dms2dfe: Comprehensive Workflow for Analysis of Deep Mutational Scanning Data

Rohan Dandage[1] and Kausik Chakraborty[1]

[1]CSIR-Institute of Genomics and Integrative Biology, New Delhi, India.

04 July 2017

Summary

`dms2dfe` is an integrative analysis workflow designed for end-to-end analysis of Deep Mutational Scanning (Fowler, Stephany, and Fields 2014) data. Using this workflow, users can implement various processing methods and downstream applications for pair-wise enrichment analysis of ultra-deep sequencing data.

Recently, owing to the evolution of sequencing and phenotyping technologies, large scale genotype to phenotype data is increasingly being generated. Along this line of research, Deep Mutational Scanning method allows comprehensive assessment of all possible amino acid substitutions of an entire gene or part of a gene. In the analysis of Deep Mutational Scanning data, `dms2dfe` addresses crucial issue of noise control using widely used DESeq2 (Love, Huber, and Anders 2014) workflow and offers variety of downstream analyses to contextualize results. In downstream analyses, `dms2dfe` workflow provides identification of potential molecular constraints, comparative analysis across different experimental conditions and generation of data-rich visualizations (Dandage and Chakraborty 2016). While a number of tools have been developed for analysis of DMS data (Fowler et al. 2011; Bloom 2015; Rubin et al. 2017), users familiar with commonly used state-of-art genomics tools such as Trimmomatic (Bolger, Lohse, and Usadel 2014), Bowtie (Langmead and Salzberg 2012), samtools (Li et al. 2009) and DESeq2 (Love, Huber, and Anders 2014) can opt for `dms2dfe` workflow for analysis of preferential enrichments. Note that `dms2dfe` workflow is designed exclusively for experimental designs in which there is a need of pair-wise analysis of samples eg. before and after selection.

As an input for the workflow, deep sequencing data (whether unaligned or aligned) or list of genotypic variants can be provided. For a demonstration purpose, sample datasets from various studies (Firnberg et al. 2014; C. A. Olson,

Wu, and Sun 2014; Melnikov et al. 2014) are available here.[1] `dms2dfe` uses DataFrames from robust Pandas library (McKinney 2010) for processing all the tabular data. For enabling downstream analyses, structural features are extracted from user-provided PDB file (Kabsch and Sander 1983; Sanner, Olson, and Spehner 1996) and conservation scores are obtained from multiple sequence alignments (Sievers and Higgins 2014; Pupko et al. 2002). As an optional step, visualizations of preferential enrichments onto PDB structure are generated using UCSF Chimera (Pettersen et al. 2004).

Source code and issue tracker is available in `dms2dfe`'s GitHub repository.[2] Documentation and API[3] are generated using Sphinx.[4]

References

Bloom, Jesse D. 2015. "Software for the analysis and visualization of deep mutational scanning data." *BMC Bioinformatics* 16 (1): 1–13. doi:10.1186/s12859-015-0590-4.

Bolger, Anthony M, Marc Lohse, and Bjoern Usadel. 2014. "Trimmomatic: A flexible trimmer for Illumina sequence data." *Bioinformatics* 30 (15): 2114–20. doi:10.1093/bioinformatics/btu170.

Dandage, Rohan, and Kausik Chakraborty. 2016. "dms2dfe : Comprehensive Workflow for Analysis of Deep Mutational Scanning Data." *BioRxiv*, 072645. doi:10.1101/072645.

Firnberg, Elad, Jason W Labonte, Jeffrey J Gray, and Marc Ostermeier. 2014. "A comprehensive, high-resolution map of a Gene's fitness landscape." *Molecular Biology and Evolution* 31 (6): 1581–92. doi:10.1093/molbev/msu081.

Fowler, Douglas M, Carlos L Araya, Wayne Gerard, and Stanley Fields. 2011. "Enrich: Software for analysis of protein function by enrichment and depletion of variants." *Bioinformatics* 27 (24): 3430–1. doi:10.1093/bioinformatics/btr577.

Fowler, Douglas M, Jason J Stephany, and Stanley Fields. 2014. "Measuring the activity of protein variants on a large scale using deep mutational scanning." *Nature Protocols* 9 (9). Nature Publishing Group: 2267–84. doi:10.1038/nprot.2014.153.

Kabsch, Wolfgang, and Christian Sander. 1983. "Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features." *Biopolymers* 22 (12). Wiley Online Library: 2577–2637.

Langmead, Ben, and Steven L Salzberg. 2012. "Fast gapped-read alignment

[1] https://github.com/rraadd88/ms_datasets
[2] https://github.com/kc-lab/dms2dfe
[3] https://kc-lab.github.io/dms2dfe
[4] http://www.sphinx-doc.org

with Bowtie 2." *Nat Methods* 9 (4): 357–59. doi:10.1038/nmeth.1923.

Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. 2009. "The Sequence Alignment/Map format and SAMtools." *Bioinformatics* 25 (16): 2078–9. doi:10.1093/bioinformatics/btp352.

Love, Michael I, Wolfgang Huber, and Simon Anders. 2014. "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2." *Genome Biology* 15 (12): 550. doi:10.1186/s13059-014-0550-8.

McKinney, Wes. 2010. "Data Structures for Statistical Computing in Python." In *Proceedings of the 9th Python in Science Conference*, edited by Stéfan van der Walt and Jarrod Millman, 51–56.

Melnikov, Alexandre, Peter Rogov, Li Wang, Andreas Gnirke, and Tarjei S. Mikkelsen. 2014. "Comprehensive mutational scanning of a kinase in vivo reveals substrate-dependent fitness landscapes." *Nucleic Acids Research* 42 (14). doi:10.1093/nar/gku511.

Olson, C Anders, Nicholas C Wu, and Ren Sun. 2014. "A comprehensive biophysical description of pairwise epistasis throughout an entire protein domain." *Current Biology* 24 (22). Elsevier Ltd: 2643–51. doi:10.1016/j.cub.2014.09.072.

Pettersen, Eric F, Thomas D Goddard, Conrad C Huang, Gregory S Couch, Daniel M Greenblatt, Elaine C Meng, and Thomas E Ferrin. 2004. "UCSF Chimera—a Visualization System for Exploratory Research and Analysis." *Journal of Computational Chemistry* 25 (13). Wiley Online Library: 1605–12.

Pupko, Tal, Rachel E Bell, Itay Mayrose, Fabian Glaser, and Nir Ben-Tal. 2002. "Rate4Site: An Algorithmic Tool for the Identification of Functional Regions in Proteins by Surface Mapping of Evolutionary Determinants Within Their Homologues." *Bioinformatics* 18 (suppl_1). Oxford University Press: S71–S77.

Rubin, Alan F, Hannah Gelman, Nathan Lucas, Sandra M Bajjalieh, Anthony T Papenfuss, Terence P Speed, and Douglas M Fowler. 2017. "A statistical framework for analyzing deep mutational scanning data." *Genome Biology* 18 (1). Genome Biology: 150. doi:10.1186/s13059-017-1272-5.

Sanner, Michel F, Arthur J Olson, and Jean-Claude Spehner. 1996. "Reduced Surface: An Efficient Way to Compute Molecular Surfaces." *Biopolymers* 38 (3). Wiley Online Library: 305–20.

Sievers, Fabian, and Desmond G Higgins. 2014. "Clustal Omega, Accurate Alignment of Very Large Numbers of Sequences." *Multiple Sequence Alignment Methods*. Springer, 105–16.