PAPER

popDMS infers mutation effects from deep mutational scanning data

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

[1]Department of Physics and Astronomy, University of California, Riverside, CA, USA, [2]Department of Computational and Systems Biology, University of Pittsburgh School of Medicine, PA, USA, [3]Department of Physics and Astronomy, University of Pittsburgh, PA, USA and *Equal contributions.

[†]Corresponding author. jpbarton@pitt.edu

FOR PUBLISHER ONLY Received on Date Month Year; revised on Date Month Year; accepted on Date Month Year

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.

Contact: jpbarton@pitt.edu

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

Supplementary information: Supplementary data are available at Bioinformatics online.

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; Lee et al., 2018; Doud 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.

Results

We developed a method, popDMS, to estimate the functional effects of mutations in DMS experiments using statistical methods from population genetics (Methods). 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,

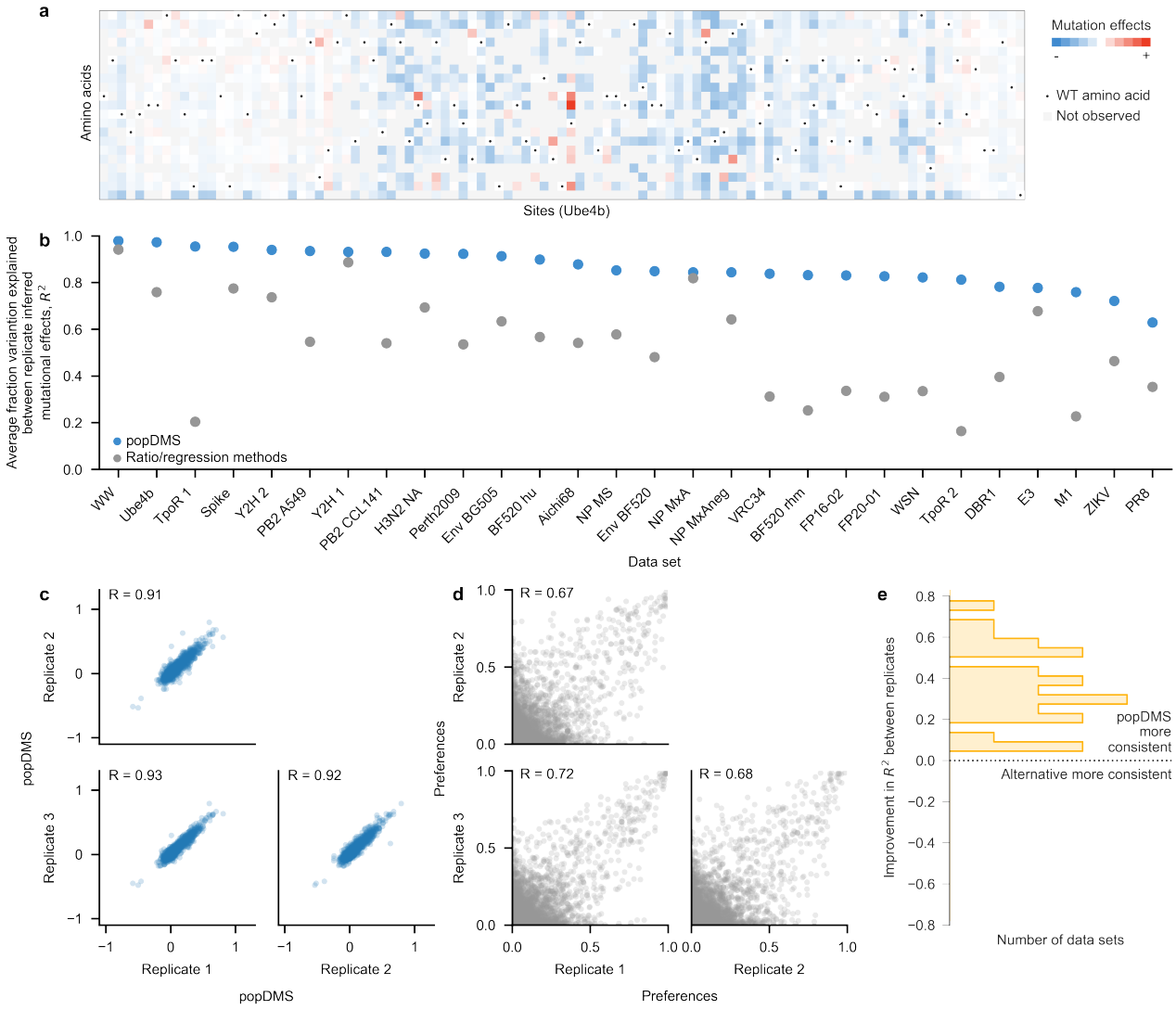

Fig. 1. popDMS overview. **a**, Example of the effects of mutations inferred by popDMS for the Ube4b protein(Starita et al., 2013). **b**, Across 28 data sets, popDMS infers more consistent mutational effects than previous ratio/regression-based methods. To illustrate consistency between replicates, we show (**c**) selection coefficients inferred across replicates for the HIV-1 envelop BF520 data set(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.

$\mathcal{L}((\boldsymbol{z}(t_k))_{k=0}^K \mid \boldsymbol{s})$ (see Methods for details). The $\boldsymbol{z}(t_k)$ represent vectors of variant frequencies $\boldsymbol{z}$ at different times t_k. The WF 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 $P_{\text{prior}}(\boldsymbol{s})$ 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,

$$\hat{\boldsymbol{s}} = \arg\max_{\boldsymbol{s}} \mathcal{L}\left(\boldsymbol{s} \mid (\boldsymbol{z}(t_k))_{k=0}^K\right) P_{\text{prior}}(\boldsymbol{s}) . \quad (1)$$

Typically, we adjust the width of the prior distribution based on the data, but a fixed value can also be specified (Methods). 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. 1**). The variant effects inferred by popDMS were also more

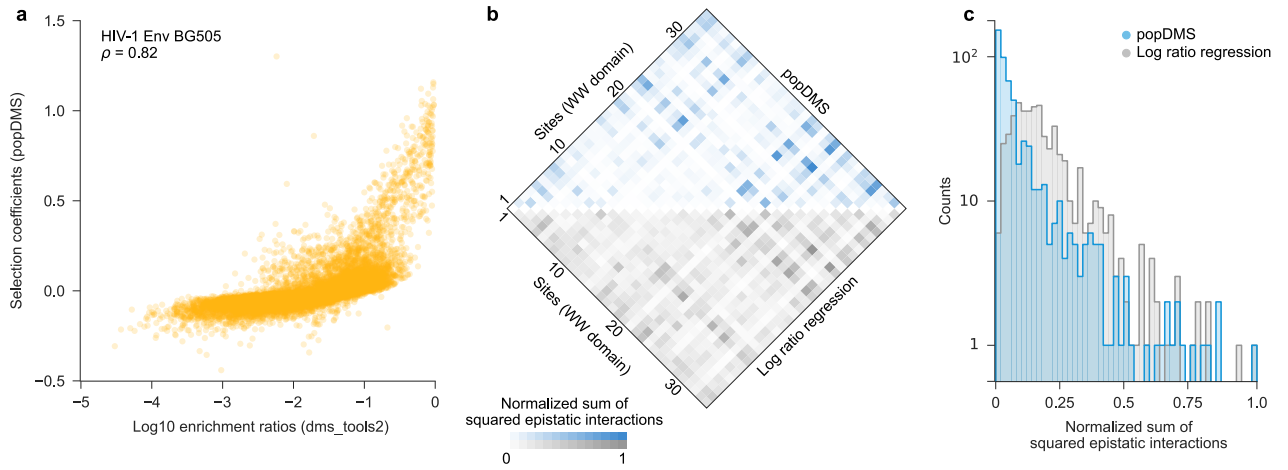

Fig. 2. Mutation effects inferred by popDMS are broadly consistent with alternative methods. a, For the HIV-1 Env BG505 data set, selection coefficients inferred by popDMS are congruent with enrichment ratios computed using dms_tools2 (Spearman's $\rho = 0.84$). At some sites, significant differences are observed (see **Supplementary Fig. 5**). **b**, In the hYAP65 WW domain data set, 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 (Methods). **c**, Epistatic interactions inferred by popDMS are substantially sparser than those inferred with the regression-based approach (Araya et al., 2012).

similar to true, underlying ones than alternative approaches, even with the addition of negative binomial sampling noise (**Supplementary Fig. 2**, see Methods).

Next, we analyzed a collection of 28 DMS data sets with popDMS (Araya et al., 2012; Starita et al., 2013; Findlay et al., 2014; Starita et al., 2015; Bridgford et al., 2020; Doud et al., 2015; Hom et al., 2019; Soh et al., 2019; Ashenberg et al., 2017; Haddox et al., 2018; Roop et al., 2020; Dingens et al., 2018; Li et al., 2016; Starr et al., 2020; Lei et al., 2023). These data sets were generated and analyzed using a variety of experimental techniques and analytical methods (see **Supplementary Table 1**). Like the functional metrics introduced by previous methods, selection coefficients provide an intuitive visualization of the functional effects of mutations (**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 data sets 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. 3**). 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 Methods).

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 data sets 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. 4**). 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 data sets 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 data set 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 data set (Dingens et al., 2018).

While the inferred mutation effects agreed for most sites, some showed qualitative differences (**Supplementary Fig. 5**). 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 Methods). 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 data set 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 (Methods). 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.

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 of data sets, 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, Methods).

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.

Acknowledgements

The work of Z.H., K.S.S, and J.P.B. reported in this publication was supported by the National Institute of General Medical Sciences of the National Institutes of Health under Award Number R35GM138233.

Contributions

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

References

Carlos L. Araya, Douglas M. Fowler, Wentao Chen, Ike Muniez, Jeffery W. Kelly, and Stanley Fields. A fundamental protein property, thermodynamic stability, revealed solely from large-scale measurements of protein function. *Proceedings of the National Academy of Sciences of the United States of America*, 109, 2012. ISSN 00278424. doi: 10.1073/pnas.1209751109.

Orr Ashenberg, Jai Padmakumar, Michael B. Doud, and Jesse D. Bloom. Deep mutational scanning identifies sites in influenza nucleoprotein that affect viral inhibition by mxa. *PLoS Pathogens*, 13, 2017. ISSN 15537374. doi: 10.1371/journal.ppat.1006288.

Jesse D. Bloom. Software for the analysis and visualization of deep mutational scanning data. *BMC Bioinformatics*, 16, 2015. ISSN 14712105. doi: 10.1186/s12859-015-0590-4.

Jessica L. Bridgford, Su Min Lee, Christine M.M. Lee, Paola Guglielmelli, Elisa Rumi, Daniela Pietra, Stephen Wilcox, Yash Chhabra, Alan F. Rubin, Mario Cazzola, Alessandro M. Vannucchi, Andrew J. Brooks, Matthew E. Call, and Melissa J. Call. Novel drivers and modifiers of MPL-dependent oncogenic transformation identified by deep mutational scanning. *Blood*, 135, 2020. ISSN 15280020. doi: 10.1182/blood.2019002561.

Bob Carpenter, Andrew Gelman, Matthew D Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus A Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: A probabilistic programming language. *Journal of statistical software*, 76, 2017.

Adam S Dingens, Hugh K Haddox, Julie Overbaugh, and Jesse D Bloom. Comprehensive mapping of hiv-1 escape from a broadly neutralizing antibody. *Cell host & microbe*, 21(6): 777–787, 2017.

Adam S. Dingens, Priyamvada Acharya, Hugh K. Haddox, Reda Rawi, Kai Xu, Gwo Yu Chuang, Hui Wei, Baoshan Zhang, John R. Mascola, Bridget Carragher, Clinton S. Potter, Julie Overbaugh, Peter D. Kwong, and Jesse D. Bloom. Complete functional mapping of infection- and vaccine-elicited antibodies against the fusion peptide of HIV. *PLoS Pathogens*, 14, 2018. ISSN 15537374. doi: 10.1371/journal.ppat.1007159.

Michael B. Doud, Orr Ashenberg, and Jesse D. Bloom. Site-specific amino acid preferences are mostly conserved in two closely related protein homologs. *Molecular Biology and Evolution*, 32, 2015. ISSN 15371719. doi: 10.1093/molbev/msv167.

Michael B Doud, Juhye M Lee, and Jesse D Bloom. How single mutations affect viral escape from broad and narrow antibodies to h1 influenza hemagglutinin. *Nature communications*, 9(1):1386, 2018.

Daniel Esposito, Jochen Weile, Jay Shendure, Lea M Starita, Anthony T Papenfuss, Frederick P Roth, Douglas M Fowler, and Alan F Rubin. Mavedb: an open-source platform to distribute and interpret data from multiplexed assays of variant effect. *Genome biology*, 20:1–11, 2019.

Warren John Ewens. *Mathematical population genetics: theoretical introduction*, volume 27. Springer, 2004.

Gregory M. Findlay, Evan A. Boyle, Ronald J. Hause, Jason C. Klein, and Jay Shendure. Saturation editing of genomic regions by multiplex homology-directed repair. *Nature*, 513, 2014. ISSN 14764687. doi: 10.1038/nature13695.

Douglas M Fowler and Stanley Fields. Deep mutational scanning: a new style of protein science. *Nature methods*, 11(8):801–807, 2014.

Douglas M Fowler, Carlos L Araya, Sarel J Fleishman, Elizabeth H Kellogg, Jason J Stephany, David Baker, and Stanley Fields. High-resolution mapping of protein sequence-function relationships. *Nature methods*, 7(9):741–746, 2010.

Douglas M Fowler, Carlos L Araya, Wayne Gerard, and Stanley Fields. Enrich: software for analysis of protein function by enrichment and depletion of variants. *Bioinformatics*, 27 (24):3430–3431, 2011.

Molly Gasperini, Lea Starita, and Jay Shendure. The power of multiplexed functional analysis of genetic variants. *Nature protocols*, 11(10):1782–1787, 2016.

Hugh K Haddox, Adam S Dingens, and Jesse D Bloom. Experimental estimation of the effects of all amino-acid

mutations to hiv's envelope protein on viral replication in cell culture. *PLoS pathogens*, 12(12):e1006114, 2016.

Hugh K Haddox, Adam S Dingens, Sarah K Hilton, Julie Overbaugh, and Jesse D Bloom. Mapping mutational effects along the evolutionary landscape of HIV envelope. *eLife*, 7, 3 2018. doi: 10.7554/elife.34420.

Ryan T Hietpas, Jeffrey D Jensen, and Daniel NA Bolon. Experimental illumination of a fitness landscape. *Proceedings of the National Academy of Sciences*, 108(19):7896–7901, 2011.

Nancy Hom, Lauren Gentles, Jesse D. Bloom, and Kelly K. Lee. Deep Mutational Scan of the Highly Conserved Influenza A Virus M1 Matrix Protein Reveals Substantial Intrinsic Mutational Tolerance. *Journal of Virology*, 93, 2019. ISSN 0022-538X. doi: 10.1128/jvi.00161-19.

Justus M Kebschull and Anthony M Zador. Sources of pcr-induced distortions in high-throughput sequencing data sets. *Nucleic acids research*, 43(21):e143–e143, 2015.

Motoo Kimura. Diffusion models in population genetics. *Journal of Applied Probability*, 1(2):177–232, 1964.

Brian Lee, Muhammad Saqib Sohail, Elizabeth Finney, Syed Faraz Ahmed, Ahmed Abdul Quadeer, Matthew R McKay, and John P Barton. Inferring effects of mutations on sars-cov-2 transmission from genomic surveillance data. *medRxiv*, pages 2021–12, 2022.

Juhye M Lee, John Huddleston, Michael B Doud, Kathryn A Hooper, Nicholas C Wu, Trevor Bedford, and Jesse D Bloom. Deep mutational scanning of hemagglutinin helps predict evolutionary fates of human h3n2 influenza variants. *Proceedings of the National Academy of Sciences*, 115(35): E8276–E8285, 2018.

Ruipeng Lei, Andrea Hernandez Garcia, Timothy JC Tan, Qi Wen Teo, Yiquan Wang, Xiwen Zhang, Shitong Luo, Satish K Nair, Jian Peng, and Nicholas C Wu. Mutational fitness landscape of human influenza h3n2 neuraminidase. *Cell reports*, 42(1), 2023.

Chuan Li, Wenfeng Qian, Calum J Maclean, and Jianzhi Zhang. The fitness landscape of a trna gene. *Science*, 352(6287): 837–840, 2016.

Sebastian Matuszewski, Marcel E Hildebrandt, Ana-Hermina Ghenu, Jeffrey D Jensen, and Claudia Bank. A statistical guide to the design of deep mutational scanning experiments. *Genetics*, 204(1):77–87, 2016.

Takahiro Nemoto, Tommaso Ocari, Arthur Planul, Muge Tekinsoy, Emilia A Zin, Deniz Dalkara, and Ulisse Ferrari. Acides: on-line monitoring of forward genetic screens for protein engineering. *Nature Communications*, 14(1):8504, 2023.

Vladimir Potapov and Jennifer L Ong. Examining sources of error in pcr by single-molecule sequencing. *PloS one*, 12(1): e0169774, 2017.

Matthew S Rich, Celia Payen, Alan F Rubin, Giang T Ong, Monica R Sanchez, Nozomu Yachie, Maitreya J Dunham, and Stanley Fields. Comprehensive analysis of the SUL1 promoter of Saccharomyces cerevisiae. *Genetics*, 203(1):191–202, 2016.

Philip A Romero, Tuan M Tran, and Adam R Abate. Dissecting enzyme function with microfluidic-based deep mutational scanning. *Proceedings of the National Academy of Sciences*, 112(23):7159–7164, 2015.

Jeremy I. Roop, Noah A. Cassidy, Adam S. Dingens, Jesse D. Bloom, and Julie Overbaugh. Identification of HIV-1 envelope mutations that enhance entry using macaque CD4 and CCR5. *Viruses*, 12, 2020. ISSN 19994915. doi: 10.3390/v12020241.

Alan F Rubin, Hannah Gelman, Nathan Lucas, Sandra M Bajjalieh, Anthony T Papenfuss, Terence P Speed, and Douglas M Fowler. A statistical framework for analyzing deep mutational scanning data. *Genome Biology*, 18(1): 1–15, 2017.

John Maynard Smith and John Haigh. The hitch-hiking effect of a favourable gene. *Genetics Research*, 23(1):23–35, 1974.

Yq Shirleen Soh, Louise H. Moncla, Rachel Eguia, Trevor Bedford, and Jesse D. Bloom. Comprehensive mapping of adaptation of the avian influenza polymerase protein PB2 to humans. *eLife*, 8, 2019. ISSN 2050084X. doi: 10.7554/eLife.45079.

Muhammad Saqib Sohail, Raymond H.Y. Louie, Matthew R. McKay, and John P. Barton. MPL resolves genetic linkage in fitness inference from complex evolutionary histories. *Nature Biotechnology*, 39, 2021. ISSN 15461696. doi: 10.1038/s41587-020-0737-3.

Muhammad Saqib Sohail, Raymond HY Louie, Zhenchen Hong, John P Barton, and Matthew R McKay. Inferring epistasis from genetic time-series data. *Molecular biology and evolution*, 39(10):msac199, 2022.

Marion Sourisseau, Daniel J. P. Lawrence, Megan C. Schwarz, Carina H. Storrs, Ethan C. Veit, Jesse D. Bloom, and Matthew J. Evans. Deep mutational scanning comprehensively maps how zika envelope protein mutations affect viral growth and antibody escape. *Journal of Virology*, 93(23):10.1128/jvi.01291–19, 2019.

Lea M. Starita, Jonathan N. Pruneda, Russell S. Lo, Douglas M. Fowler, Helen J. Kim, Joseph B. Hiatt, Jay Shendure, Peter S. Brzovic, Stanley Fields, and Rachel E. Klevit. Activity-enhancing mutations in an E3 ubiquitin ligase identified by high-throughput mutagenesis. *Proceedings of the National Academy of Sciences of the United States of America*, 110, 2013. ISSN 00278424. doi: 10.1073/pnas.1303309110.

Lea M. Starita, David L. Young, Muhtadi Islam, Jacob O. Kitzman, Justin Gullingsrud, Ronald J. Hause, Douglas M. Fowler, Jeffrey D. Parvin, Jay Shendure, and Stanley Fields. Massively parallel functional analysis of BRCA1 RING domain variants. *Genetics*, 200, 2015. ISSN 19432631. doi: 10.1534/genetics.115.175802.

Tyler N Starr, Allison J Greaney, Sarah K Hilton, Daniel Ellis, Katharine HD Crawford, Adam S Dingens, Mary Jane Navarro, John E Bowen, M Alejandra Tortorici, Alexandra C Walls, et al. Deep mutational scanning of sars-cov-2 receptor binding domain reveals constraints on folding and ace2 binding. *cell*, 182(5):1295–1310, 2020.

Paula Tataru, Thomas Bataillon, and Asger Hobolth. Inference under a wright-fisher model using an accurate beta approximation. *Genetics*, 201(3):1133–1141, 2015.

Bargavi Thyagarajan and Jesse D Bloom. The inherent mutational tolerance and antigenic evolvability of influenza hemagglutinin. *elife*, 3:e03300, 2014.