bioRxiv preprint doi: https://doi.org/10.1101/072645; this version posted August 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

Abstract

Deep Mutational Scanning is a robust method for massive scale assessment of genotypic variants and is being applied to wide domains of research. *dms2dfe* (Deep Mutational Scanning to Distribution of Fitness Effects) is a comprehensive computational workflow designed to streamline analysis of such data on the basis of evolutionary principles. *dms2dfe* assists in contextualizing data from Deep Mutational Scanning experiment in terms of Distribution of Fitness Effects which is a powerful indicator of evolutionary dynamics. In addition to estimations of preferential enrichments of experimentally determined mutations, *dms2dfe* utilizes a novel application of robust random forest modeling to infer of preferential enrichments of mutants which are not empirically determined. This helps to deduce biologically relevant interpretations from population level dynamics of DFEs across different experimental conditions by solving normalization issue and sampling bias. *dms2dfe* is available at https://kc-lab.github.io/dms2dfe .

Keywords: Deep Mutational Scanning, Distribution of Fitness Effects, Deep sequencing, Machine learning

bioRxiv preprint doi: https://doi.org/10.1101/072645; this version posted August 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Rationale

Recent developments in high throughput mutagenesis and massively parallel DNA sequencing culminated in a method popularly known as Deep Mutational Scanning (DMS) [1, 2]. A conventional Deep Mutational Scanning experiment involves selection of a pool of a mutagenized gene under selection pressure of interest followed by deep sequencing of amplicons by next-generation sequencing. Due to the combination of ease of experimentation, cost effectiveness and generation of massive scale information rich data, Deep Mutational Scanning is being widely employed in diverse fields of research [3–22]. In order to streamline and standardize analysis of such data, a comprehensive analysis pipeline is needed which would additionally also help in design and optimization of DMS experiment.

Comprehensive fitness landscapes produced from Deep Mutational Scanning provide information about all possible evolutionary paths available for a gene - including the ones which are not accessible through natural selection - which makes it valuable in challenging task of predicting effects of new mutations [23]. The distribution of fitnesses along fitness axis of fitness landscape i.e. distribution of fitness effects (DFE) is a powerful estimator of the underlying evolutionary dynamics and has been traditionally used to contextualize molecular evolution [24, 25]. One can easily characterize the nature of selection pressure in effect and potential susceptibility, evolvability or robustness of the individuals in the population by comparing DFEs. Deep Mutational Scanning provides empirical fitnesses of large scale of mutations which useful for fundamental studies in population genetics which are heavily dependent on the analysis of Distribution of Fitness Effects (DFEs).

As compared to previously reported workflow [26], *dms2dfe* provides a more tunable framework which renders wider applicability across variety of modifications of Deep Mutational Scanning method. *dms2dfe* also provides utilities to infer the fitnesses of mutants which are not empirically recorded, by utilizing a robust machine learning approach of random forest decision trees. We find that this ensemble approach performs better than previously reported approach [27] across independent datasets. The resultant cumulative DFE encompassing empirical and inferred fitnesses increase the information available per fitness landscape which is critical to make population level interpretations. *dms2dfe* workflow also provides mechanistic insights underlying observed changes in organismal fitnesses, in terms of ranked relative importances of molecular features involved in the activity, folding or molecular dynamics of the gene in concern.

In the comprehensive workflow of *dms2dfe*, following key aspects of the analysis of Deep Mutational Scanning experiment are addressed.

bioRxiv preprint doi: https://doi.org/10.1101/072645; this version posted August 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

1. Assessment of quality of mutational data in terms of sequencing depth and biological and technical noise.
2. Variant calling and estimation of frequency of mutants per amino acid and codon.
3. Estimation of fitnesses of mutants from preferential enrichments of mutants under selection pressure with respect to frequency of mutants under no selection pressure.
4. Inference of fitness of mutants which are not empirically determined.
5. Comparison of resultant DFEs across different conditions to characterize nature of condition specific selection pressure in effect.

Results

Applicability of *dms2dfe*

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 pool or mutants used as an input pool or a pool under no selection pressure. *dms2dfe* is applicable to all such approaches wherein a basic DMS experiment involves a comparative analysis of input and selected pool of mutants. In addition to the shot gun and full length ultra-deep sequencing, *dms2dfe* also supports concatamer based approach [13] as well as multiplexing strategy using barcoded amplicons.

dms2dfe workflow can also import sequencing data (in unaligned fastq or aligned sam or bam formats) or alternatively, a mutation matrix of frequencies of mutants can provided. Also pre-estimated fold changes or fitness scores of mutants can be provided to infer fitness scores of the mutants that are not empirically determined. Along with input sequencing or mutation data, *dfe2dfe* requires the reference sequence (in fasta format) and PDB structure of the protein. Alongside informative visualizations, *dms2dfe* also stores the data associated with each analysis in widely applicable comma separated text (csv) format.

Estimation of preferential enrichments

Variant calling

dms2dfe provides utilities to widely popular preprocessing tools for quality filtering (Trimmomatic [28]) and alignment (bowtie2 [29] and samtools [30]) to process raw sequencing data to get aligned reads which are used for subsequently variant calling. By iterating through aligned sequences, based on position in the

bioRxiv preprint doi: https://doi.org/10.1101/072645; this version posted August 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

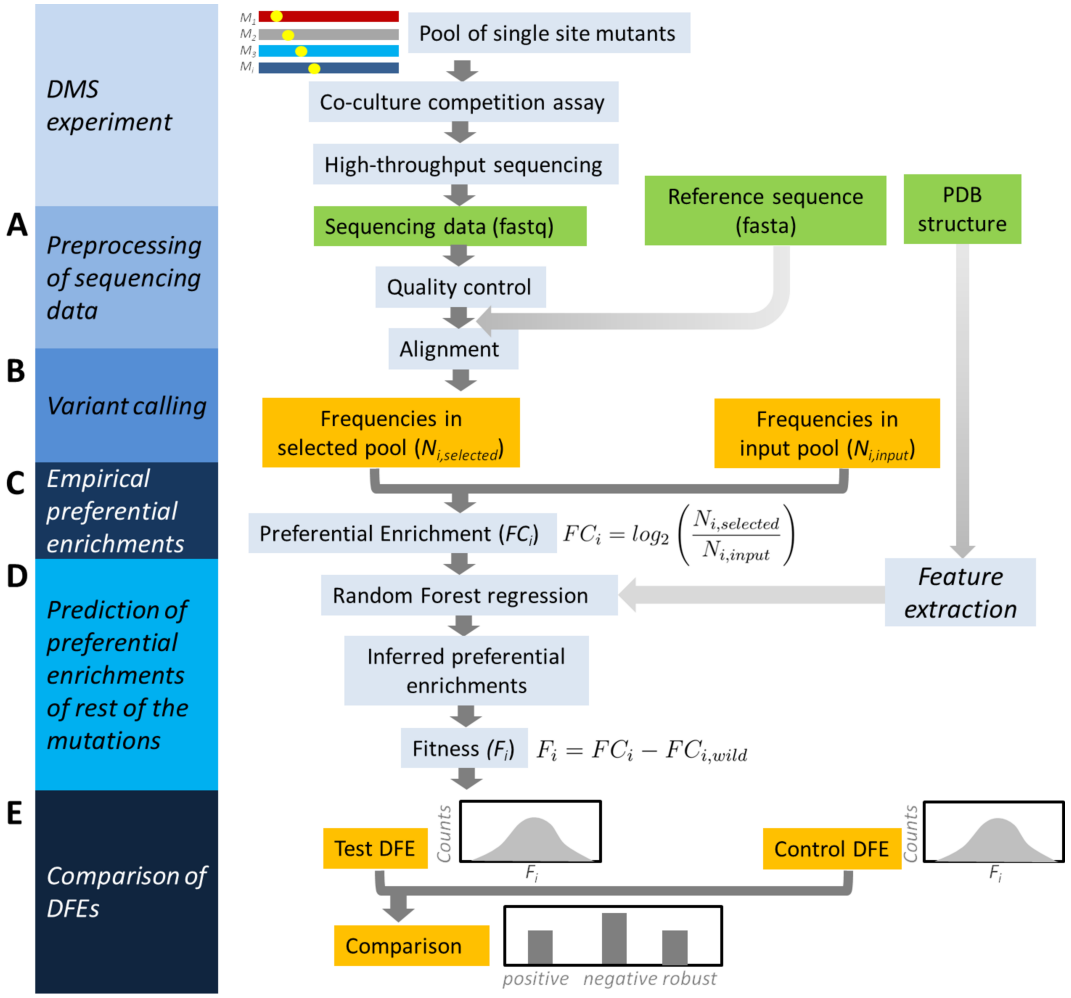

Figure 1: Structure of *dms2dfe*'s analysis workflow. Shown in green and yellow boxes are inputs and outputs of the workflow respectively. (A) In the first step, preprocessing of sequencing data is carried out by quality filtering to filter out reads with lower Q-score than threshold and subsequent alignment to reference sequence. (B) Variants are then identified from the aligned sequences and stored in the form of frequencies of mutations (N_i). (C) Empirical preferential enrichments (FC_i) are estimated as log fold change of frequencies of mutants in selected and input libraries. (D) Preferential enrichments which are not empirically recorded are inferred by random forest regressor using existing empirical fold changes as a training set and molecular features extracted from the feature extraction tools. Fitness scores are estimated by rescaling preferential enrichments with respect to preferential enrichments of wild type (synonymous alleles). (E) In the downstream analysis, DFEs of test and control experimental conditions are compared and accordingly mutants are categorized into 'positive', 'negative' or 'robust' classes, the relative proportions of which provide information about the condition specific selection pressure. The full documentation of *dms2dfe* is available at https://kc-lab.github.io/dms2dfe.

bioRxiv preprint doi: https://doi.org/10.1101/072645; this version posted August 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

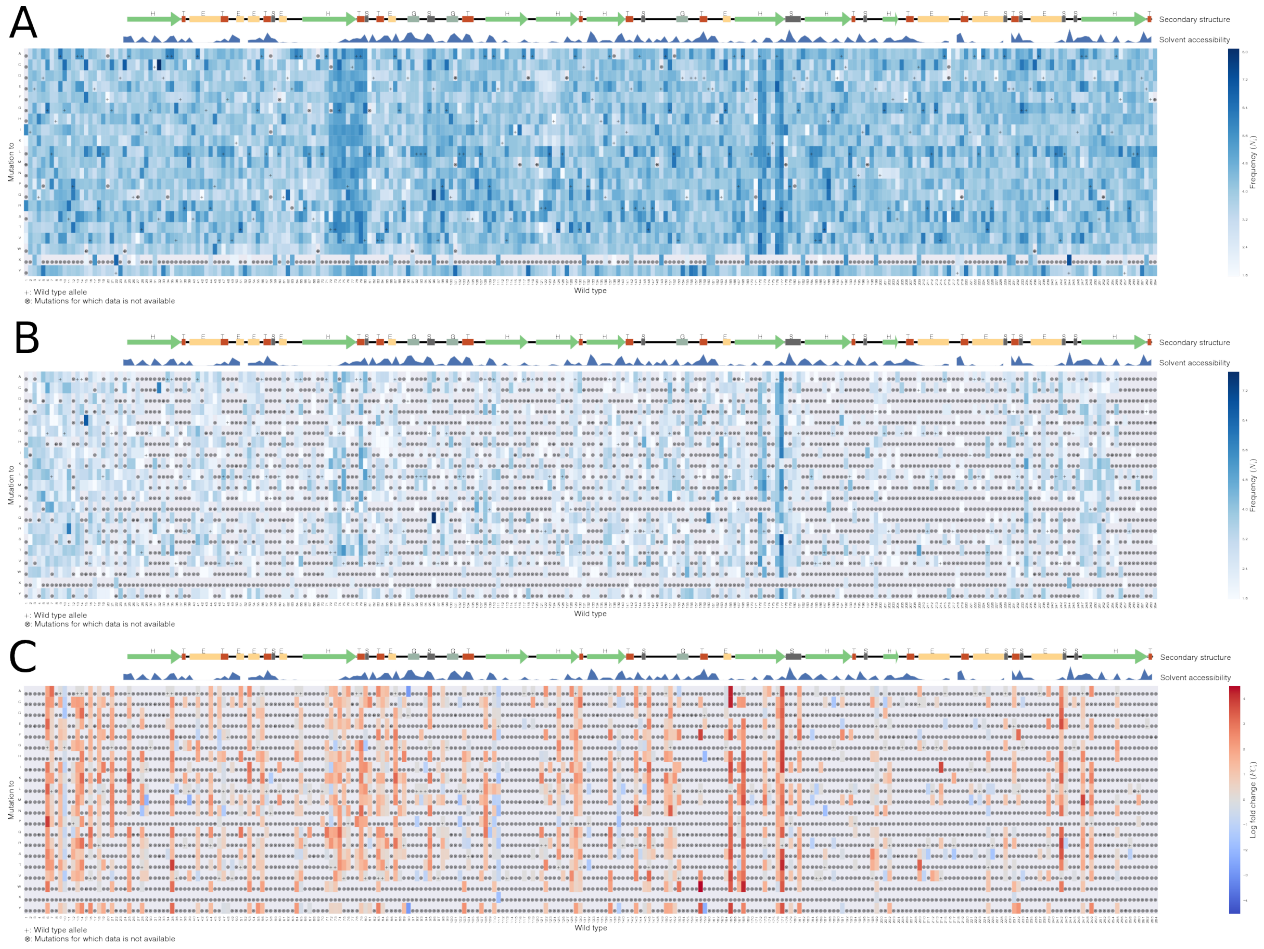

Figure 2: Visualizations of mutational data in the form of mutational matrix. Positions of wild type residues are represented on x-axis and corresponding mutations are set on y-axis. Here, for an example dataset - APH2, the frequencies (log transformed) of mutants in input pool, selected pool and preferential enrichments are represented in panels A B and C respectively.

bioRxiv preprint doi: https://doi.org/10.1101/072645; this version posted August 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

corresponding reading frames, nucleotide variants are called and are reported at codon level. Mutations at codon level are then transformed into a mutation matrix (also known as sequence-function map) in which each locus identifies read depth of a unique mutation (Figure 2A and B). Codon level mutations are scored only if the average Q-score of the codon is greater than a threshold which can be provided through input configuration of *dms2dfe*. Also to avoid biases in the ratio based estimation of preferential enrichments a threshold can be set to filter out mutants with low frequencies.

Preferential enrichments

Preferential enrichments are estimated as log ratios of the frequencies of the mutants in selected pool with respect to input pool (Equation 1). The accuracy of the fitness estimations in DMS experiment is dependent on levels of frequencies of mutants which in turn depend on the sequencing depth of the respective libraries [31, 32]. Given that the frequencies of mutants are at optimal levels - which can be assessed through *dms2dfe* s quality check utilities (described in subsequent sections) - a ratio based approach serve as both efficient and intuitive way to estimate preferential enrichments.

Estimation fitness scores

Normalization of preferential enrichments

In order to determine whether a mutant is beneficial or deleterious its fitness is compared with the fitness of wild type allele which serves as a reference. Two approaches are implemented in *dms2dfe* to normalize preferential enrichments of mutants with respect to wild type. Either preferential enrichments of wild type (synonymous alleles) or sequencing depths of wild type sequences can be employed as normalization factor.

Since all the synonymous mutants would express into the same wild-type protein, in first method, preferential enrichments of synonymous mutations are used as a normalizing factor. With the conventional methods used to generate mutation libraries e.g. using degenerate (NNK or NNS) codons, along with non-synonymous mutations, a proportion of synonymous mutations are produced. For normalization, preferential enrichments of all mutants are z-score normalized with respect to maximum likelihood and standard deviation of distribution of the preferential enrichment of synonymous mutations (Equation 2). This method is applicable in cases where proportions of synonymous mutations are relatively invariable across experimental conditions.

Alternatively, through *dms2dfe*, preferential enrichments of mutants can also be normalized with respect to

bioRxiv preprint doi: https://doi.org/10.1101/072645; this version posted August 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

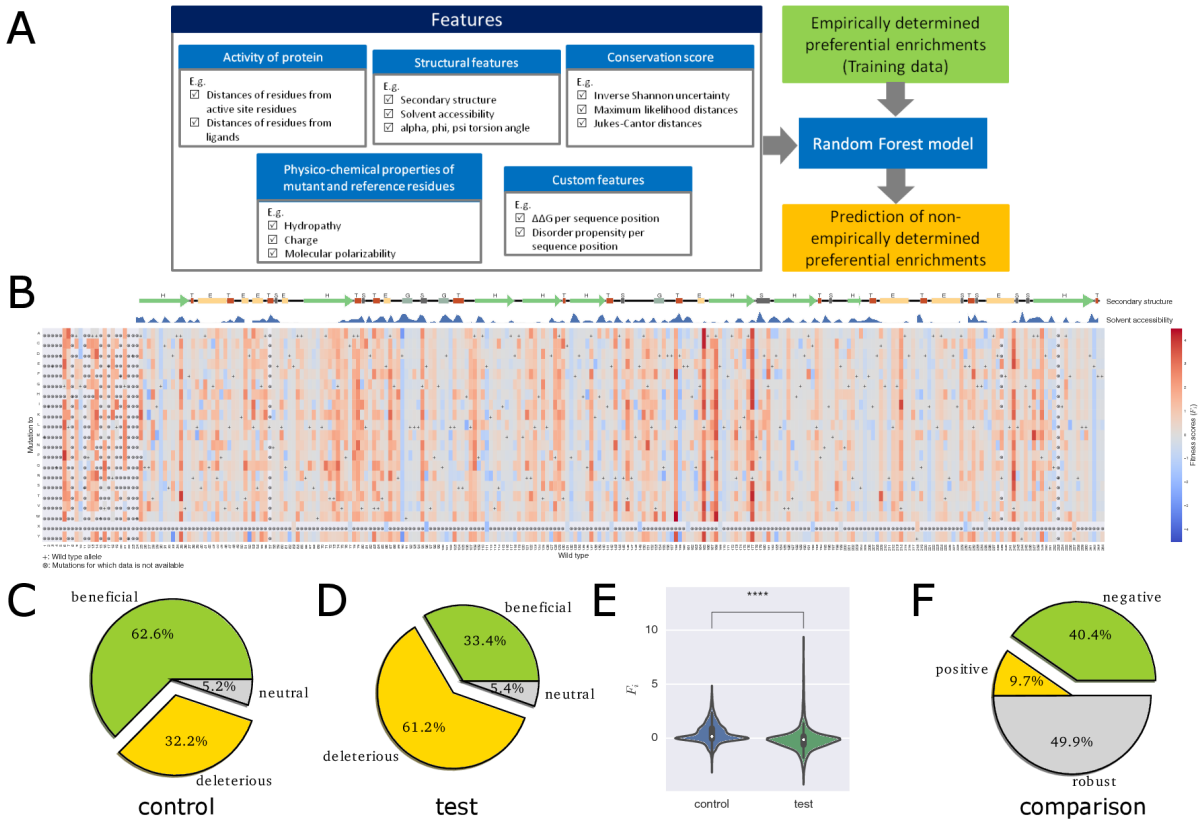

Figure 3: Estimation of fitness scores and downstream analysis. (A) Molecular features relevant to the fitness of mutants such as activity, folding of protein, evolutionary conservation and physico-chemical properties of amino acids and user provided custom features are used to train random forest regressor to infer preferential enrichments of the non-empirically determined mutants from empirically determined ones. (B) The preferential enrichments of all mutants are rescaled with respect to that of wild type allele. Resultant cumulative mutation matrix of empirical and inferred fitness scores allows more coverage of fitness landscape. *dms2dfe* suite also provides modules to compare fitness of mutants across of multiple test conditions. Here, for demonstration, from APH2 dataset, test (Kanamycin concentration of 1:1 MIC) and control (Kanamycin concentration of 1:4 MIC) conditions are used. Panels (C) and (D) proportions of beneficial, neutral and deleterious mutations in control and test conditions respectively. (E) owing to evident more deleteriousness of the test condition the DFE of test condition show a significant ($p<10^{-4}$) disparity as compared to DFE of control. (F) Depending on the comparative dynamics of DFEs, the mutants can be categorized as under negative or positive selection pressure or robust (Equation 6). Relative proportions of mutants in such categories for test dataset APH2 shows that the selection pressure of test (higher antibiotic concentration) shifts mutants to the negative class and hence characterizing the nature of the selection pressure of the condition.

bioRxiv preprint doi: https://doi.org/10.1101/072645; this version posted August 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

sequencing depths of wild-type alleles as previously described by Melnikov et al. [13] (Equation 3). This approach is especially useful in scenarios where frequencies of synonymous mutants are considered to vary across experimental conditions or in cases where synonymous mutations are not generated by methods used for cloning mutants.

Estimation of fitness scores by rescaling preferential enrichments relative to wild type

Since DMS studies are often based on co-culture competition assays, preferential enrichment of mutants can be best interpreted in terms of corresponding changes in their relative fitnesses. Preferential enrichments of mutants at a position are rescaled with respect to the wild type (synonymous allele) at that position (Equation 4). Since fitness scores scaled with respect to that of wild type alleles, they are classified into categories beneficial or deleterious if there fitness is greater or lower than that of the wild type (Equation 5 and Figure 3C and D).

Need to infer preferential enrichments

Normalization issues

Since their functional importances would be conserved and equal to wild type allele, fold changes of synonymous mutations serve as benchmark of preferential enrichment of wild type and hence can be used as a reference to estimated fitness of mutants. But due limited efficiencies of widely used cloning or transformation methods, fraction of synonymous mutants are lost before sequencing. In such cases, normalization of preferential enrichments and subsequent rescaling relative to wild type (synonymous alleles) to determine whether the mutation under consideration is beneficial or deleterious becomes impossible and preferential enrichment values of the rest of the mutants at that position are rendered unusable.

In other popularly used normalization method, frequencies of mutants are normalized with respect to the total sequencing depth at that position. The fitness values estimated by this approach are rendered sensitive to changes in levels of sequencing depth along the length of reference sequence. This is especially prevalent the case in shot gun sequencing methods e.g. tagmentation based library preparations.

Possible sampling biases

To characterize the condition specific selection pressure without a sampling bias, fitnesses of majority of individual representatives from all the populations being compared are required. In case of highly deleterious

bioRxiv preprint doi: https://doi.org/10.1101/072645; this version posted August 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

selection pressures, fraction of the mutants can be killed so their frequencies are not empirically determined. Additionally, current widely used method for generation of mutation libraries including degenerate (NNK or NNS) codon is inefficient in cloning all the substitutions at every position. Even with the use of synthetic oligos due to reasons such as transformation efficiencies of host cells, a fraction of mutants are lost from the input pool of mutants. As a result the left out mutations is a loss of information in in the analysis of the comprehensive DFE. The resulting data can be used for pairwise analysis of mutants across experimental conditions but due to inherent sampling bias, interpretation of DFE of the gene is compromised.

To solve this issue, in the workflow of *dms2dfe*, preferential enrichments of mutants which are empirically left out, are estimated by inferring them by training a random forest regressor model on preferential enrichments of empirically determined mutants.

Inferring preferential enrichments

Extraction of molecular features

In order to constrain the random forest regressor model to the biological significance of the underlying mechanisms that can potentially guide dynamics in DFEs, molecular features which are potentially relevant to the fitness of mutants are collected through *dms2dfe* feature extraction suite (Figure 3A). They collectively cover following aspects of factors that can determine the fitness of mutants.

1. Activity of protein. In co-culture competition assays, since fitness of mutants is directly linked to activity of genes, features that can roughly reflect changes in the activity of proteins are of utmost importance. Since residues at the active site of protein tend to be more conserved than the rest, distance of the mutated residue from the active site can be used as an indicator of potential changes in activity [8, 18]. Though 'feature extraction' modules of *dms2dfe*, distances of all amino acids are measured from all the atoms of known residues lining active sites and substrates.

2. Structural features Folding of protein is coupled to the activity of protein and is guided by structural constraints such as solvent accessible surface area of amino acids and dihedral angles. Structural features such as secondary structure, solvent accessibility and values of dihedral angles of residues are extracted from user provided PDB structure by utilizing DSSP [33]. Since solvent accessibility is found to be the crucial factor in earlier DMS experiments, additionally a closely related molecular feature - residue depth per residue is estimated by utilizing MSMS [34].

bioRxiv preprint doi: https://doi.org/10.1101/072645; this version posted August 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

3. Physico-chemical properties of residues In terms of molecular dynamics of the proteins, changes due to mutation would affect non-covalent inter-residue interactions. Such interactions would in turn depend on the physico chemical properties of the interacting residues such as solvent accessibility, pI, pKa, charge, hydropathy etc. For this reason physico-chemical properties of the reference and mutated amino acids predicted by chemaxon (http://www.chemaxon.com) and are used to account for the perturbations of local interactions.

4. Conservation score Since conservation scores accounts for the molecular evolutionary history of protein and represents the level of tolerance to mutations. This factor can potentially be of higher relative importance in prediction of fitness of mutations. Through modules of *dms2dfe*'s feature extraction utilities, a multiple sequence alignment (MSA) is generated by a protein BLAST. Conservation scores of the residues of the protein are determined from inverse Shannon entropy of each residue position of protein the MSA. Additionally, using the efficient libraries of Rate4site [35], conservation score of residues of the protein is determined using maximum likelihood distance and Jukes-Cantor distance approaches.

5. Custom features In addition to the above auto-generated features, the custom features which could be potentially relevant to the experiment can be provided through input configuration of a *dms2dfe* run. Relevant custom features such as per residue aggregation propensity can potentially improve the predictive power for an aggregation prone protein.

Regression model

To circumvent the possible normalization issues and sampling biases in the interpretation of DFEs, through *dms2dfe* suite, empirical preferential enrichments are used as training set and preferential enrichments of rest of the mutants is inferred. *dms2dfe* utilizes random forest among machine learning approaches because of its inherent robustness.

Firstly, using all the extracted molecular features, a random forest classifier is trained to predict whether the preferential enrichment value of mutant is greater or lesser that the median of preferential enrichments of all mutants in the training set. This provides ranking of molecular features based on their relative importances, from which top most important features are selected and used along with the empirical preferential enrichments to train a random forest regressor to predict the preferential enrichments of rest of the mutations.

bioRxiv preprint doi: https://doi.org/10.1101/072645; this version posted August 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Although popularly known as a black box method, critical information such as relative importances of the features is helpful in contextualizing the mechanistic insights guiding dynamics of DFEs. Shown in Figure 3B is a mutation matrix of the resultant fitness scores the coverage of mutational space of which is more as compared to that solely from empirical preferential enrichments (Figure 2C). This increase in the information per fitness landscape would assist user in making interpretations with efficient normalization and reduced sampling bias.

Downstream analysis

Comparison of DFEs across experimental conditions

Depending on condition specific selection pressure, the shape and position of DFEs may exhibit a relative dynamics. *dms2dfe* generates visualization of the DFEs control and test conditions along with significance of the dynamics estimated by two tailed Mann–Whitney U test (Figure 3E).

Depending on the relative changes in the fitness of mutants, they are classified in positive, negative and robust categories (Equation 6). The mutants under positive selection pressure gain fitness while those under negative selection pressure lose their fitness as compared to control condition. The mutants which do not gain or lose their fitness are classified as robust. This comparison based on the relative proportion of mutants in each categories can be visualized though *dms2dfe* visualization utilities as shown in Figure 3F. This way, the condition specific nature of selection pressure can be characterized.

Factors guiding dynamics in DFEs

Since molecular features underlying dynamics of DFEs are often coupled [36] and their relation to organismal fitness is a complex, we use ensemble approach implemented through random Forest decision trees to infer preferential enrichments. Along with the predictions of preferential enrichments, this approach also provides relative importances of molecular features used for predictions, which in turn provide a mechanistic insight as well as helps to contextualize the interpretation of DMS experiment. The levels of relative importance of a molecular feature can be used to characterize the nature of respective selection pressure under effect.

Visualizations

dms2dfe generates following visualizations to represent the analysis of DMS data.

1. Distribution of fitness scores i.e. DFE (Figure 3E).

bioRxiv preprint doi: https://doi.org/10.1101/072645; this version posted August 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

2. Heatmaps of mutation matrix of frequencies and fitness scores of mutants (Figure 2A and 3B).
3. Average fitness scores per residue position of the protein projected onto PDB structure. Along with that a PDB file is generated with Average fitness scores incorporated in place of b-factor (Figure 5B).
3. Comparative analysis of DFEs across experimental conditions (Figure 3E and F).
4. Substitution matrix in which each locus represents averaged fitness scores for a particular mutation.

Additionally *dms2dfe* also generates visualizations for assessing the quality of DMS data (Figure 4A and B) and clustering of mutation matrices.

Quality checks for a DMS experiment

Because of the multiple level of complexities involved in the co-culture competition assays, DMS experiments are prone to biological and technical noise which ultimately reflects in the accuracies of the fitness scores. Therefore to gain accurate estimation of fitness of mutants it is crucial to check the quality of DMS data before its interpretation. *dms2dfe* provides a suite of tools to account for following quality checks.

Monitoring sequencing depth

In order to estimate fitness scores with optimum accuracy, sequencing depth is a crucial factor [32]. Depending on the total number of mutants considered for a DMS experiment, sequencing depth determines the frequencies of mutants. *dms2dfe* analyzes the depth of sequencing and the cumulative frequencies along the length of protein which gives a rough estimation of the expected accuracy of fitness scores (Figure 4A).

Monitoring reproducibility across replicates

In addition, given the possibilities of biological or technical the noise technical and biological replicates are used to monitor their levels. *dms2dfe* provides utilities to generate a correlation matrix of such biological or technical replicates (4B) to help user get an estimate of the level of noise in the experimentation.

Performance validation

Across different datasets

To benchmark the efficiency of *dms2dfe* workflow, we used two different datasets of DMS experiments of two different proteins aminoglycoside-2"-phosphotransferase (APH2) [13] and TEM-1 beta-lactamase (TEM1) [14]. As described earlier, the implementation of random forest classifier produced accuracy scores (AUC)

bioRxiv preprint doi: https://doi.org/10.1101/072645; this version posted August 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

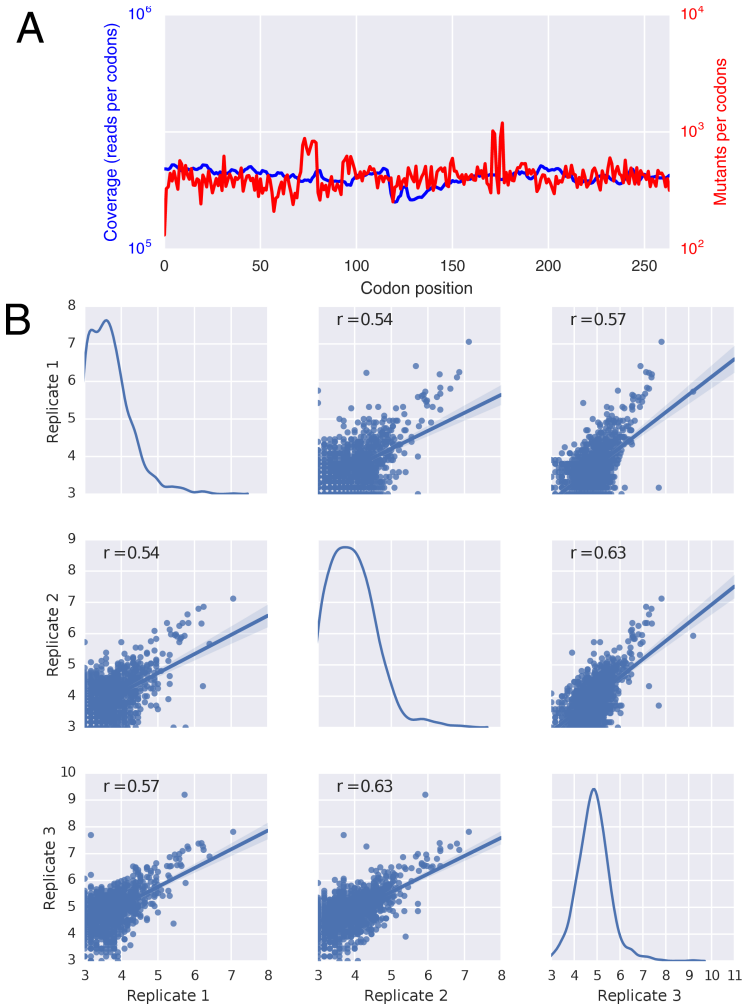

Figure 4: Quality check utilities provided by *dms2dfe*. (A) *dms2dfe* generates a visualization of sequencing depth of the sequencing data along with the cumulative frequencies of mutants at respective positions along the length of the gene. Here, sequencing depth and frequencies of mutants represented in the panel are generated from APH2 dataset. (B) Additionally *dms2dfe* generates a correlation matrix of frequencies of mutants in the replicates (technical or biological) to assess the noise in the input data. Shown in panel (B) are correlations between frequencies (log transformed) of mutants in replicates from the example APH2 dataset. 'Replicate 1', 'Replicate 2' and 'Replicate 3' are representing 3 rounds of selections of first, second and first background libraries respectively at Kanamycin concentration of 1:4 MIC of wild type APH2.

bioRxiv preprint doi: https://doi.org/10.1101/072645; this version posted August 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

of 0.79 and 0.83 for APH2 and TEM1 dataset respectively (Figure 5A and C respectively). This signifies the efficiency of the feature extraction tools and robustness random forest approach of prediction. Additionally as variously reported earlier in alignment with the earlier DMS studies, conservation score, surface accessibility and residue depth are found to the most important features of the models, thus underscoring the biological significance of the predictions from the *dms2dfe* suite. Through *dms2dfe*, the average fitness values per position are projected onto the PDB structure by utilizing visualization modules from UCSF-Chimera [37]. As shown in Figure 5B and D, the depletion of fitness scores for buried residues signifies the efficiency of the inferences of the fitness scores by *dms2dfe*.

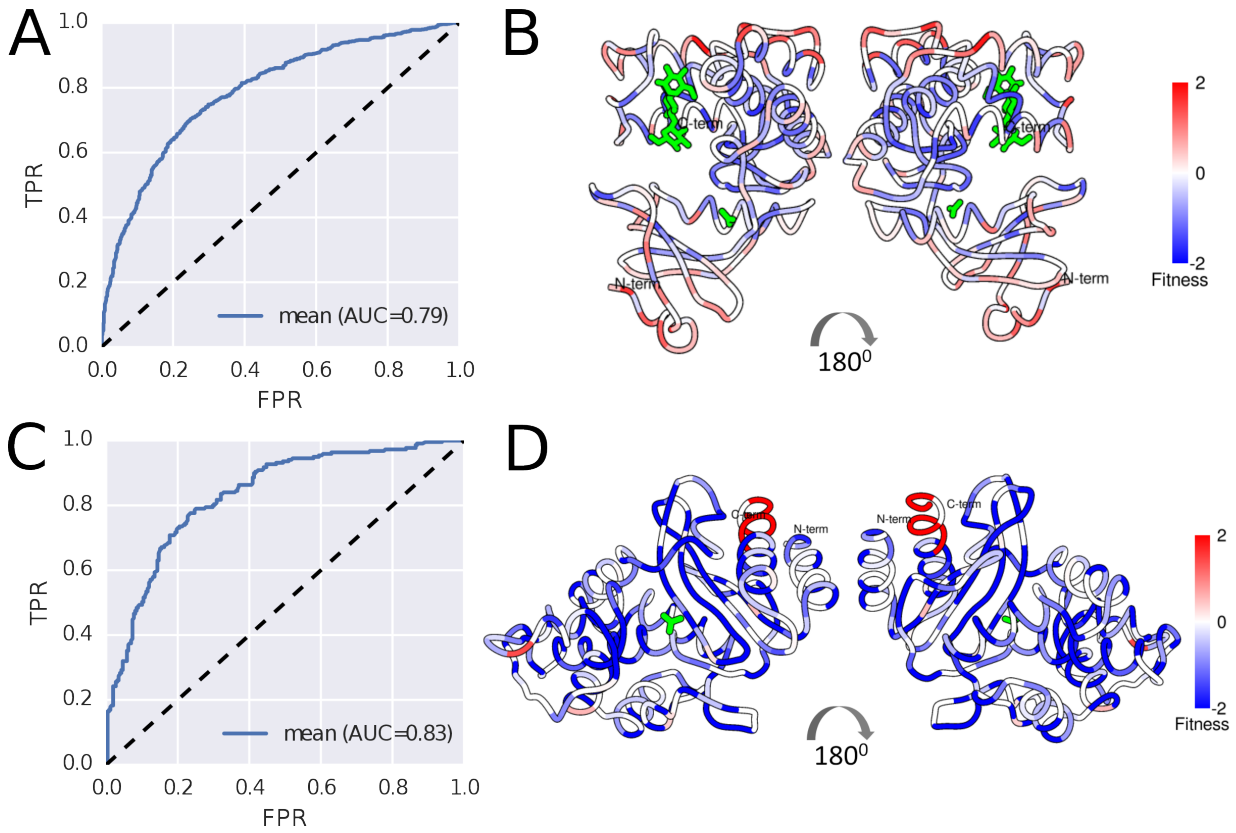

Figure 5: Performance validation across different datasets. ROC curve generated from random forest modeler to predict preferential enrichments of APH2 and TEM1 dataset are shown in panel (A) and (C) respectively. The estimated fitness scores of APH2 and TEM1 datasets are projected onto respective PDB structures as shown in panel (B) and (D) through visualization utilities.

Comparison with *dms_tools*

Next, we compared *dms2dfe*'s inference of fitness scores with that by *dms_tools*. We randomly sampled empirical (true) preferential enrichments of the two datasets (APH2 and TEM1) and tested the extent of conformity of inferred fitnesses with randomly sampled empirical preferential enrichments. As shown in Figure 6A and B, the pearson correlation coefficients of the inferred fitness values of APH2 dataset (F_i

bioRxiv preprint doi: https://doi.org/10.1101/072645; this version posted August 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

in case of *dms2dfe* and $\Phi_{r,x}$ in case of *dms_tools*) versus true values are found to be higher for *dms2dfe*'s predictions than that by *dms_tools* across 5 random samplings of 100 mutants each. In case of TEM1 dataset with one exception all other randomly sampled mutations are better correlated to fitness inferred by *dms2dfe* as compared to *dms_tools*. As a result, these comparisons indicate that for independent example datasets, the random forest approach of *dms2dfe* performs better than Markov chain Monte Carlo (MCMC) approach used in *dms_tools*.

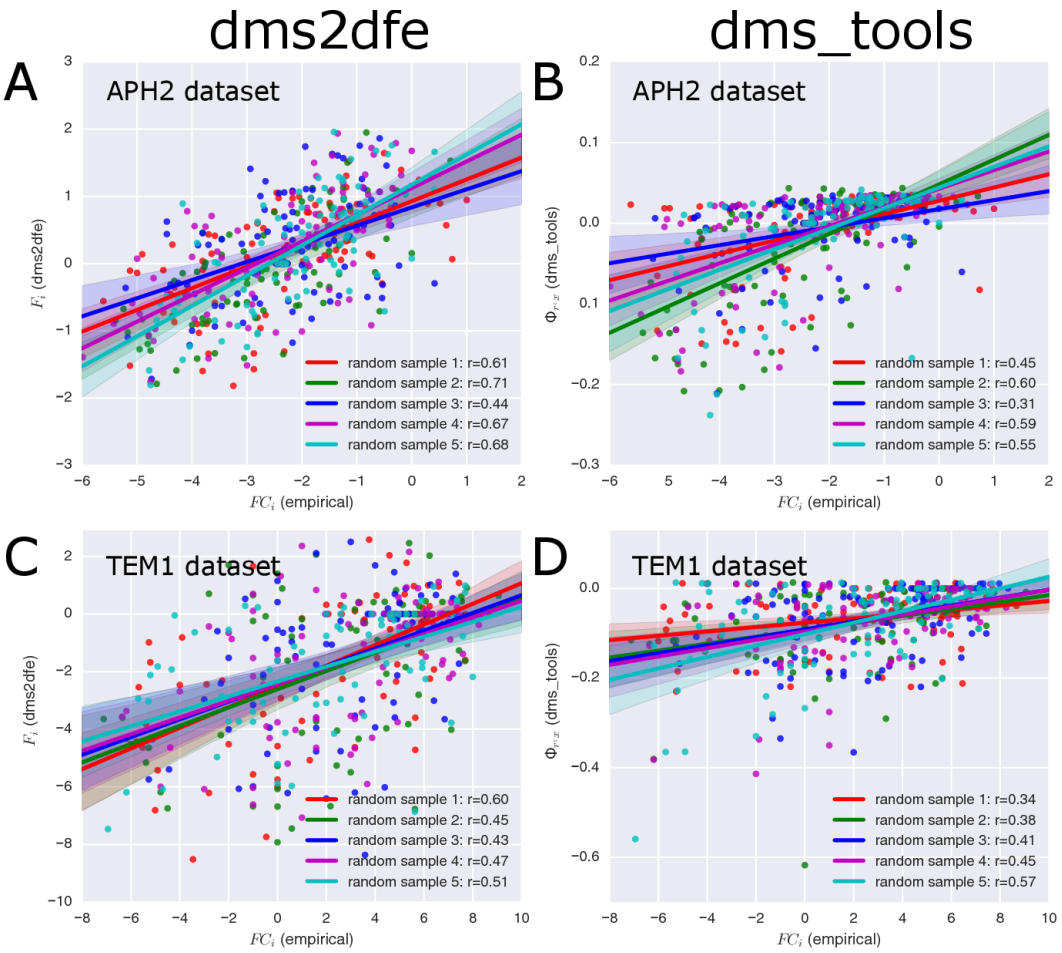

Figure 6: Comparison of *dms2dfe* with *dms_tools*. Correlations of fitness scores of randomly sampled mutants inferred using *dms2dfe* (panel (A) and (C)) show better correlation with empirical preferential enrichments than that by *dms_tools* (panel (B) and (D)). First row i.e. panel (A) and (B) represents data from APH2 dataset while second row i.e. panel (C) and (D) represents data from TEM1 dataset.

Discussion

The broad appeal of DMS method is evident by the wide range of applicability [38] inclusive of questions relevant to structural modeling [4], protein stability [3], substrate specificity [13, 16, 18], protein-protein interactions [6, 7, 10, 11, 19, 20], mutational effects on organismal fitness [8, 12, 21], environmental effects

bioRxiv preprint doi: https://doi.org/10.1101/072645; this version posted August 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[22], biology of viruses [5, 15, 17], fitness landscapes of antibiotic resistance [14] and oncogenes [9]. As the usage of DMS reaches even wider scientific community, there is a need to make analysis utilities available in a user friendly format with easily interpretable data formats and visualizations. Scalability of *dms2dfe* workflow renders it readily applicable to this need. With adaptability to wide range of modifications and developments in DMS experiments, *dms2dfe* can potentially serve as a robust alternative to existing tools designed for analysis of DMS data.

dms2dfe allows assessment of quality of the DMS data by accounting for biological and technical noises and sequencing depth per position of gene. It also provides solution to normalization issues and sampling biases in interpreting relative dynamics of DFEs by inferring preferential enrichments of mutations. *dms2dfe* workflow also characterizes the nature of experimental condition specific selection pressure and provides insights into mechanistic insights into the factors guiding the dynamics in DFEs in terms of relative importances of molecular features.

Biological significance of the random forest modeler is evident by the higher accuracy scores for both the independent datasets used. The better accuracy of the predictor (AUC score) for unrelated datasets indicates that the molecular feature extracted by the feature extraction tools of *dms2dfe* efficiently capture the biological significance of the fitness effects. Additionally, better correlation between random sampling of empirical preferential enrichments and inferred fitness scores shows that *dms2dfe* approach offers better performance than *dms_tools*. Collectively, the validation of *dms2dfe* workflow underscores the efficient use of robust random forest approach and a definite advance in the methods used for analysis of DMS data.

dms2dfe through its comprehensively documented python programming library (API) (available at http://kc-lab.github.io/dms2dfe) would allow further developments and improvements to synchronize with advances and modifications in DMS experiments. In conclusion, a direct link from sequencing outputs to the analysis of DFEs provided by *dms2dfe* suite enables reproducibility and scalability which would assist user in design, optimization and execution of DMS experiments.

bioRxiv preprint doi: https://doi.org/10.1101/072645; this version posted August 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Materials and methods

Example datasets

For generating visualization for the variant calling, estimation of preferential enrichment, inference of fitnesses, subsequent comparison of DFEs and quality assessments of sequencing data as shown in Figures 2,3 and 4, sequencing data from Deep Mutational Scanning of aminoglycoside-2"-phosphotransferase (APH2)[13] was used. The sequencing data for the input pool of mutants (KKA2_Bkg1) and pools selected in Kanamycin selection at concentration of 1:4 (KKA2_S1_Kan14_L1) and 1:1 (KKA2_S1_Kan11_L1) MIC of wild type protein are obtained from SRA (SRR1292901, SRR1292881 and SRR1292709 respectively).

In order to benchmark the efficiency of *dms2dfe* (Figure 5), we used mutational data of TEM-1 beta-lactamase (TEM1) [14] in addition to APH2. In case of APH2 dataset, we used data for pool of mutants selected in the Kanamycin concentration of 1:4 the MIC of wild type protein as selected pool and background pool as the input library (Figure 5 panel A and B). While in case of TEM1 dataset, we used a subset of data in which pool of mutants was selected for resistance against 256 μg/ml Ampicilin and we regard the sample selected at minimal concentration of Ampicilin (0.5 μg/ml) as input library (Figure 5 panel C and D).

Estimation and normalization of preferential enrichments

Preferential enrichments are estimated as a log fold change of frequencies of the mutants in selected and input pool as follows.

$$FC_i = log_2\left(\frac{N_{i,selected}}{N_{i,input}}\right) \tag{1}$$

where FC_i is the fold change of frequencies of i^{th} mutant in selected pool ($N_{i,selected}$) and input pool ($N_{i,input}$).

Following two approaches used to normalize empirical preferential enrichments in the implementation of *dms2dfe*.

In the first approach, preferential enrichments are Z-score normalized employing population mean and standard deviation of distribution of preferential enrichments of synonymous mutations.

bioRxiv preprint doi: https://doi.org/10.1101/072645; this version posted August 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

$$FC_{i,normalized} = \frac{FC_i - \mu_{syn}}{\sigma_{syn}} \tag{2}$$

where $FC_{i,normalized}$ is the normalized preferential enrichment of i^{th} mutant. μ_{syn} and σ_{syn} are the population mean and standard deviation respectively of Gaussian fitted distribution of preferential enrichments of synonymous mutants.

Alternatively, through *dms2dfe*, normalization can also be carried out with respect to sequencing depth of wild type sequences as described by Melnikov et al. [13].

$$FC_{i,normalized} = FC_i - FC_{i,wild,depth} \tag{3}$$

where $FC_{i,normalized}$ and FC_i are normalized, and non-normalized preferential enrichments of i^{th} mutant respectively. $FC_{i,wild,depth}$ fold change of sequencing depth of wild type alleles.

Rescaling of preferential enrichments with respect to synonymous mutations

In order to interpret fitness of mutants relative to the wild type allele, preferential enrichments are rescaled with respect to preferential enrichments of the wild type allele at that position.

$$F_i = FC_i - FC_{i,wild} \tag{4}$$

where F_i and FC_i are fitness score and preferential enrichment of i^{th} mutant respectively. $FC_{i,wild}$ is the fold change of wild type sequences at that position.

Classification of mutants based on fitness scores

Since the fitness scores are scaled based on wild type, relatively, the mutants can be classified into beneficial and deleterious categories as follows,

$$M_i \in \begin{cases} beneficial & if\ F_i > 0 \\ neutral & if\ F_i = 0 \\ deleterious & if\ F_i < 0 \end{cases} \tag{5}$$

where, M_i is i^{th} kind of survived mutant and F_i is its respective fitness score.

bioRxiv preprint doi: https://doi.org/10.1101/072645; this version posted August 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Classification of mutants based on comparison of DFEs

For comparison of DFEs between experimental conditions, mutants can be classified based on their relative fitnesses into following 3 categories.

$$M_{i,test} \in \begin{cases} positive & if\ (M_{i,control} \in deleterious\ and\ M_{i,test} \in beneficial) \\ negative & if\ (M_{i,control} \in beneficial\ and\ M_{i,test} \in deleterious) \\ robust & else \end{cases} \quad (6)$$

where $M_{i,test}$ and $M_{i,control}$ are i^{th} kind of survived mutant in test and control conditions respectively.

Software availability

dms2dfe is open source and published under the GNU General Public License. The python package is freely available at https://github.com/kc-lab/dms2dfe. Accompanying documentation with installation and usage information along with examples is available at https://kc-lab.github.io/dms2dfe .

Abbreviations

dms2dfe : Deep Mutational Scanning to Distribution of Fitness Effects;

DMS : Deep Mutational Scanning;

DFE : Distribution of Fitness Effects;

MSA : Multiple Sequence Alignment;

APH2 : aminoglycoside-2"-phosphotransferase;

TEM1 : TEM-1 beta-lactamase.

SRA : Sequence Read Archive

Additional files

SI_01: Fitness scores of APH2 dataset.

SI_02: Fitness scores of TEM1 dataset.

SI_03: Empirical preferential enrichments (FC_i), inferred fitness scores by *dms2dfe* (F_i) and by *dms_tools* ($\Phi_{r,x}$) of randomly sampled mutations from APH2 and TEM1 datasets.

bioRxiv preprint doi: https://doi.org/10.1101/072645; this version posted August 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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.

bioRxiv preprint doi: https://doi.org/10.1101/072645; this version posted August 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

References

1. Fowler DM, Araya CL, Fleishman SJ, Kellogg EH, Stephany JJ, Baker D, Fields S: **High-resolution mapping of protein sequence-function relationships.** *Nature methods* 2010, **7**:741–6.

2. Fowler DM, Fields S: **Deep mutational scanning: a new style of protein science.** *Nature methods* 2014, **11**:801–7.

3. 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**. *Proceedings of the National Academy of Sciences* 2012, **109**:16858–16863.

4. Adkar BV, Tripathi A, Sahoo A, Bajaj K, Goswami D, Chakrabarti P, Swarnkar MK, Gokhale RS, Varadarajan R: **Protein model discrimination using mutational sensitivity derived from deep sequencing**. *Structure* 2012, **20**:371–381.

5. Whitehead TA, Chevalier A, Song Y, Dreyfus C, Fleishman SJ, De Mattos C, Myers CA, Kamisetty H, Blair P, Wilson I a, Baker D: **Optimization of affinity, specificity and function of designed influenza inhibitors using deep sequencing.** *Nature biotechnology* 2012, **30**:543–8.

6. Fujino Y, Fujita R, Wada K, Fujishige K, Kanamori T, Hunt L, Shimizu Y, Ueda T: **Robust in vitro affinity maturation strategy based on interface-focused high-throughput mutational scanning**. *Biochemical and Biophysical Research Communications* 2012, **428**:395–400.

7. McLaughlin RN, Poelwijk FJ, Raman A, Gosal WS, Ranganathan R: **The spatial architecture of protein function and adaptation.** *Nature* 2012, **491**:138–42.

8. Roscoe BP, Thayer KM, Zeldovich KB, Fushman D, Bolon DNA: **Analyses of the effects of all ubiquitin point mutants on yeast growth rate**. *Journal of Molecular Biology* 2013, **425**:1363–1377.

9. Starita LM, Pruneda JN, Lo RS, Fowler DM, Kim HJ, Hiatt JB, Shendure J, Brzovic PS, Fields S, Klevit RE: **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* 2013, **110**:E1263–72.

10. Forsyth CM, Juan V, Akamatsu Y, DuBridge RB, Doan M, Ivanov AV, Zhiyuan M, Polakoff D, Razo J, Wilson K, Powers DB: **Deep mutational scanning of an antibody against epidermal growth factor receptor using mammalian cell display and massively parallel pyrosequencing**. *mAbs*

bioRxiv preprint doi: https://doi.org/10.1101/072645; this version posted August 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

2013, **5**:523–532.

11. Moretti R, Fleishman SJ, Agius R, Torchala M, Bates PA, Kastritis PL, Rodrigues JPGLM, Trellet M, Bonvin AMJJ, Cui M, Rooman M, Gillis D, Dehouck Y, Moal I, Romero-Durana M, Perez-Cano L, Pallara C, Jimenez B, Fernandez-Recio J, Flores S, Pacella M, Praneeth Kilambi K, Gray JJ, Popov P, Grudinin S, Esquivel-Rodr??guez J, Kihara D, Zhao N, Korkin D, Zhu X, et al.: **Community-wide evaluation of methods for predicting the effect of mutations on protein-protein interactions**. *Proteins: Structure, Function and Bioinformatics* 2013, **81**:1980–1987.

12. 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 (New York, NY)* 2013, **19**:1537–51.

13. 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 Research* 2014, **42**:1–8.

14. Firnberg E, Labonte JW, Gray JJ, Ostermeier M: **A comprehensive, high-resolution map of a Gene's fitness landscape**. *Molecular Biology and Evolution* 2014, **31**:1581–1592.

15. Bloom JD: **An experimentally determined evolutionary model dramatically improves phylogenetic fit**. *Molecular Biology and Evolution* 2014, **31**:1956–1978.

16. Thyme SB, Song Y, Brunette TJ, Szeto MD, Kusak L, Bradley P, Baker D: **Massively parallel determination and modeling of endonuclease substrate specificity**. *Nucleic Acids Research* 2014, **42**:13839–13852.

17. Thyagarajan B, Bloom JD: **The inherent mutational tolerance and antigenic evolvability of influenza hemagglutinin**. *eLife* 2014, **2014**.

18. Stiffler MA, Hekstra DR, Ranganathan R: **Evolvability as a Function of Purifying Selection in TEM-1 ??-Lactamase**. *Cell* 2015, **160**:882–892.

19. Van Blarcom T, Rossi A, Foletti D, Sundar P, Pitts S, Bee C, Melton Witt J, Melton Z, Hasa-Moreno A, Shaughnessy L, Telman D, Zhao L, Cheung WL, Berka J, Zhai W, Strop P, Chaparro-Riggers J, Shelton DL, Pons J, Rajpal A: **Precise and efficient antibody epitope determination through library design, yeast display and next-generation sequencing**. *Journal of Molecular Biology* 2015, **427**:1513–1534.

bioRxiv preprint doi: https://doi.org/10.1101/072645; this version posted August 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

20. Doolan KM, Colby DW: **Conformation-dependent epitopes recognized by prion protein antibodies probed using mutational scanning and deep sequencing**. *Journal of Molecular Biology* 2015, **427**:328–340.

21. Mishra P, Flynn JM, Starr TN, Bolon DN: **Systematic Mutant Analyses Elucidate General and Client-Specific Aspects of Hsp90 Function**. *Cell Reports* 2016, **15**:588–598.

22. Mavor D, Barlow K, Thompson S, Barad BA, Bonny AR, Cario CL, Gaskins G, Liu Z, Deming L, Axen SD, Caceres E, Chen W, Cuesta A, Gate RE, Green EM, Hulce KR, Ji W, Kenner LR, Mensa B, Morinishi LS, Moss SM, Mravic M, Muir RK, Niekamp S, Nnadi CI, Palovcak E, Poss EM, Ross TD, Salcedo EC, See SK, et al.: **Determination of ubiquitin fitness landscapes under different chemical stresses in a classroom setting**. *eLife* 2016, **5**:1–23.

23. Katsonis P, Lichtarge O: **A formal perturbation equation between genotype and phenotype determines the Evolutionary Action of protein-coding variations on fitness**. *Genome Research* 2014, **24**:2050–2058.

24. Kimura M: **Model of effectively neutral mutations in which selective constraint is incorporated.** *Proceedings of the National Academy of Sciences of the United States of America* 1979, **76**:3440–3444.

25. Ohta T: **The Nearly Neutral Theory Of Molecular Evolution**. *Annual Review of Ecology and Systematics* 1992, **23**:263–286.

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**:3430–3431.

27. Bloom JD: **Software for the analysis and visualization of deep mutational scanning data**. *BMC Bioinformatics* 2015, **16**:1–13.

28. Bolger AM, Lohse M, Usadel B: **Trimmomatic: A flexible trimmer for Illumina sequence data**. *Bioinformatics* 2014, **30**:2114–2120.

29. Langmead B, Salzberg SL: **Fast gapped-read alignment with Bowtie 2**. *Nat Methods* 2012, **9**:357–359.

30. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R: **The Sequence Alignment/Map format and SAMtools**. *Bioinformatics* 2009, **25**:2078–2079.

bioRxiv preprint doi: https://doi.org/10.1101/072645; this version posted August 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

31. Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP: **Sequencing depth and coverage: key considerations in genomic analyses.** *Nature reviews Genetics* 2014, **15**:121–32.

32. Matuszewski S, Hildebrandt ME, Ghenu A-H, Jensen JD, Bank C: **A Statistical Guide to the Design of Deep Mutational Scanning Experiments.** *Genetics* 2016, **XXX**(September):1–23.

33. Kabsch W, Sander C: **Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features**. *Biopolymers* 1983, **22**:2577–2637.

34. Sanner MF, Olson a J, Spehner JC: **Reduced surface: an efficient way to compute molecular surfaces.** *Biopolymers* 1996, **38**:305–320.

35. Pupko T, Bell RE, Mayrose I, Glaser F, Ben-Tal N: **Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues.** *Bioinformatics (Oxford, England)* 2002, **18 Suppl 1**:S71–S77.

36. Manhart M, Morozov AV: **Protein folding and binding can emerge as evolutionary spandrels through structural coupling.** *Proceedings of the National Academy of Sciences of the United States of America* 2015, **112**:1797–1802.

37. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE: **UCSF Chimera–a visualization system for exploratory research and analysis.** *Journal of computational chemistry* 2004, **25**:1605–12.

38. Shendure J, Fields S: **Massively Parallel Genetics**. *Genetics* 2016, **203**(June):617–619.