Systematic Analysis of Variant Effect Maps

BCB330 Proposal & Literature Review

by

Bilin Nong

Supervisors: Jochen Weile & Fritz Roth

University of Toronto
June 5, 2023

1 Introduction

Linking phenotype to genotype is a complex problem. Due to this complexity, the interpretation of clinical variants is a difficult process. To classify variant effects, the American College of Medical Genetics and Genomics (ACMG) guidelines[1] provides the categories "pathogenic", "likely pathogenic", "benign", "likely benign", and "variants of uncertain significance" (VUS) to describe variants found in genes causing Mendelian disorders. As a result of lack of evidence, the majority of clinical variants are currently classified as "variants of uncertain significance" (VUS), meaning there is insufficient evidence to conclude a relationship between genetic variants and the disease. Moreover, since numerous strategies such as computational predictors and laboratory assays are used to interrogate variant effects, the ACMG guideline assigns different evidence strength of variant classification to *in silico* methods and laboratory studies.

On one hand, *in silico* methods such as PolyPhen-2 (Polymorphism Phenotyping v2)[2], SIFT (Sorting Intolerant From Tolerant)[3], and PROVEAN (PROtein Variation Effect ANalyzer)[4] can be applied at genome scale to predict variant effects. But the ACMG guidelines consider these *in silico* algorithms to be only supporting evidence strength in classifying pathogenic (code PP3) and benign (code BP4) variants. Thus, computational predictors alone are insufficient for clinical interpretation of variant effects. To justify this choice, the ACMG guidelines[1] cite underlying problems of computational evidence: First, different *in silico* algorithms may rely on similar data to support predictions. Second, the majority of algorithms have not been tested against known pathogenic variants. Additionally, the prediction power of algorithms for various genes can vary greatly. However, the ACMG guidelines for variant classification have not yet

caught up with the development of newer approaches such as VARITY[5] and ESM1-v[6], which have been shown to be much more reliable[7]. Therefore, these classification standards might be in need of revision.

On the other hand, laboratory assays such as functional complementation[8] and Y2H (Yeast-2-Hybrid)[9] are more trusted, since they tend to have relatively high sensitivity that outperform the computational predictors: Y2H assays are able to detect around two out of three disease variants associated with protein interactions[10]; the yeast complementation assays reached a recall rate over 60%[11] at 90% precision threshold in terms of identifying known pathogenic mutations. Additionally, assays in human cell lines with fluorescence-based readouts are used increasingly to analyze variant effect, such as the study conducted by Matreyek et al.[12], in which the authors measured the effects of PTEN variants on protein abundance. The ACMG guidelines classify such *in vitro/in vivo* functional studies to be strong evidence for pathogenic (code PS3) or benign (code BS3) variant effects. However, laboratory assays until recently did not scale easily and were thus generally used re-actively, i.e., to test variants after they were encountered in patients. In addition, only a small portion of human genes can be accommodated in these laboratory assays.

1.1 MAVE: Multiplex Assay of Variant Effect

To tackle the problem of scaling laboratory assays, a proactive approach called Multiplexed Assays of Variant Effect (MAVEs)[13] was developed. MAVEs harness high-throughput sequencing to apply laboratory assays at scale and often also integrate machine learning methods and clinical expertise.

MAVE studies typically include four main components: mutagenesis, selection of variants via assay, sequencing, and computational analysis. In the mutagenesis step, a commonly used method is POPCode mutagenesis(Precision Oligo-Pool based Code Alteration), which aims to yield a complete spectrum of possible amino acid changes in all protein positions[14]. Following mutagenesis, the major component of a MAVE is the selection step, enriching or depleting variants based on their effects on protein functionality. There are many selection schemes that can be applied in a MAVE, such as functional complementation, Y2H assays, and FACS (fluorescence-activated cell sorting). Sequencing follows the selection step, aiming to count the enrichment or depletion of variants as a result of selection. There are different sequencing approaches including Tileseq[1] and Barseq[2] can be used in this step. The last step in MAVE is computational analysis, where scripts are employed to analyze the sequencing readout and calculate the selection advantage for each variant.

Nowadays, MAVEs harvest lots of variant effect maps targeting clinically relevant genes for clinical research. However, there are some issues related to MAVEs. First, going from MAVEs to clinical interpretation is not straightforward, since the selection advantage for each variant may not reflect their pathogenicity. For this reason, including the Log-likelihood Ratio (LLR) approach in the downstream analysis of MAVEs is necessary, since this approach can transform the fitness scores into the tendency towards pathogenicity of variants. Second, computational analysis pipelines have undergone many iterations of developments. Different versions of MAVEs may adopt different implementations, leading to varying outcomes. Therefore, it is important to analyze different versions of MAVEs and evaluate their performance systematically.

[1] https://github.com/rothlab/tileseqMave
[2] https://github.com/rothlab/pacybara

2 Goals and Objectives

Based on the existing issues with MAVEs, this BCB330 project aims to re-evaluate the performance of variant effect maps based on different versions of MAVE pipelines with respect to precision and sensitivity on reliable benchmarks. The detailed goals include:

1. Re-process the raw data underlying existing variant effect maps with the latest versions of their respective analysis pipelines.
    (a) Inspect the QC outputs for the maps to identify potential quality issues.
2. Compile benchmark sets of variants with known pathogenicity from online databases and literature for each map.
    (a) Explore alternative reference sets of non-disease genes.
3. Compare the predictions made by different versions of variant effect maps using the benchmark sets and use them to infer evidence strength for clinical interpretation.
    (a) Identify disagreeing variant effect outputs, and establish their computational provenance.
    (b) Produce Precision-Recall Curves to evaluate the performance of updated version and old version of MAVEs.
    (c) Calculate Log-likelihood Ratio transformations and identify the fitness score intervals that corresponding to different evidence levels towards "pathogenic" and "benign" classifications.
4. Provide recommendations for optimizing the implementation of MAVEs based on the evaluation result.

3 Approaches

3.1 Reprocessing Maps

To reprocess the raw data sets using various versions of MAVEs, we use the TileSeq pipeline. These pipelines take the experimental results of a MAVE, calculate fitness scores from sequencing reads, and generate diagnostic Quality-Control plots.

There are two main components in the TileSeq workflow: TileseqMut and TileSeqMave. To begin, TileSeqMut aligns the sequencing reads to the reference sequence and uses a Bayesian method to call the most likely true variants. Following this, TileSeqMave, the second component of the TileSeq pipeline, analyzes the variant calling result obtained from the previous steps. It starts by translating variant calls into amino acid changes at the protein-level. Next, it calculates the enrichment ratios, $log(\phi)$, i.e. the log-ratio between pre- and post-selection frequencies for each given variant and performs error-modeling and quality filtering. Lastly, the "scaleScore" step re-scales the $\log(\phi)$ scores relative to pivot points which represent typical nonsense and synonymous variants. As a result, the final scores are distributed such that 0 represents full loss of function, while 1 represents wildtype-like fitness. The pipeline also includes two QC steps, which generates a series of diagnostic plots.

3.2 Compiling Benchmark Sets

To generate reference sets of variants with known pathogenicity, we use a script which offers automated generation of benchmark sets tailored to specific disease-causing genes, drawing from reliable sources such as ClinVar[3] and gnomAD[4] controls. Since ClinVar tends to contain more pathogenic variants than benign, gnomAD is used for supple-

[3] https://www.ncbi.nlm.nih.gov/clinvar/
[4] https://gnomad.broadinstitute.org

menting the benign variants collect at population level. Moreover, it provides options that allows users to define specific criteria, including allele frequency threshold, quality, and trait of interest. These features allows for the refinement of reference sets, which enhances their applicability in the downstream analysis, includes drawing precision-recall curve, and calculating the log-likelihood ratio for pathogenicity. At the same time, there exists some limitations when using gnomAD and ClinVar as reference sources. First, there can be variations in the reliability of ClinVar submissions, since ClinVar collects submissons of interpretation with varying standards of provenance, and some submissions might lack detailed evidence. Second, variants in gnomAD controls can only serve as a proxy-benign set, as they have not been officially classified. Additionally, both ClinVar and gnomAD focus on variants related to diseases, the validation of maps for non-disease genes will require manual curation of reference variants.

3.3 Evaluation approaches

3.3.1 Precision-Recall Curve

There are multiple methods available to assess classification performance, including the Precision-Recall Curve (PRC), Receiver-operator characteristic (ROC)[15], Matthew's correlation coefficient (MCC)[16], and F-scores[17]. However, when evaluating different maps, the PRC stands out as the most suitable approach due to its ability to handle class imbalance, a common occurrence in maps where reference set sizes vary. To further alleviate potential biases, a prior-balancing approach[5] is used to compensate for differences in reference set sizes.

To compare the predictions made by the old and updated versions of MAVE, the

precision-recall curve (PRC) serves as a straightforward and informative visualization tool. Precision is defined as the fraction of true positive calls out of all positive calls, or in the context of variant effects, the proportion of correctly predicted pathogenic variants out of all predicted pathogenic variants. On the other hand, recall represents the fraction of true positive calls out of all actual cases, or in the context of variant effects, the fraction of variants correctly identified as pathogenic among all existing pathogenic variants.[18] Utilizing the PRC offers numerous advantages in evaluating the performance of these prediction maps. First, since high precision is crucial in the context of clinical decision based on prediction, we can compare the recall level of these maps under the threshold of 90% precision: the precision-recall curve provides a numerical values (REC90) describing this information. Second, an overall assessment can be made by analyzing the total area under the curve (AUPRC), which provides a summary measurement of the precision-recall curve. For instance, Figure 1 illustrates the precision-recall curve of Calmodulin.

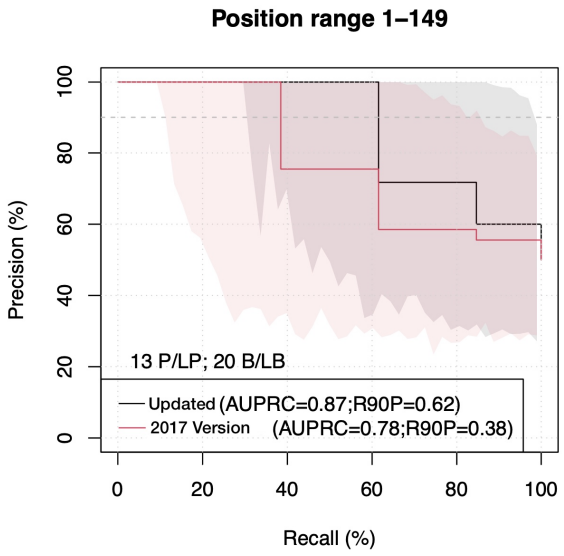

Figure 1: The PRC for the old and updated version of Calmodulin Map. The updated version of Calmodulin map (shown in black) outperforms the original version published in 2017[14] (shown in red).

3.3.2 Log-likelihood Ratio of Pathogenicity

In the context of pathogenicity assessment, the log-likelihood ratio (LLR) for pathogenicity[19] can be used to evaluate the likelihood of a variant being pathogenic versus benign based on the data. While we can get the fitness scores of the variants from the maps, these scores only represent the effect of variants on protein function, and they do not necessarily reflect onto the pathogenicity (i.e. how likely they will cause diseases)[20]. This makes it potentially difficult to use MAVEs for clinical variant interpretation. LLRs can be used to address this problem. We can first estimate the respective probability densities across the scores of presumed pathogenic and benign reference variants. Subsequently, these densities are used to calculate the LLR of pathogenicity, expressing how much more likely a variant at a fitness score is to be pathogenic than it is benign[20].

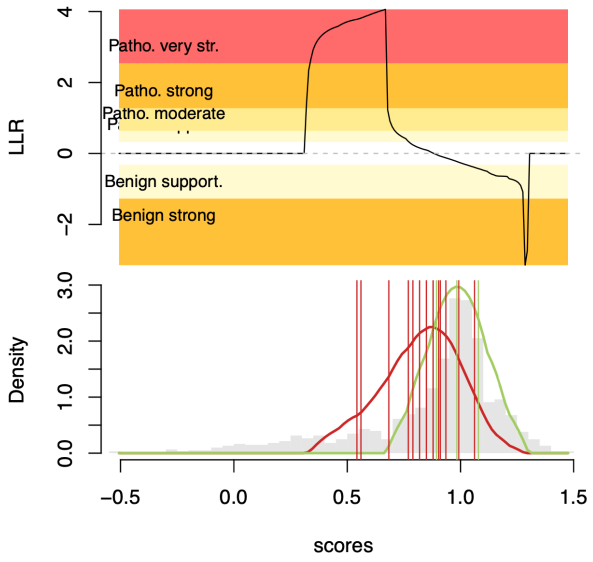

Figure 2: A two-panel plot mapping variant fitness score to log-likelihood ratio to pathogenicity of Calmodulin. The bottom panel represents the distribution of known pathogenic (red) and benign (green) benchmark variant sets. The top panel illustrates the estimated log-likelihood ratio as a function of variant scores.

An LLR plot for Calmodulin is demonstrated in Figure 2. Notably, in a fitness scores range between 0.3 and 0.8, peaking at a score of around 0.669, the LLR function

is positive – indicating a propensity towards pathogenicity. Whereas in the fitness range between 0.9 and 1.4, peaking at 1.286, the LLR function is negative, indicating a propensity to benignity. It is worth noting the absence of high LLRs for scores near 0, which may initially seem contrary to expectations. However, this can be explained by the dominant-negative inheritance pattern observed for Calmodulin.

4 Potential Problems and Their Solutions

4.1 Deviation Between New Scores and Old Scores

Sometimes we might encounter a situation where the new and old scores significantly disagree for certain variants. For example, it is possible to observe instances where the fitness score of the new map approaches 1 while the fitness score of the old map tends to be close to 0, or vice versa. To solve this problem, several approaches can be used. One potential solution involves tuning the scaling pivots based on the $log(\phi)$ distribution. Alternatively, another option is to adjust the filter value to control over the level of inclusion (or exclusion) of variants.

4.2 Reference Sets for non-disease genes

Though there are scripts for compelling benchmark sets for disease-causing genes, we do not have straight-forward references for non disease-causing genes. One solution may be to compile benchmark sets for non-disease genes manually. An alternative solution is to compare the scores of these genes with other computational predictors, such as VARITY[5] and ESM1-v[6]. This comparative analysis have potential to serve as a proxy reference for non-disease genes, offering valuable information for bench-marking and evaluation purposes.

References

[1] Sue Richards, Nazneen Aziz, Sherri Bale, David Bick, Soma Das, Julie Gastier-Foster, Wayne W. Grody, Madhuri Hegde, Elaine Lyon, Elaine Spector, Karl Voelkerding, and Heidi L. Rehm. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genetics in Medicine*, 17(5):405–424, May 2015.

[2] Ivan Adzhubei, Daniel M. Jordan, and Shamil R. Sunyaev. Predicting Functional Effect of Human Missense Mutations Using PolyPhen-2. *Current Protocols in Human Genetics*, 76(1), January 2013.

[3] P. C. Ng. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Research*, 31(13):3812–3814, July 2003.

[4] Linnea Sandell and Nathaniel P Sharp. Fitness Effects of Mutations: An Assessment of PROVEAN Predictions Using Mutation Accumulation Data. *Genome Biology and Evolution*, 14(1):evac004, January 2022.

[5] Yingzhou Wu, Hanqing Liu, Roujia Li, Song Sun, Jochen Weile, and Frederick P. Roth. Improved pathogenicity prediction for rare human missense variants. *The American Journal of Human Genetics*, 108(10):1891–1906, October 2021.

[6] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a lan-

guage model. *Science*, 379(6637):1123–1130, 2023. Earlier versions as preprint: bioRxiv 2022.07.20.500902.

[7] Vikas Pejaver, Alicia B. Byrne, Bing-Jian Feng, Kymberleigh A. Pagel, Sean D. Mooney, Rachel Karchin, Anne O'Donnell-Luria, Steven M. Harrison, Sean V. Tavtigian, Marc S. Greenblatt, Leslie G. Biesecker, Predrag Radivojac, Steven E. Brenner, Leslie G. Biesecker, Steven M. Harrison, Ahmad A. Tayoun, Jonathan S. Berg, Steven E. Brenner, Garry R. Cutting, Sian Ellard, Marc S. Greenblatt, Peter Kang, Izabela Karbassi, Rachel Karchin, Jessica Mester, Anne O'Donnell-Luria, Tina Pesaran, Sharon E. Plon, Heidi L. Rehm, Natasha T. Strande, Sean V. Tavtigian, and Scott Topper. Calibration of computational tools for missense variant pathogenicity classification and ClinGen recommendations for PP3/BP4 criteria. *The American Journal of Human Genetics*, 109(12):2163–2177, December 2022.

[8] Eva Trevisson, Alberto Burlina, Mara Doimo, Vanessa Pertegato, Alberto Casarin, Luca Cesaro, Placido Navas, Giuseppe Basso, Geppo Sartori, and Leonardo Salviati. Functional Complementation in Yeast Allows Molecular Characterization of Missense Argininosuccinate Lyase Mutations. *Journal of Biological Chemistry*, 284(42):28926–28934, October 2009.

[9] Stanley Fields and Ok-kyu Song. A novel genetic system to detect protein–protein interactions. *Nature*, 340(6230):245–246, July 1989.

[10] Nidhi Sahni, Song Yi, Mikko Taipale, Juan I. Fuxman Bass, Jasmin Coulombe-Huntington, Fan Yang, Jian Peng, Jochen Weile, Georgios I. Karras, Yang Wang, István A. Kovács, Atanas Kamburov, Irina Krykbaeva, Mandy H. Lam, George Tucker, Vikram Khurana, Amitabh Sharma, Yang-Yu Liu, Nozomu Yachie, Quan Zhong, Yun Shen, Alexandre Palagi, Adriana San-Miguel, Changyu Fan, Dawit

Balcha, Amelie Dricot, Daniel M. Jordan, Jennifer M. Walsh, Akash A. Shah, Xinping Yang, Ani K. Stoyanova, Alex Leighton, Michael A. Calderwood, Yves Jacob, Michael E. Cusick, Kourosh Salehi-Ashtiani, Luke J. Whitesell, Shamil Sunyaev, Bonnie Berger, Albert-László Barabási, Benoit Charloteaux, David E. Hill, Tong Hao, Frederick P. Roth, Yu Xia, Albertha J.M. Walhout, Susan Lindquist, and Marc Vidal. Widespread Macromolecular Interaction Perturbations in Human Genetic Disorders. *Cell*, 161(3):647–660, April 2015.

[11] Song Sun, Fan Yang, Guihong Tan, Michael Costanzo, Rose Oughtred, Jodi Hirschman, Chandra L. Theesfeld, Pritpal Bansal, Nidhi Sahni, Song Yi, Analyn Yu, Tanya Tyagi, Cathy Tie, David E. Hill, Marc Vidal, Brenda J. Andrews, Charles Boone, Kara Dolinski, and Frederick P. Roth. An extended set of yeast-based functional assays accurately identifies human disease mutations. *Genome Research*, 26(5):670–680, May 2016.

[12] Kenneth A. Matreyek, Jason J. Stephany, Ethan Ahler, and Douglas M. Fowler. Integrating thousands of PTEN variant activity and abundance measurements reveals variant subgroups and new dominant negatives in cancers. *Genome Medicine*, 13(1):165, December 2021.

[13] Lea M. Starita, Muhtadi M. Islam, Tapahsama Banerjee, Aleksandra I. Adamovich, Justin Gullingsrud, Stanley Fields, Jay Shendure, and Jeffrey D. Parvin. A multiplexed homology-directed dna repair assay reveals the impact of 1,700 brca1 variants on protein function. *bioRxiv*, 2018.

[14] Jochen Weile, Song Sun, Atina G Cote, Jennifer Knapp, Marta Verby, Joseph C Mellor, Yingzhou Wu, Carles Pons, Cassandra Wong, Natascha Lieshout, Fan Yang, Murat Tasan, Guihong Tan, Shan Yang, Douglas M Fowler, Robert Nuss-

baum, Jesse D Bloom, Marc Vidal, David E Hill, Patrick Aloy, and Frederick P Roth. A framework for exhaustively mapping functional missense variants. *Molecular Systems Biology*, 13(12):957, December 2017.

[15] Kelly H. Zou, A. James O'Malley, and Laura Mauri. Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models. *Circulation*, 115(5):654–657, 2007.

[16] Davide Chicco and Giuseppe Jurman. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics*, 21:1–13, 2020.

[17] Steven A. Hicks, Inga Strümke, Vajira Thambawita, Malek Hammou, Michael A. Riegler, Pål Halvorsen, and Sravanthi Parasa. On evaluation metrics for medical applications of artificial intelligence. *Scientific Reports*, 12(1):5979, April 2022.

[18] Kai Ming Ting. *Precision and Recall*, pages 781–781. Springer US, Boston, MA, 2010.

[19] Jochen Weile, Nishka Kishore, Song Sun, Ranim Maaieh, Marta Verby, Roujia Li, Iosifina Fotiadou, Julia Kitaygorodsky, Yingzhou Wu, Alexander Holenstein, Céline Bürer, Linnea Blomgren, Shan Yang, Robert Nussbaum, Rima Rozen, David Watkins, Marinella Gebbia, Viktor Kozich, Michael Garton, D. Sean Froese, and Frederick P. Roth. Shifting landscapes of human MTHFR missense-variant effects. *The American Journal of Human Genetics*, 108(7):1283–1300, July 2021.

[20] Brendan Floyd, Jochen Weile, Prince Kannankeril, Andrew Glazer, Chloe Reuter, Calum MacRae, Euan Ashley, Dan Roden, Frederick Roth, and Victoria Parikh.

Proactive Variant Effect Mapping to Accelerate Genetic Diagnosis for Pediatric Cardiac Arrest. preprint, MEDICINE & PHARMACOLOGY, February 2023.