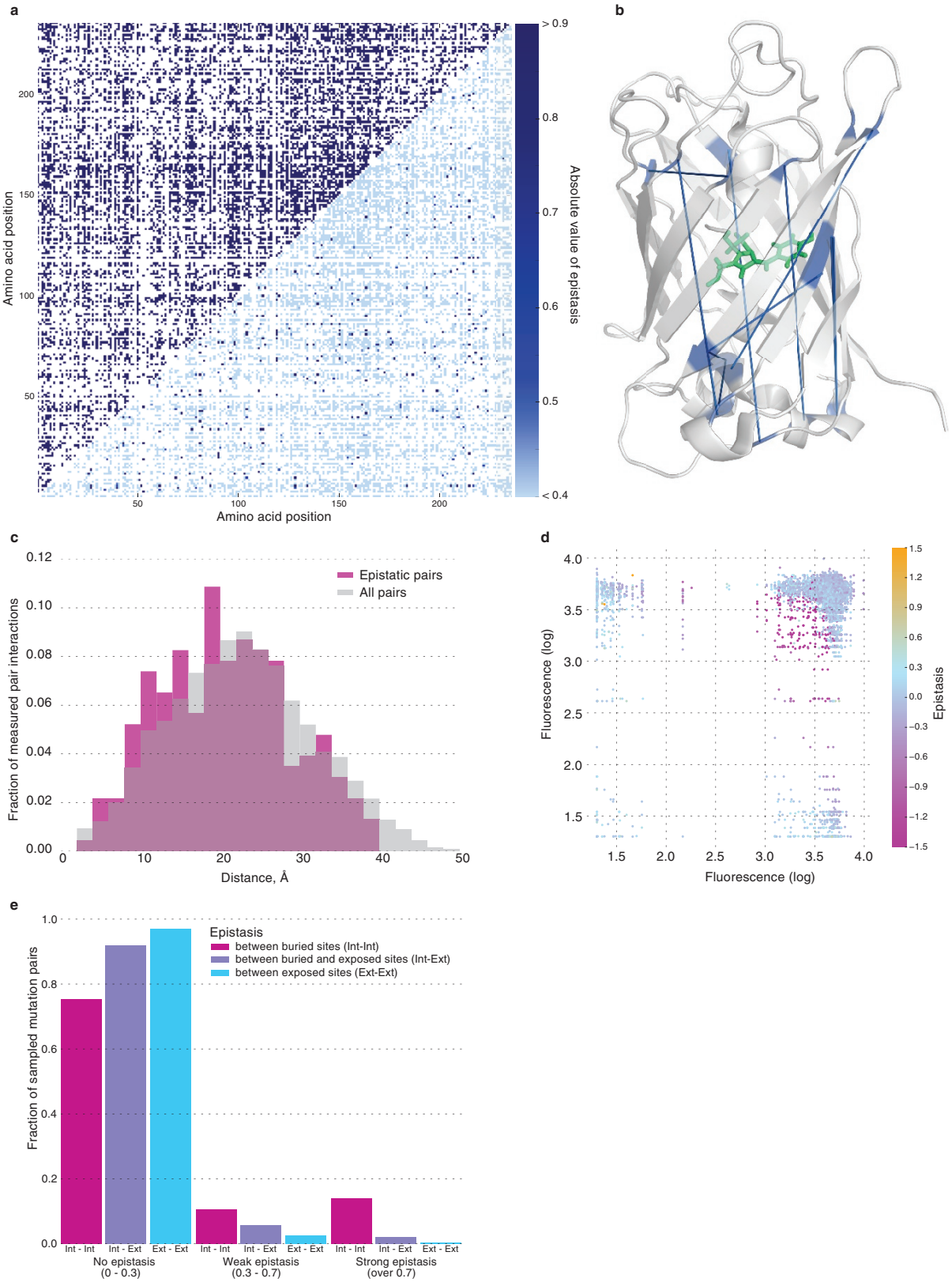
Extended Data Figure 1 | Scheme of the experimental approach. The depiction of the construct design, expression and cell sorting.



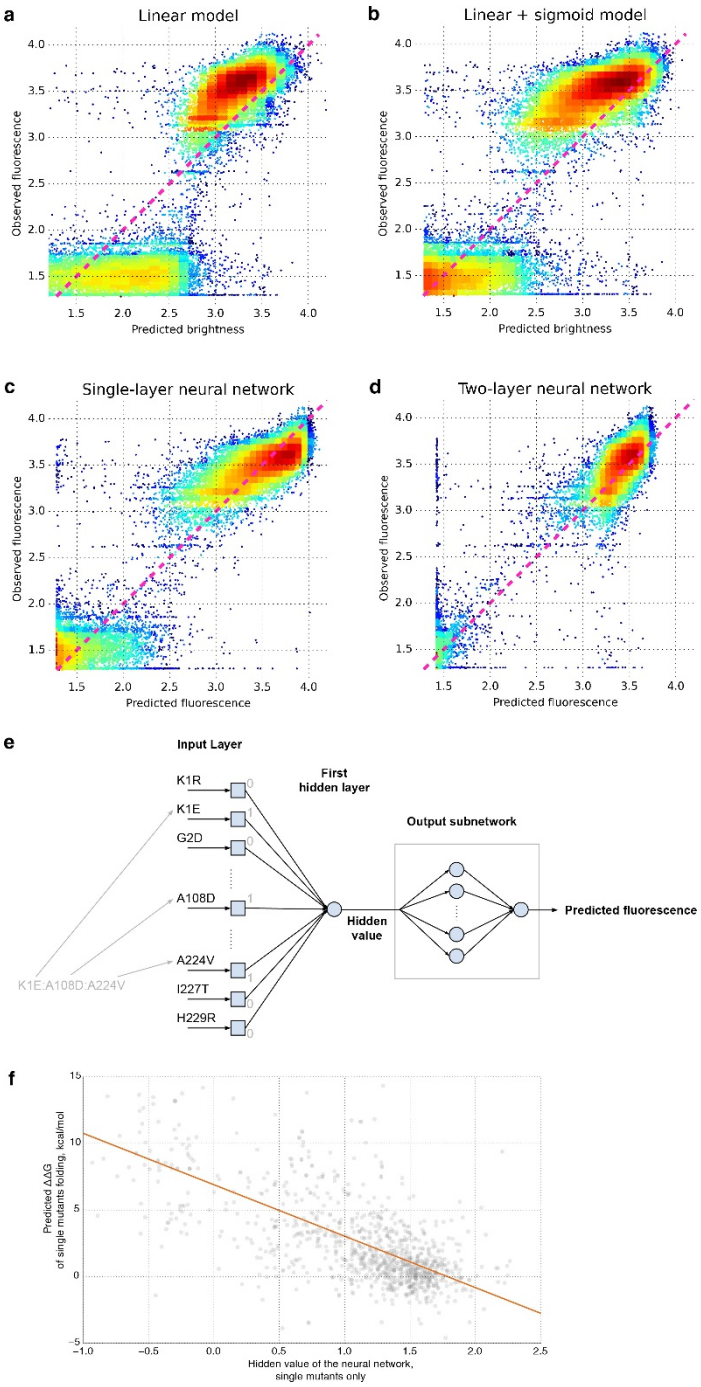
Extended Data Figure 2 | Fluorescence and impact of mutations. A violin plot of the measured levels of fluorescence for genotypes carrying different numbers of missense mutations.



Extended Data Figure 3 | Mutant genotypes and evolution. a, **b**, The log-fluorescence and evolutionary conservation expressed as Shannon entropy (**a**), and fraction of mutant amino acid states found in avGFP orthologues (**b**). The *y*-axis error bars in **b** show the binomial proportion confidence interval level (68%), and other error bars denote s.e.m.



Extended Data Figure 4 | Epistatically interacting pairs of sites in the GFP structure. a, Pairs of amino acid sites for which we assayed at least one combination of mutations (in blue, top left). The distribution of the maximum level of epistasis observed between sites (blue scale, bottom right) and unknown values (white). **b**, Pairs of sites under exceptionally strong epistatic interaction ($e < -2$) connected by a blue line on the GFP structure. **c**, The distribution of distances in the GFP structure between sites with at least one pair of epistatically interacting mutations (red) and all pairs of sites in the structure (grey). **d**, Epistasis between pairs of mutations as a function of their individual fluorescence. **e**, The contribution of internally and externally oriented amino acid residues in the avGFP structure relative to pairs of missense mutations showing no epistasis ($|e| < 0.3$), weak ($0.3 < |e| < 0.7$) and strong ($|e| > 0.7$) epistasis.



Extended Data Figure 5 | Modelling effect of mutations on fluorescence. **a**, A multiple linear regression in which fluorescence is linear combination of effects of individual single mutations. **b**, A multiple regression in which mutations contribute linearly to a fitness potential and fluorescence is a sigmoidal function of p where $F \approx e^{-p}/(1+e^{-p})$. **c**, **d**, The predicted fluorescence by a neural network approach. Predicted fitness function by a neural network with one hidden neuron and two neurons in the outer layer. **e**, The scheme of our neural network approach. The genotype data was passed to the input layer of the neural network as an array of 0s or 1s corresponding to the absence or presence of amino acid mutations in the genotype, respectively. The first hidden layer consisted of a single neuron that calculated the weighted sum of inputs using weights obtained during training. The output of the first hidden layer was passed through an output subnetwork that transformed this value with a nonlinear function to make the final prediction of fluorescence. The output subnetwork consisted of several neurons with a sigmoidal transfer function, allowing the subnetwork to approximate a broad range of nonlinear functions. The final mapping of the hidden value to fluorescence was determined by the weights of connections between neurons inside the output subnetwork. During training all weights were optimized to find the best prediction of fluorescence from the hidden value. The resulting function that was defined during training is shown in Fig. 4. **f**, Correlation between the hidden value of the neural network and Rosetta-predicted $\Delta\Delta G$ for single mutants.

Extended Data Table 1 | Genotypes with measured fluorescence in our data set

Number of mutations from the wild type	Number of amino acid sequences assayed	Number of possible amino acid sequences*	Fraction of amino acid sequences sampled	Number of nucleotide sequences assayed	Number of possible nucleotide sequences	Fraction of nucleotide sequences sampled
1	1,114	1,233	0.90	1,336	2,133	0.626
2	13,010	759,528	0.02	11,555	2,271,645	0.005
3	12,683	311,659,656	4.1×10^{-5}	12,654	1.61×10^{9}	7.86×10^{-6}
4	9,759	9.6×10^{10}	1.0×10^{-7}	10,633	8.55×10^{11}	1.24×10^{-8}
5	7,215	2.4×10^{13}	3.1×10^{-10}	8,164	3.63×10^{14}	2.25×10^{-11}
6	4,643	4.8×10^{15}	9.6×10^{-13}	5,869	1.28×10^{17}	4.58×10^{-14}
7	2,783	8.5×10^{17}	3.3×10^{-15}	3,664	3.87×10^{19}	9.47×10^{-17}
8	1,526	1.3×10^{20}	1.2×10^{-17}	2,212	1.02×10^{22}	2.17×10^{-19}
9	714	1.8×10^{22}	4.1×10^{-20}	1,229	2.39×10^{24}	5.13×10^{-22}
10	352	2.2×10^{24}	1.6×10^{-22}	624	5.04×10^{26}	1.24×10^{-24}

*We considered possible amino acid sequences as sequences in which every amino acid sequence is the result of single nucleotide substitutions from the wild-type avGFP sequence.