dms2dfe: Comprehensive Workflow for Analysis of Deep Mutational Scanning Data

Rohan Dandage[1,2,*], **Kausik Chakraborty**[1,2,*]

[1]CSIR Institute of Genomics and Integrative Biology, Mathura Road Campus, New Delhi 110020,India.

[2]Academy of Scientific and Innovative Research (AcSIR), New Delhi 110001, India.

***Corresponding authors**: kausik@igib.in, rohan@igib.in

Summary

High throughput genotype to phenotype (G2P) data is increasingly being generated by widely applicable Deep Mutational Scanning (DMS) method. dms2dfe is a comprehensive end-to-end workflow that addresses critical issue with noise reduction and offers variety of crucial downstream analyses. Noise reduction is carried out by normalizing counts of mutants by depth of sequencing and subsequent dispersion shrinkage at the level of calculation of preferential enrichments. In downstream analyses, dms2dfe workflow provides identification of relative selection pressures, potential molecular constraints and generation of data-rich visualizations.

Availability: dms2dfe is implemented as a python package and it is available at https://kc-lab.github.io/dms2dfe .

Contact: kausik@igib.in, rohan@igib.in

Supplementary information: Supplementary data are available at Bioinformatics online.

1 Introduction

Recent developments in high throughput mutagenesis and massively parallel DNA sequencing culminated in a method popularly known as Deep Mutational Scanning (DMS) (1,2). The broad appeal of DMS method is evident by the wide range of applicability (3) relevant to structural modeling (4), protein stability (5), substrate specificity (6–8), protein-protein interactions (9–14), mutational effects on organismal fitness (15–17), environmental effects (18), biology of viruses (19–21), fitness landscapes of antibiotic resistance (7,22) and proto-oncogenes(23). The distribution of fitness scores along fitness axis of fitness landscape i.e. distribution of fitness effects (DFE) is known as a powerful estimator of the underlying evolutionary dynamics and has been traditionally used to contextualize molecular evolution (24,25). One can characterize the nature of selection pressure in effect and potential susceptibility, evolvability or robustness of the individuals in the population by comparing DFEs.

Accounting and reducing for the noise is a major primary challenge in the analysis of differential count data. Adopting an alternative approach to the analysis than available tools (26–28), dms2dfe implements noise reduction utilizing both sequencing depth (Fig S1 A) and empirical Bayes shrinkage (29) (Fig S1 B). Along with upstream analysis to obtain variant counts, preferential enrichments and fitness estimates, dms2dfe offers utilities for downstream analysis such as identification of relative selection pressure, molecular constraints and visualizations of the data; thus constituting an end-to-end workflow. Comprehensively documented python programming library (API) (available at http://kc-lab.github.io/dms2dfe) provides a platform for further developments to adapt with advances and modifications in DMS experiments.

2 dms2dfe workflow

In a conventional DMS experiment, a pool of mutants is selected in a co-culture competition assay under selection pressure of interest and frequencies of the mutants are compared with a pool or mutants used as a reference (also referred as input or background). dms2dfe is applicable to all such approaches wherein a basic DMS experiment involves a comparative analysis of selected and reference pool of mutants. In addition to the shot gun and full length ultra-deep sequencing, dms2dfe also supports concatamer based approach (7) as well as multiplexing strategy using barcoded amplicons (30). Sequencing data whether aligned (.sam, .bam) or not (.fastq) can be provided as input of dms2dfe workflow. Alternatively, mutational data (frequencies, preferential enrichments or fitness score) can be provided to exclusively carry out downstream analyses.

As shown in Fig 1 A, in the comprehensive workflow of dms2dfe, following key aspects of the analysis of DMS experiment are addressed.

2.1 identification of relative selection pressure

As previously implemented (8), we classify the statistically neutral (least effect) mutants upper and lower thresholds from the preferential enrichments of reference (unselected) pool. Depending on the nature of experiment, fitness scores can be estimated by one of several methods based on preferential enrichments of wild type alleles or otherwise preferential enrichments themselves can be used as fitness scores (see S1 *Supporting Methods*). So, mutants can be classified as beneficial, deleterious or neutral as follows,

$$M_i \in \begin{cases} enriched & if\ F_i > \mu + 2\sigma \\ neutral & if\ \mu + 2\sigma \geq F_i \geq \mu - 2\sigma \\ depleted & if\ F_i < \mu - 2\sigma \end{cases} \quad (1)$$

where, M_i is i^{th} kind of survived mutant and F_i is its respective fitness score. Note that class neutral exclusively represents the synonymous mutants. Metrics such as relative number of mutations under the classes of fitness (Fig 1B), relative survivability i.e. difference in number of mutants survived in test condition with respect to control condition (Δn) and relative fitness changes i.e. difference in average fitness of mutants in test condition with respect to control condition (ΔF) assist users to identify the directionality and strength of the relative selection pressures.

2.2 Identification of molecular constraints

Due to multiple reasons (31), effects of mutation at one site are dependent on residues at other sites. This results in a highly interdependent ensemble of molecular constraints which is difficult to decipher (32). Here we use ensemble machine learning algorithm - Gradient Boosting which allows modeling of the complex mutational data. It provides critical advantage of determination by relative importances of the features (Fig 1 C) and partial dependence i.e. marginal effect of features on the target i.e. fitness scores (Fig S2). Such information would help users in contextualizing G2P interactions in terms of how fitness scores are mechanistically constrained by molecular constraints (Fig S1 D).

2.3 Visualizations

Fitness data too can be visualized in three different formats. Fitness per individual mutations is visualized in the form of a heatmap also called sequence-function map, wherein each locus represents fitness of individual mutant (Fig 1 E). Fitness per substitutions too can be visualized in the form of matrix of average fitness scores per individual substitution (Fig 1 F). Average fitness values per position are projected onto the PDB structure by utilizing visualization modules of UCSF-Chimera (34)(Fig 1 G). Alongside, an integrated visualization of the change in

frequencies of mutants at the level of individual mutations, substitutions and positions is generated (Fig S1 B).

3 Conclusions

dms2dfe is a tunable, open-source workflow that integrates state-of-the-art methods of genomics for analysis of DMS data. Accounting variable sequencing depth across the length of the gene and implementing dispersion shrinkage though empirical Bayes approach results in reduction in the noise associated with empirical data. Downstream analyses such as identification of relative selection pressures, identification of molecular constraints, and intuitive visualizations help in contextualizing the DMS data. Collectively, dms2dfe offers experimenters a comprehensive solution for the analysis of DMS data and an open-source platform for future developments to accommodate potential modifications of DMS method.

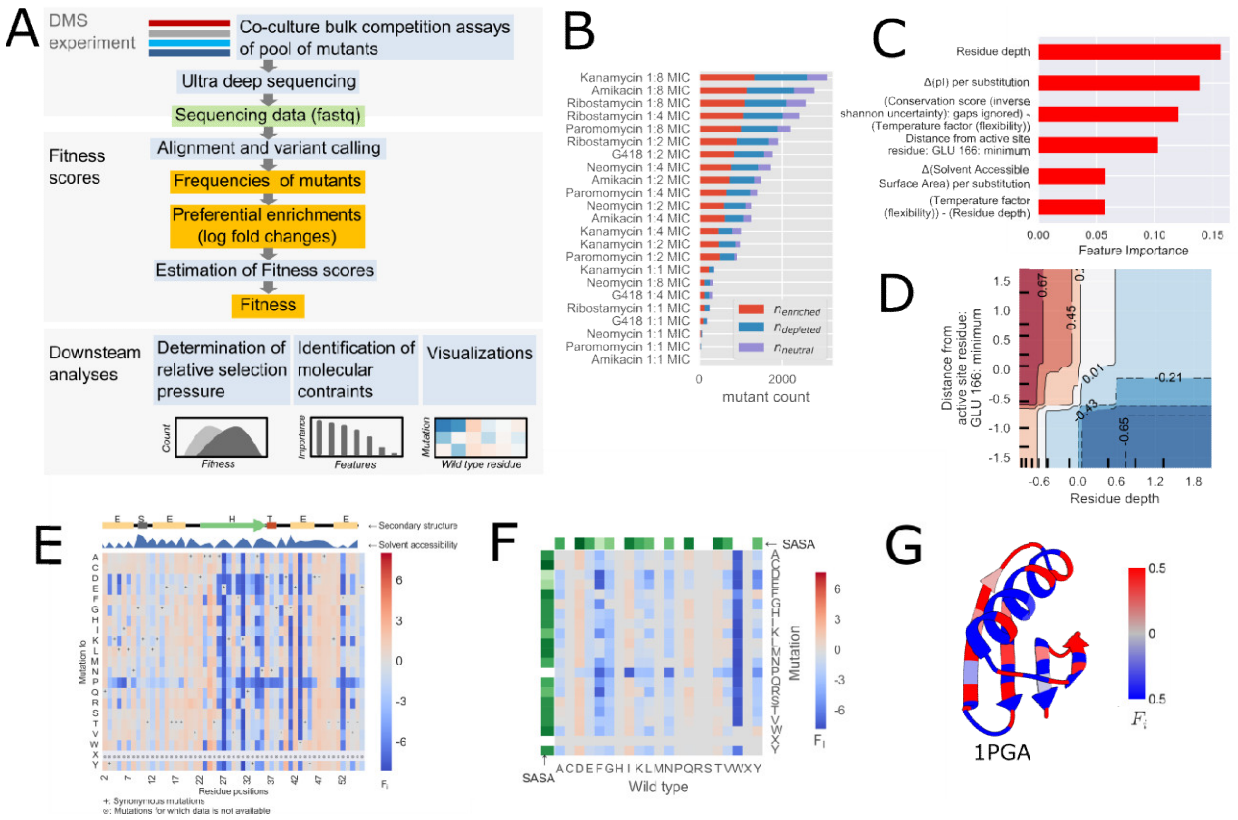

Fig. 1. Workflow and functionalities of dms2dfe. (A) Sequencing data produced by DMS experiment is preprocessed, aligned and variants are identified. Preferential enrichments (log fold changes) are used to estimate Fitness scores according to several methods (see S1 *Supporting methods*). Among downstream analyses, dms2dfe provides methods to identify the selection pressure among multiple DFEs, determine molecular constraints of the fitness landscape and generate various visualizations of the mutational data. (B) Demonstration of determination of relative selection pressure in the form of counts of mutants classified as enriched ($n_{enriched}$), depleted ($n_{depleted}$), and neutral ($n_{neutral}$). (C and D) Identification of molecular constraints of fitness data. (C) Feature importance of top features with highest feature importance as determined a binary classification model (S1 Materials and Methods) are indentified as strong constraints. (D) Contour plot lined by top molecular constraints with partial dependence of the predictors is scaled along z axis. Axis ticks in represent decile values of the features. (E-F) Visualizations of mutational data in three different formats i.e. (E) fitness per

mutations (F) average fitness per substitutions and (G) average fitness per positions of the protein. In panel E, each locus of the heatmap represents fitness of individual mutant. In panel F, each locus of the heatmap represents average fitness score per substitution. In panel G, average fitness score per residue position is projected onto PDB structure (1PGA) of the protein. Here, for generating plot B, APH2 dataset was used. For generating plots C and D, APH2 dataset was used. For generating plot E-H, GB1 dataset was used.

Additional files

S1 text: Supporting Methods, Figures and Tables.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

RD and KC designed the method. RD implemented the method. RD and KC wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgments

We acknowledge CSIR for its funding through EMPOWER project and infrastructural support from CSIR IGIB. R.D. acknowledges UGC for graduate funding. We thank the members of KC lab for critically reviewing the manuscript.

References

1. Fowler DM, Araya CL, Fleishman SJ, Kellogg EH, Stephany JJ, Baker D, et al. High-resolution mapping of protein sequence-function relationships. Nat Methods [Internet]. 2010;7(9):741–6. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2938879

2. Fowler DM, Fields S. Deep mutational scanning: a new style of protein science. Nat Methods [Internet]. 2014;11(8):801–7. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4410700

3. Shendure J, Fields S. Massively Parallel Genetics. Genetics. 2016;203(June):617–9.

4. Adkar B V, Tripathi A, Sahoo A, Bajaj K, Goswami D, Chakrabarti P, et al. Protein model discrimination using mutational sensitivity derived from deep sequencing. Structure. 2012;20(2):371–81.

5. Araya CL, Fowler DM, Chen W, Muniez I, Kelly JW, Fields S. A fundamental protein property, thermodynamic stability, revealed solely from large-scale measurements of protein function. Proc Natl Acad Sci. 2012;109(42):16858–63.

6. Thyme SB, Song Y, Brunette TJ, Szeto MD, Kusak L, Bradley P, et al. Massively parallel determination and modeling of endonuclease substrate specificity. Nucleic Acids Res. 2014;42(22):13839–52.

7. Melnikov A, Rogov P, Wang L, Gnirke A, Mikkelsen TS. Comprehensive mutational scanning of a kinase in vivo reveals substrate-dependent fitness landscapes. Nucleic Acids Res. 2014;42(14).

8. Stiffler MA, Hekstra DR, Ranganathan R. Evolvability as a Function of Purifying Selection in TEM-1 β-Lactamase. Cell [Internet]. 2015;160(5):882–92. Available from:

http://dx.doi.org/10.1016/j.cell.2015.01.035

9. McLaughlin RN, Poelwijk FJ, Raman A, Gosal WS, Ranganathan R. The spatial architecture of protein function and adaptation. Nature [Internet]. 2012;491(7422):138–42. Available from: http://www.nature.com/nature/journal/v491/n7422/abs/nature11500.html%7B#%7Daffil-auth

10. Fujino Y, Fujita R, Wada K, Fujishige K, Kanamori T, Hunt L, et al. Robust in vitro affinity maturation strategy based on interface-focused high-throughput mutational scanning. Biochem Biophys Res Commun [Internet]. 2012;428(3):395–400. Available from: http://dx.doi.org/10.1016/j.bbrc.2012.10.066

11. Moretti R, Fleishman SJ, Agius R, Torchala M, Bates PA, Kastritis PL, et al. Community-wide evaluation of methods for predicting the effect of mutations on protein-protein interactions. Proteins Struct Funct Bioinforma. 2013;81(11):1980–7.

12. Forsyth CM, Juan V, Akamatsu Y, DuBridge RB, Doan M, Ivanov A V, et al. Deep mutational scanning of an antibody against epidermal growth factor receptor using mammalian cell display and massively parallel pyrosequencing. MAbs. 2013;5(4):523–32.

13. Starita LM, Pruneda JN, Lo RS, Fowler DM, Kim HJ, Hiatt JB, et al. Activity-enhancing mutations in an E3 ubiquitin ligase identified by high-throughput mutagenesis. Proc Natl Acad Sci U S A [Internet]. 2013;110(14):E1263--72. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3619334

14. Van Blarcom T, Rossi A, Foletti D, Sundar P, Pitts S, Bee C, et al. Precise and efficient antibody epitope determination through library design, yeast display and next-generation sequencing. J Mol Biol [Internet]. 2015;427(6):1513–34. Available from:

http://dx.doi.org/10.1016/j.jmb.2014.09.020

15. Roscoe BP, Thayer KM, Zeldovich KB, Fushman D, Bolon DNA. Analyses of the effects of all ubiquitin point mutants on yeast growth rate. J Mol Biol [Internet]. 2013;425(8):1363–77. Available from: http://dx.doi.org/10.1016/j.jmb.2013.01.032

16. Melamed D, Young DL, Gamble CE, Miller CR, Fields S. Deep mutational scanning of an RRM domain of the Saccharomyces cerevisiae poly(A)-binding protein. RNA [Internet]. 2013;19(11):1537–51. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3851721

17. Mishra P, Flynn JM, Starr TN, Bolon DNA. Systematic Mutant Analyses Elucidate General and Client-Specific Aspects of Hsp90 Function. Cell Rep [Internet]. 2016;15(3):588–98. Available from: http://linkinghub.elsevier.com/retrieve/pii/S2211124716303175

18. Mavor D, Barlow K, Thompson S, Barad BA, Bonny AR, Cario CL, et al. Determination of ubiquitin fitness landscapes under different chemical stresses in a classroom setting. Elife [Internet]. 2016;5:1–23. Available from: http://elifesciences.org/lookup/doi/10.7554/eLife.15802

19. Whitehead TA, Chevalier A, Song Y, Dreyfus C, Fleishman SJ, De Mattos C, et al. Optimization of affinity, specificity and function of designed influenza inhibitors using deep sequencing. Nat Biotechnol [Internet]. 2012;30(6):543–8. Available from: http://dx.doi.org/10.1038/nbt.2214,

20. Bloom JD. An experimentally determined evolutionary model dramatically improves phylogenetic fit. Mol Biol Evol. 2014;31(8):1956–78.

21. Thyagarajan B, Bloom JD. The inherent mutational tolerance and antigenic evolvability of

influenza hemagglutinin. Elife. 2014;2014(3).

22. Firnberg E, Labonte JW, Gray JJ, Ostermeier M. A comprehensive, high-resolution map of a Gene's fitness landscape. Mol Biol Evol. 2014;31(6):1581–92.

23. Starita LM, Young DL, Islam M, Kitzman JO, Gullingsrud J, Hause RJ, et al. Massively parallel functional analysis of BRCA1 RING domain variants. Genetics. 2015;200(2):413–22.

24. Kimura M. Model of effectively neutral mutations in which selective constraint is incorporated. Proc Natl Acad Sci U S A. 1979;76(7):3440–4.

25. Ohta T. The Nearly Neutral Theory Of Molecular Evolution. Annu Rev Ecol Syst [Internet]. 1992;23(1992):263–86. Available from: http://www.jstor.org/stable/2097289

26. Fowler DM, Araya CL, Gerard W, Fields S. Enrich: Software for analysis of protein function by enrichment and depletion of variants. Bioinformatics. 2011;27(24):3430–1.

27. Bloom JD. Software for the analysis and visualization of deep mutational scanning data. BMC Bioinformatics [Internet]. 2015;16(1):1–13. Available from: http://www.biomedcentral.com/1471-2105/16/168

28. Rubin AF, Lucas N, Bajjalieh SM, Papenfuss AT, Speed TP, Fowler DM. Enrich2: a statistical framework for analyzing deep mutational scanning data. 2016; Available from: http://biorxiv.org/lookup/doi/10.1101/075150

29. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol [Internet]. 2014;15(12):550. Available from: http://genomebiology.biomedcentral.com/articles/10.1186/s13059-014-0550-8

30. Dandage R, Pandey R, Jayaraj G, Chakraborty K. Differential Strengths Of Molecular

Determinants Guide Environment Specific Mutational Fates. bioRxiv [Internet]. 2017; Available from: http://biorxiv.org/content/early/2017/05/05/134569?rss=1&utm_source=dlvr.it&utm_medium=twitter

31. Starr TN, Thornton JW. Epistasis in protein evolution. Protein Science. 2016. p. 1204–18.

32. Boucher JI, Bolon DNA, Tawfik DS. Quantifying and understanding the fitness effects of protein mutations: Laboratory versus nature. Protein Sci. 2016;25:1219–26.

33. Katsonis P, Lichtarge O. A formal perturbation equation between genotype and phenotype determines the Evolutionary Action of protein-coding variations on fitness. Genome Res. 2014;24(12):2050–8.

34. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, et al. UCSF Chimera--a visualization system for exploratory research and analysis. J Comput Chem [Internet]. 2004;25(13):1605–12. Available from: http://www.ncbi.nlm.nih.gov/pubmed/15264254