Title

popDMS infers mutation effects from deep mutational scanning data.

Journal

Computer applications in the biosciences : CABIOS, 40(8)

Authors

Hong, Zhenchen

Shimagaki, Kai

Barton, John

Publication Date

2024-08-02

DOI

10.1093/bioinformatics/btae499

Peer reviewed

Bioinformatics, 2024, **40(8)**, btae499
https://doi.org/10.1093/bioinformatics/btae499
Advance Access Publication Date: 8 August 2024
Applications Note

OXFORD

Sequence analysis

popDMS infers mutation effects from deep mutational scanning data

Zhenchen Hong[1,†], **Kai S. Shimagaki**[2,†], **John P. Barton**[1,2,3,*]

[1]Department of Physics and Astronomy, University of California, Riverside, CA 92521, United States
[2]Department of Computational and Systems Biology, University of Pittsburgh School of Medicine, PA 15260, United States
[3]Department of Physics and Astronomy, University of Pittsburgh, PA 15260, United States

*Corresponding author. Department of Computational and Systems Biology, University of Pittsburgh, 800 Murdoch Bldg, 3420 Forbes Avenue, Pittsburgh, PA 15260, United States. E-mail: jpbarton@pitt.edu (J.P.B)

†Equal contribution.

Associate Editor: Pier Luigi Martelli

Abstract

Summary: Deep mutational scanning (DMS) experiments provide a powerful method to measure the functional effects of genetic mutations at massive scales. However, the data generated from these experiments can be difficult to analyze, with significant variation between experimental replicates. To overcome this challenge, we developed popDMS, a computational method based on population genetics theory, to infer the functional effects of mutations from DMS data. Through extensive tests, we found that the functional effects of single mutations and epistasis inferred by popDMS are highly consistent across replicates, comparing favorably with existing methods. Our approach is flexible and can be widely applied to DMS data that includes multiple time points, multiple replicates, and different experimental conditions.

Availability and implementation: popDMS is implemented in Python and Julia, and is freely available on GitHub at https://github.com/bartonlab/popDMS.

1 Introduction

Understanding the relationship between protein sequence and phenotype is a central question in evolution and protein engineering. In recent years, a new family of experimental methods, commonly referred to as deep mutational scanning (DMS) or multiplexed assays for variant effects (MAVEs), have been developed to measure the functional effects of large numbers of mutations simultaneously (Fowler *et al.* 2010, Gasperini *et al.* 2016). DMS experiments typically work by generating a vast library of protein variants that are then passed through rounds of selection that favor functional variants while eliminating deleterious ones (Fowler and Fields 2014). One can then compare variant frequencies in the pre- and post-selection libraries to estimate the functional effects of mutations. This approach has been successfully applied in a wide variety of contexts, from studying the function of enzymes (Romero *et al.* 2015) and tRNAs (Li *et al.* 2016) to measuring the mutational tolerance of influenza (Thyagarajan and Bloom 2014, Doud *et al.* 2018, Lee *et al.* 2018) and human immunodeficiency virus (HIV-1) (Haddox *et al.* 2016, Dingens *et al.* 2017, Haddox *et al.* 2018) surface proteins.

Despite the success of DMS experiments, popular approaches for analyzing DMS data yield modest correlations between the inferred functional effects of mutations in experimental replicates. Thus, a significant amount of variance in the data remains unexplained. Some methods use the ratios between post- and pre-selection variant frequencies, known as enrichment ratios, to estimate mutation effects (Fowler *et al.* 2011, Hietpas *et al.* 2011, Bloom 2015). Ratio-based methods may be sensitive to noise when variant counts are low, a common occurrence in DMS experiments. Methods based on regression (Araya *et al.* 2012, Starita *et al.* 2015, Matuszewski *et al.* 2016, Rich *et al.* 2016, Rubin *et al.* 2017) provide improved performance, but substantial uncertainty in the inferred effects of different mutations persists.

2 Results

We developed a method, popDMS, to estimate the functional effects of mutations in DMS experiments using statistical methods from population genetics (Supplementary Information). In our approach, we view rounds of phenotypic selection in experiments as analogous to rounds of reproduction in natural populations. We quantify the effect of each mutation i by a selection coefficient s_i, which describes the relative advantage or disadvantage of the mutation for surviving selection in the experiment. For simplicity, we assume that the total fitness of a sequence with multiple mutations is the sum of the corresponding selection coefficients. We then use the Wright-Fisher (WF) model, an evolutionary model from population genetics, to quantify the likelihood of the experimentally observed variant frequencies over time as a function of the selection coefficients, $\mathcal{L}((\boldsymbol{z}(t_k))_{k=0}^{K}|\boldsymbol{s})$ (see Supplementary Information for details). The $\boldsymbol{z}(t_k)$ represent vectors of variant frequencies $\boldsymbol{z}$ at different times t_k. The WF

Received: 19 February 2024; **Revised:** 10 July 2024; **Editorial Decision:** 31 July 2024; **Accepted:** 6 August 2024

model defines the relationship between "fitness" and frequency change, and allows us to model competition between variants. We then use sequence data to estimate the effects of mutations on fitness in experiments.

To regularize our estimates, we introduce a Gaussian prior distribution $P_{\text{prior}}(s)$ for the selection coefficients. Leveraging recently developed computational methods (Sohail *et al.* 2021, 2022, Lee *et al.* 2022), we can identify the selection coefficients that represent the best compromise between fitting the data and minimizing the prior distribution,

$$\hat{s} = \underset{s}{\text{argmax}}\ \mathcal{L}\left(s|(z(t_k))_{k=0}^{K}\right) P_{\text{prior}}(s). \quad (1)$$

Typically, we adjust the width of the prior distribution based on the data, but a fixed value can also be specified (Supplementary Information). The Gaussian prior is equivalent to an L_2-norm penalty on the selection coefficients, or ridge regression.

popDMS has several computational strengths. First, the use of regularization for the selection coefficients curbs the inference of strong functional effects in the absence of strong statistical evidence. Our likelihood framework further allows us to derive joint estimates of selection coefficients across replicates that are guided by levels of evidence in the data, rather than simply averaging the inferred functional effects of mutations across replicates. When information about sequencing error rates is available, we can perform error correction for variant frequencies.

In simulations, we found that popDMS was robust to sampling noise and provided stronger correlations between inferred variant effects across replicates than common methods based on enrichment ratios or regression (Supplementary Fig. S1). The variant effects inferred by popDMS were also more similar to true, underlying ones than alternative approaches, even with the addition of negative binomial sampling noise (Supplementary Fig. S2, see Supplementary Information).

Next, we analyzed a collection of 28 DMS datasets with popDMS (Araya *et al.* 2012, Starita *et al.* 2013, Findlay *et al.* 2014, Doud *et al.* 2015, Starita *et al.* 2015, Li *et al.* 2016, Ashenberg *et al.* 2017, Dingens *et al.* 2018, Haddox *et al.* 2018, Hom *et al.* 2019, Soh *et al.* 2019, Bridgford *et al.* 2020, Roop *et al.* 2020, Starr *et al.* 2020, Lei *et al.* 2023). These datasets were generated and analyzed using a variety of experimental techniques and analytical methods (see Supplementary Table S1). Like the functional metrics introduced by previous methods, selection coefficients provide an intuitive visualization of the functional effects of mutations

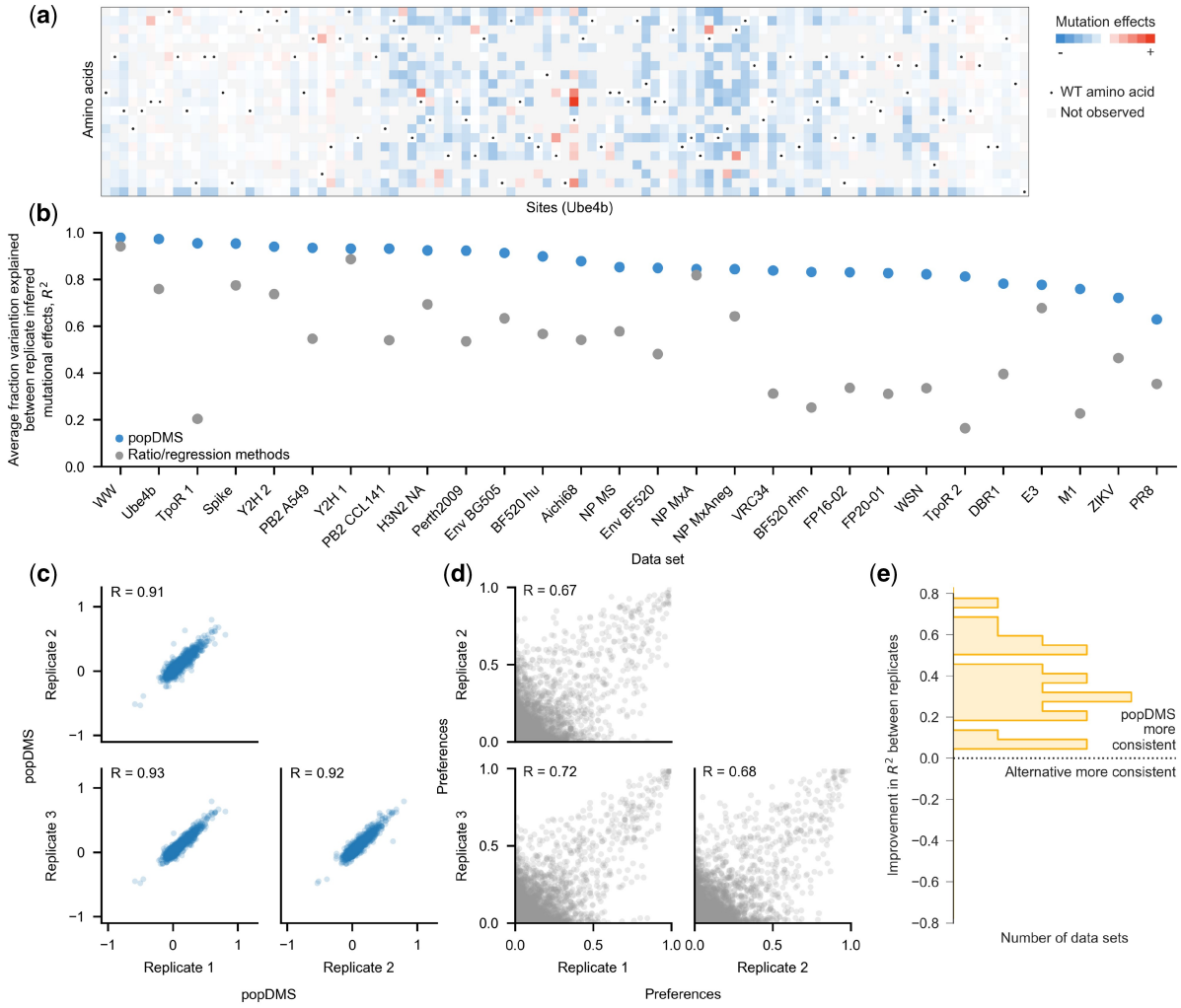

Figure 1. popDMS overview. (a) Example of the effects of mutations inferred by popDMS for the Ube4b protein (Starita *et al.* 2013). (b) Across 28 datasets, popDMS infers more consistent mutational effects than previous ratio/regression-based methods. To illustrate consistency between replicates, we show (c) selection coefficients inferred across replicates for the HIV-1 envelop BF520 dataset (Haddox *et al.* 2018), compared with (d) enrichment ratios for the same data. (e) popDMS gains in consistency across replicates are often substantial, improving R^2 by an average of 0.34

(Fig. 1a). To quantify the consistency of different analytical methods, we computed the Pearson correlation R between mutation effects inferred from replicates of the same experiment. We found that mutation effects inferred by popDMS had higher correlations between replicates than those inferred by prior methods for all the datasets that we considered (Fig. 1b). The rank correlations between replicates were also typically higher for popDMS than for other approaches, showing that the consistency of the inferred mutational effects is not simply due to rescaling (Supplementary Fig. S3). Furthermore, our selection coefficients compared favorably with the frequencies of amino acid variants in influenza viruses in a natural population (Thyagarajan and Bloom 2014) (see Supplementary Information).

To illustrate performance in a typical case, we show selection coefficients inferred for mutations in the HIV-1 envelope protein BF520 (Fig. 1c) compared with enrichment ratios (Fig. 1d) for the same data (Haddox *et al.* 2018). Improvements in consistency across replicates with popDMS were often substantial. The mean improvement in R^2 for variant effects was 0.35, with 6 out of 28 datasets showing an improvement in R^2 of >0.50 (Fig. 1e).

In addition to the modified form of our estimator for variant effects, regularization also contributes to the improved correlation between replicates by shrinking effects with little support in the data toward zero (see Supplementary Fig. S4). As we discuss below, we also treat wild-type (WT) amino acids differently than most ratio- or regression-based approaches. Because WT residues are typically among the fittest at each site, changes to these terms can have particularly large effects on consistency between replicates.

We then asked how similar the selection coefficients inferred by popDMS are to mutation effects inferred by previous methods. Across the experimental datasets that we tested, popDMS results were broadly consistent with existing metrics (average Pearson's $R = 0.74$). This correlation is similar to the average correlation between replicates of the same dataset using current ratio- or regression-based methods (average Pearson's $R = 0.70$). Figure 2a shows a typical example, comparing selection coefficients inferred by popDMS with enrichment ratios for the HIV-1 Env BG505 dataset (Dingens *et al.* 2018).

While the inferred mutation effects agreed for most sites, some showed qualitative differences (Supplementary Fig. S5). One factor underlying this result is that popDMS models variants with high initial frequencies, such as WT or reference amino acids, in the same way as other, low-frequency variants (see Supplementary Information). In alternative methods, the statistical treatment for WT amino acids is often different than for other variants.

Beyond inferring the effects of individual mutations, we can apply popDMS to estimate pairwise epistatic interactions between variants at different sites. We inferred epistatic interactions in an hYAP65 WW domain dataset using popDMS, which we also compared with previous results (Araya *et al.* 2012). Due to different conventions in defining epistasis, we transformed the functional measurements defined in ref. (Araya *et al.* 2012) to more directly compare with our results (Supplementary Information). To more clearly identify strongly interacting pairs of sites, we computed the sum of squared epistatic interactions between all pairs of amino acids at each pair of sites in the WW domain, using both popDMS and the previous regression-based approach. Our results showed good agreement with the pairs of sites that were previously inferred to have the strongest epistatic interactions (Fig. 2b). However, epistatic interactions inferred by popDMS were substantially sparser than those that had been inferred before (Fig. 2c). Given the enormous number of possible epistatic interactions between amino acid variants at different sites, sparsity is an attractive statistical feature that can facilitate focus on a smaller number of biologically important interactions.

3 Discussion

In summary, popDMS is an efficient, reliable approach for inferring mutation effects from DMS data, which is grounded in evolutionary theory. Across simulations and a wide array

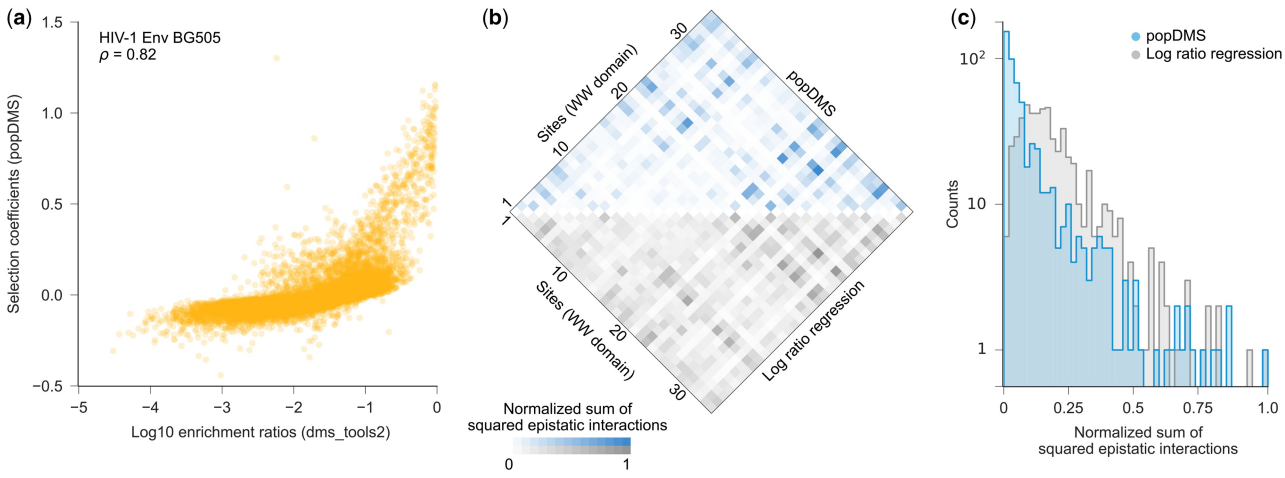

Figure 2. Mutation effects inferred by popDMS are broadly consistent with alternative methods. (a) For the HIV-1 Env BG505 dataset, selection coefficients inferred by popDMS are congruent with enrichment ratios computed using dms_tools2 (Spearman's $\rho = 0.84$). At some sites, significant differences are observed (see Supplementary Fig. S5). (b) In the hYAP65 WW domain dataset, similar sites are inferred to have strong epistatic interactions using popDMS and log ratio regression (Araya *et al.* 2012). Interactions inferred in ref. (Araya *et al.* 2012) have been transformed to compare more directly with interactions inferred by popDMS, and both sets of interactions are normalized to scale between zero and one (Supplementary Information). (c) Epistatic interactions inferred by popDMS are substantially sparser than those inferred with the regression-based approach (Araya *et al.* 2012)

of datasets, we found that popDMS infers more consistent mutation effects than the popular alternatives used here. Our approach allows us to combine statistical power across multiple replicates, and it is also capable of inferring epistatic interactions given appropriate data. popDMS is written in Python3 and C++, and uses codon counts in dms_tools format (Bloom 2015) or sequence counts in MaveDB format (Esposito *et al.* 2019) as input, with code and example visualizations freely available on GitHub (https://github.com/bartonlab/popDMS, Supplementary Information).

Here, we have focused on the correlations of inferred mutational effects between experimental replicates to quantify the consistency of different inference methods. By this statistical measure, popDMS is more consistent on average than current ratio- and regression-based methods, including both correlations between values (Pearson correlations) and the ranks of mutational effects (Spearman correlations). We also found that selection coefficients inferred by popDMS more closely matched with underlying fitness parameters in simulations. However, greater biological relevance could only be established through experiments. Future studies that experimentally test the predictions of different inference methods would be of great interest.

Author contributions

All authors contributed to methods development, data analysis, interpretation of results, and writing the paper. J.P.B. supervised the project.

Supplementary data

Supplementary data are available at *Bioinformatics* online.

Conflict of interest

None declared.

Funding

This work was supported by the National Institute of General Medical Sciences of the National Institutes of Health [Award Number R35GM138233] to J.P.B.

References

Araya CL, Fowler DM, Chen W *et al.* A fundamental protein property, thermodynamic stability, revealed solely from large-scale measurements of protein function. *Proc Natl Acad Sci USA* 2012;**109**:16858–63. https://doi.org/10.1073/pnas.1209751109.

Ashenberg O, Padmakumar J, Doud MB *et al.* Deep mutational scanning identifies sites in influenza nucleoprotein that affect viral inhibition by mxa. *PLoS Pathog* 2017;**13**:e1006288. https://doi.org/10.1371/journal.ppat.1006288.

Bloom JD. Software for the analysis and visualization of deep mutational scanning data. *BMC Bioinformatics* 2015;**16**:168. https://doi.org/10.1186/s12859-015-0590-4.

Bridgford JL, Lee SM, Lee CMM *et al.* Novel drivers and modifiers of MPL-dependent oncogenic transformation identified by deep mutational scanning. *Blood* 2020;**135**:287–92. https://doi.org/10.1182/blood.2019002561.

Dingens AS, Haddox HK, Overbaugh J *et al.* Comprehensive mapping of HIV-1 escape from a broadly neutralizing antibody. *Cell Host Microbe* 2017;**21**:777–87.e4.

Dingens AS, Acharya P, Haddox HK *et al.* Complete functional mapping of infection- and vaccine-elicited antibodies against the fusion peptide of HIV. *PLoS Pathog* 2018;**14**:e1007159. https://doi.org/10.1371/journal.ppat.1007159.

Doud MB, Ashenberg O, Bloom JD. Site-specific amino acid preferences are mostly conserved in two closely related protein homologs. *Mol Biol Evol* 2015;**32**:2944–60. https://doi.org/10.1093/molbev/msv167.

Doud MB, Lee JM, Bloom JD. How single mutations affect viral escape from broad and narrow antibodies to h1 influenza hemagglutinin. *Nat Commun* 2018;**9**:1386.

Esposito D, Weile J, Shendure J *et al.* Mavedb: an open-source platform to distribute and interpret data from multiplexed assays of variant effect. *Genome Biol* 2019;**20**:223.

Findlay GM, Boyle EA, Hause RJ *et al.* Saturation editing of genomic regions by multiplex homology-directed repair. *Nature* 2014;**513**:120–3. https://doi.org/10.1038/nature13695.

Fowler DM, Fields S. Deep mutational scanning: a new style of protein science. *Nat Methods* 2014;**11**:801–7.

Fowler DM, Araya CL, Fleishman SJ *et al.* High-resolution mapping of protein sequence-function relationships. *Nat Methods* 2010;**7**:741–6.

Fowler DM, Araya CL, Gerard W *et al.* Enrich: software for analysis of protein function by enrichment and depletion of variants. *Bioinformatics* 2011;**27**:3430–1.

Gasperini M, Starita L, Shendure J. The power of multiplexed functional analysis of genetic variants. *Nat Protoc* 2016;**11**:1782–7.

Haddox HK, Dingens AS, Bloom JD. Experimental estimation of the effects of all amino-acid mutations to hiv's envelope protein on viral replication in cell culture. *PLoS Pathog* 2016;**12**:e1006114.

Haddox HK, Dingens AS, Hilton SK *et al.* Mapping mutational effects along the evolutionary landscape of HIV envelope. *Elife* 2018;**7**:3. https://doi.org/10.7554/elife.34420.

Hietpas RT, Jensen JD, Bolon DNA. Experimental illumination of a fitness landscape. *Proc Natl Acad Sci USA* 2011;**108**:7896–901.

Hom N, Gentles L, Bloom JD *et al.* Deep mutational scan of the highly conserved influenza a virus M1 matrix protein reveals substantial intrinsic mutational tolerance. *J Virol* 2019;**93**:e00161-19. https://doi.org/10.1128/jvi.00161-19.

Lee B, Sohail MS, Finney E *et al.* Inferring effects of mutations on sars-cov-2 transmission from genomic surveillance data. *medRxiv* 2022. https://doi.org/10.1101/2021.12.31.21268591.

Lee JM, Huddleston J, Doud MB *et al.* Deep mutational scanning of hemagglutinin helps predict evolutionary fates of human h3n2 influenza variants. *Proc Natl Acad Sci USA* 2018;**115**:E8276–85.

Lei R, Hernandez Garcia A, Tan TJC *et al.* Mutational fitness landscape of human influenza h3n2 neuraminidase. *Cell Rep* 2023;**42**:113356.

Li C, Qian W, Maclean CJ *et al.* The fitness landscape of a trna gene. *Science* 2016;**352**:837–40.

Matuszewski S, Hildebrandt ME, Ghenu A-H *et al.* A statistical guide to the design of deep mutational scanning experiments. *Genetics* 2016;**204**:77–87.

Rich MS, Payen C, Rubin AF *et al.* Comprehensive analysis of the SUL1 promoter of *Saccharomyces cerevisiae*. *Genetics* 2016;**203**:191–202.

Romero PA, Tran TM, Abate AR. Dissecting enzyme function with microfluidic-based deep mutational scanning. *Proc Natl Acad Sci USA* 2015;**112**:7159–64.

Roop JI, Cassidy NA, Dingens AS *et al.* Identification of HIV-1 envelope mutations that enhance entry using macaque CD4 and CCR5. *Viruses* 2020;**12**:241. https://doi.org/10.3390/v12020241.

Rubin AF, Gelman H, Lucas N *et al.* A statistical framework for analyzing deep mutational scanning data. *Genome Biol* 2017;**18**:150–15.

Soh YS, Moncla LH, Eguia R *et al.* Comprehensive mapping of adaptation of the avian influenza polymerase protein PB2 to humans. *Elife* 2019;**8**:e45079. https://doi.org/10.7554/eLife.45079.

Sohail MS, Louie RHY, McKay MR *et al.* MPL resolves genetic linkage in fitness inference from complex evolutionary histories. *Nat Biotechnol* 2021;**39**:472–9. https://doi.org/10.1038/s41587-020-0737-3.

Sohail MS, Louie RHY, Hong Z *et al.* Inferring epistasis from genetic time-series data. *Mol Biol Evol* 2022;**39**:msac199.

Starita LM, Pruneda JN, Lo RS *et al.* Activity-enhancing mutations in an E3 ubiquitin ligase identified by high-throughput mutagenesis. *Proc Natl Acad Sci USA* 2013;**110**:E1263–72. https://doi.org/10.1073/pnas.1303309110.

Starita LM, Young DL, Islam M *et al.* Massively parallel functional analysis of BRCA1 RING domain variants. *Genetics* 2015;**200**:413–22. https://doi.org/10.1534/genetics.115.175802.

Starr TN, Greaney AJ, Hilton SK *et al.* Deep mutational scanning of sars-cov-2 receptor binding domain reveals constraints on folding and ace2 binding. *cell* 2020;**182**:1295–310.e20.

Thyagarajan B, Bloom JD. The inherent mutational tolerance and antigenic evolvability of influenza hemagglutinin. *Elife* 2014;**3**:e03300.

Bioinformatics, 2024, 40, 1–5
https://doi.org/10.1093/bioinformatics/btae499
Applications Note