Re-analysis of Variant Effect Maps

BCB330 Final Report

by

Bilin Nong

Supervisors: Jochen Weile & Fritz Roth

University of Toronto
August 21, 2023

1 Introduction

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 categories ranging from "pathogenic" to "benign", with uncertain cases labeled as "variants of uncertain significance" (VUS), where insufficient evidence exists. The majority of clinical variants are currently classified as the latter[2]. The ACMG guidelines assign different evidence strengths to different types of information, such as *in silico* predictors or laboratory studies. Unfortunately, while computational predictions are easily scalable, they are only considered "supporting" evidence, whereas the stronger evidence provided by laboratory assays traditionally only exists at small scale.

1.1 MAVE: Multiplex Assays of Variant Effect

To tackle the problem of scaling laboratory assays, a proactive approach called Multiplexed Assays of Variant Effect (MAVEs)[3] 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 steps: mutagenesis, selection of variants via assay, sequencing, and computational analysis. A good example for a mutagenesis method is Precision Oligo-Pool based Code Alteration(POPCode) mutagenesis, which aims to yield a complete spectrum of possible amino acid changes across the protein [4]. Following mutagenesis, the resulting variant libraries are subjected to a 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, Yeast-2-Hybrid (Y2H) assays, or sorting based on fluorescent reporter activity via FACS (fluorescence-activated cell

sorting). Sequencing is then used to quantify the enrichment or depletion of variants as a result of selection. There are different sequencing approaches including Tileseq[1] and Barseq[2] which can be used in this step. The last step in MAVE is computational analysis, where pipelines and scripts are employed to analyze the sequencing readout and calculate the selection advantage for each variant.

Nowadays, MAVEs generate lots of variant effect maps of 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 in a given assay may not reflect their pathogenicity. For this reason, including a Log-likelihood Ratio (LLR) approach in the downstream analysis of MAVEs has been proposed. This approach can transform the fitness scores into a metric of evidence strength towards or against 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.

1.2 Goals and Objectives

Based on the above 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.

1. Re-process the raw data underlying existing variant effect maps with the latest versions of their respective analysis pipelines.

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

(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.

2 Methods

2.1 Reprocessing Maps

To re-process the raw data sets using various versions of MAVEs, we use the TileSeqPro pipeline, which takes the raw sequencing output from a Tileseq MAVE experiment, applies quality filters and error corrections, calculates fitness scores, and generates diagnostic Quality-Control plots.

TileSeqPro improves upon an older "Legacy" TileSeq pipeline that was employed in the Roth Lab until recently. There are many difference of implementations between TileSeqPro and the Legacy computational pipelines. Most importantly, the Legacy pipeline collapsed the equivalent codon changes early into amino acid changes, before filtering. While this meant that lower quality data could get past the filters, and error correction could not be performed at nucleotide levels, it had the advantage of boosting the number of reads supporting each data point. In contrast, TileSeqPro calculates functional scores for individual codon changes separately, filters out low quality variants based on a number of different criteria such as the low read count (below a certain threshold), and finally combined them into amino acid changes. Thus, TileSeqPro sacrifices data to avoid systematic error, but as a result may potentially suffer from more noise.

2.2 Evaluation approaches

2.2.1 Precision-Recall Curve

There are multiple methods available to assess classification performance, including the Precision-Recall Curve (PRC), Receiver-operator characteristic (ROC)[5], Matthew's correlation coefficient (MCC)[6], and F-scores[7]. 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[8] 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.[9]

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.

2.2.2 Compiling Benchmark Sets

To evaluate variant effect maps via Precision-Recall curves, we require adequate reference sets of know pathogenic and benign benchmark variants. To generate such reference sets, we use a script that is built into TileseqPro, which offers automated generation of benchmark sets tailored to specific disease-causing genes, drawing from reliable sources such as ClinVar[10] and gnomAD[11] controls. Since ClinVar tends to contain more pathogenic variants than benign, gnomAD is used for supplementing the benign variants collect at population level. The script 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 finding transformation functions from fitness scores to 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. Finally, the validation of maps for non-disease genes will require alternative approaches, such as comparison against high-quality computational predictors.

2.2.3 Correlation against computational predictors

To compare the performance of TileseqPro and the Legacy Tileseq pipelines from another perspective, we did the moving window correlation analysis between Tileseq scores versus VARITY[8] scores along amino acid positions. VARITY is a computational method for pathogenicity prediction, it provides a probability of pathogenicity scores by harnessing gradiant boosted trees algorithm to weight input training sets, where more close to 1 for variants that are inferred to be pathogenic, and more close to 0 for variants that are inferred to be benign.

There are two VARITY models generated by the VARITY framework: VARITY_R and VARITY_ER. The VARITY_R model included rare ClinVar[10] variants with global minor allele frequency (MAF) less than 0.5% in its core training/ test set; while VARITY_ER model only included extremely rare ClinVar variants with MAF less than 10^{-6} in its core set. The score scale of VARITY is completely opposite to which of the fitness score produced by Tileseq pipelines, therefore, we measured the performance of fitness scores by comparing their anti-correlation with VARITY scores.

2.2.4 Log-likelihood Ratio of Pathogenicity

In the context of pathogenicity assessment, the log-likelihood ratio (LLR) for pathogenicity[12] 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 pathogenicity (i.e. how likely they will cause diseases)[13]. We first estimate the probability densities across the scores of pathogenic and benign reference variants, respectively, via kernel density estimation. The log ratio between the density functions is then 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[13]. However, this method is very sensitive to even small differences in metaparameter choice, particularly kernel bandwidth and requires careful manual supervision.

3 Results

3.1 SUMO1

We first re-calculated the map for SUMO1, one of the first variant effect maps created in the Roth Lab[4].

3.1.1 Observations from QC results

Examining the distributions of synonymous and nonsense variant enrichment ratios ($\log(\phi)$) relative to marginal frequency (see supplemental Fig. 1) reveals an acceptable separation between the two at a frequency of 10^{-4}. Given the sequencing depth of 1.5M reads per sample, this suggests an ideal cutoff of 150 reads. After setting filters accordingly, approximately one in three variants were filtered out by the frequency filter

and bottleneck filters.

The library shows an extrapolated average number of 3.09 amino acid changes per clone (see supplemental Fig. 2) Accordingly, the overall coverage appears decent, despite a small band with reduced coverage near amino acid position 28.

After filter application, there is a clear separation between the enrichment values of the nonsense and synonymous variants (Fig 1), indicating a reliable selection assay. The missense variants enrichment values show a bimodal distribution, two modes are located at around -1.474 and -0.3, which are slightly smaller than the nonsense and synonymous modes, which are around -1.212 and -0.025 respectively. In the case of the upper mode, this could indicate that most non-syononymous SUMO1 variants have at least a small fitness effect. For the lower mode however, it is unlikely that many variants are more deleterious than nonsense, so the shift may instead be an artifact of the harsher filtering approach.

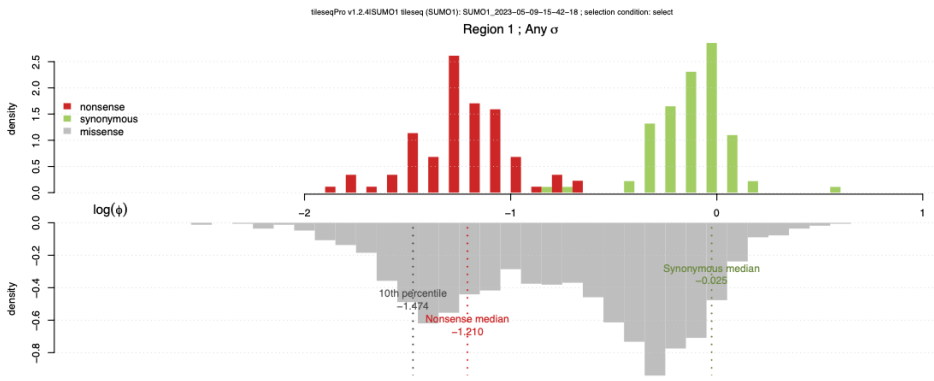
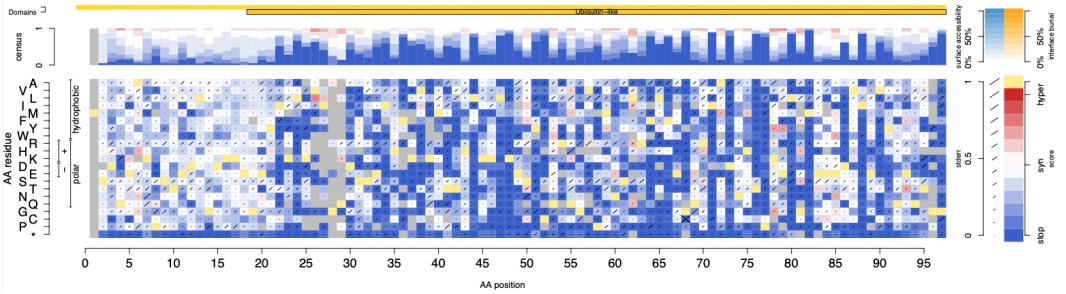


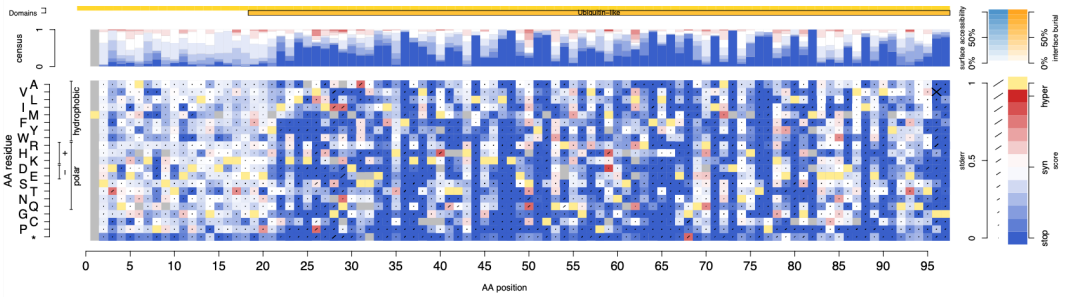
Figure 1: The enrichment ratio distributions of SUMO1 shows the overall distribution of missense(grey), synonymous(green), and nonsense(red) variants' enrichment values in region 1 of SUMO1 gene.

3.1.2 Comparison between TileSeqPro and the Legacy pipeline

A comparison of the maps created by TileSeqPro and the Legacy pipeline as visualized via Mavevis[14] is shown in Figure 2. The first twenty amino acid are insensitive to missense mutaions in both version of variant effect maps as expected, since this region of SUMO1 is intrinsically disordered[15]. The Legacy version of the map shows more available amino acid changes compared to our new version, especially near amino acid position 28, due to differences in filtering.



(a) heatmap for SUMO1 2023 scores generated by Mavevis



(b) heatmap for SUMO1 2019 scores generated by Mavevis

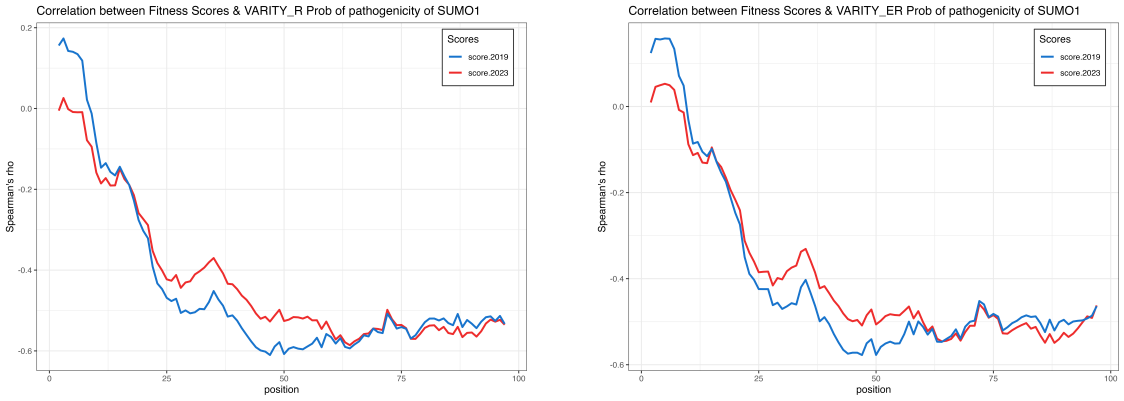
Figure 2: A comparison of new and old version of SUMO1 heatmaps. The x-axis represents the amino acid position in protein, and the y-axis includes all possible amino acid changes. Colors used in the heatmap stand for the fitness score, where blue describes the variants which are as deleterious as full deletion; white represents synonymous-like variants, red stands for increased fitness than the wildtype residue at the given position; and yellow represents the wildtype amino acid.[14]

Missense and synonymous mutations, were assigned larger estimates of standard error by TileseqPro compared to Legacy (see supplemental Fig. 3). The reason for this change might be due to the introduction of bootstrapping and more pessimistic error regularization methods.

Overall, the correlation between the fitness scores produced by the two pipelines is high with Spearman's $\rho = 0.95$ (see supplemental Fig. 4). Interestingly, a moving window analysis of the correlation also indicates slightly less agreement between two in the N-terminal disordered region.

3.1.3 Moving Window Correlation between Fitness Scores and VARITY Scores

Since SUMO1 is not a disease gene, no reference set for precision-recall analysis in terms of pathogenicity exists. As an alternative evaluation approach, we analysed the correlation of fitness scores with the computational predictor VARITY[8]. As expected, there is substantial anti-correlation between the VARITY probability of pathogenicity and fitness scores (they are scaled oppositely). Disregarding the disordered region (the first 20 amino acids), from position 25 to 70, the Legacy fitness scores are more anti-correlated with VARITY_R and VARITY_ER scores; whereas from position 70 to the end, the new fitness scores have better anti-correlation (Fig. 3). The overall value of ρ in these regions fluctuates around approximately -0.5.



(a) Correlation between VARITY_R Scores and Fitness Scores

(b) Correlation between VARITY_ER Scores and Fitness Scores

Figure 3: Moving window correlation between VARITY Scores and Fitness Scores. Blue line represents the old scores, while red line represents the updated scores.

3.1.4 Conclusion

Both version of SUMO1 maps performed similarly in terms of correlation between VARITY scores. Our updated TileseqPro pipelines used more information on generating the error estimates, whereas the harsher filtering steps in the TileseqPro pipelines caused measurement error we revealed in the variant enrichment value distribution. Additionally, amino acid changes with low marginal frequencies were filtered out in the updated map while they presented in the previous version of SUMO1 map.

3.2 CALM1

We then re-processed the map for CALM1, which was first created by the Roth Lab in 2018 and first updated by 2019[13].

3.2.1 Observations from QC results

The library shows an extrapolated average number of 2 amino acid changes per clone (see supplemental Fig. 5). The overall coverage appears low, with many low-frequency variants (white in the figure) or outright missing ones (gray in the figure).

Examining the distribution of synonymous and nonsense variant enrichment ratios ($\log(\phi)$) relative to marginal frequency (see supplemental Fig. 6 (a)) reveals that separation between $\log(\phi)$ improves noticeably at a frequency of $10^{-3.85}$. Given the sequencing depth of 1.4M reads per sample, we chose a minimum read count threshold of 200. After setting filters accordingly, around two third of variants were filtered out by frequency filter and bottleneck filters (see supplemental Fig. 6 (b)).

Following these filter settings, there is a clear separation between the enrichment values of the nonsense and synonymous variants (Fig. 4), with modes located at -1.230 and 0.001 respectively, whereas the missense variants have only one mode near 0.001

with an extended shoulder towards the low end. This agrees with previous observations that most missense variants of CALM1 are not very damaging in terms of fitness[4].

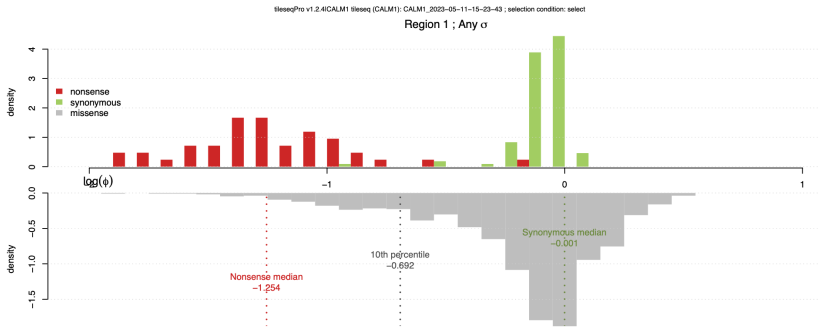
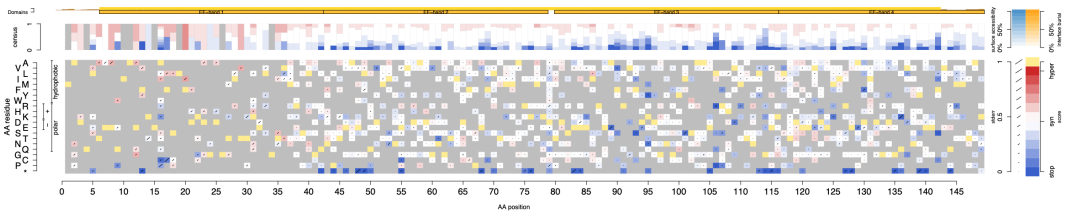
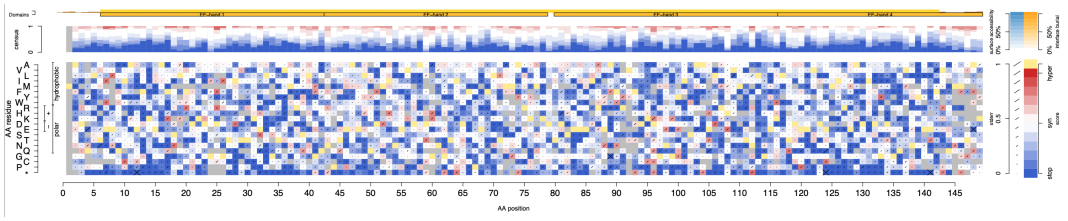


Figure 4: The enrichment ratio distributions of CALM1 shows the overall distribution of missense(grey), synonymous(green), and nonsense(red) variants' enrichment values in region 1 of CALM1 gene.

A comparison between the maps calculated via TileSeqPro and the Legacy pipeline is shown in Figure 5. The impact of harsh filtering in the new map is clearly visible, as almost 2/3 of variants were removed. This updated version CALM1 map could be a perfect candidate for machine learning imputation.



(a) heatmap for CALM1 2023 scores generated by Mavevis



(b) heatmap for CALM1 2019 scores generated by Mavevis

Figure 5: A comparison of new and old version of CALM1 heatmaps generated by Mavevis.

3.2.2 Precision-recall Curve and Log-likelihood Ratio of Pathogenicity

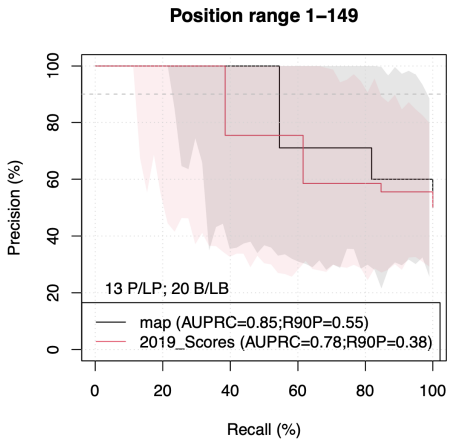
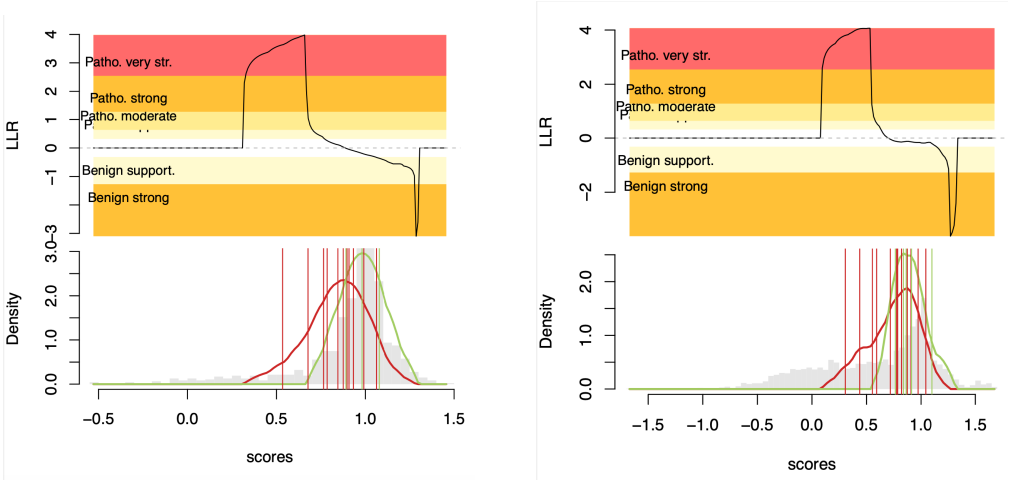


Figure 6: Precision-recall Curve for the 2023 and 2019 version of CALM1 maps. The black line represents the curve for 2023 map, and the red line represents the curve for 2019 map. Reference sets this PRC is a combined reference of CALM1, CALM2, and CALM3 (three identical copies of Calmodulin gene) from ClinVar and gnomAD.

Figure 6 shows the precision-recall curve (PRC) for both versions of the CALM1 map. Our updated TileseqPro version of CALM1 map outperformed the Legacy version, with area under PRC 0.85 compared to 0.78, and a recall of 55% at 90% precision compared to 38% for the Legacy map.

The transformations to log-likelihood ratio of pathogenicity for both CALM1 maps are compared in Figure 7. The overall shape of these LLR curves are similar, despite some small differences in fitness score ranges where LLR is positive. For the updated LLR, when the fitness scores range from 0.3 to 0.8, the LLR function is positive, indicates a tendency towards pathogenicity; while in the old version LLR, LLR function is positive when the fitness scores are between 0.2 and 0.7. We could not infer LLRs for scores near 0, due to the absence of reference variants in those regions. However the concentration of negative reference variants in the intermediate fitness range might be explained by the dominant-negative inheritance pattern for Calmodulin[13]. One potential problem with the transformation function is the negative LLR spikes at fit-

ness score above 1, as there is no supporting reference data. The spike is an artifact of the kernel density estimates and will need to be addressed in future iterations of the software.



(a) LLR for 2023 CALM1 map

(b) LLR for 2019 CALM1 map

Figure 7: The Log-likelihood ratio of pathogenicity for the CALM1 as a function of fitness score.

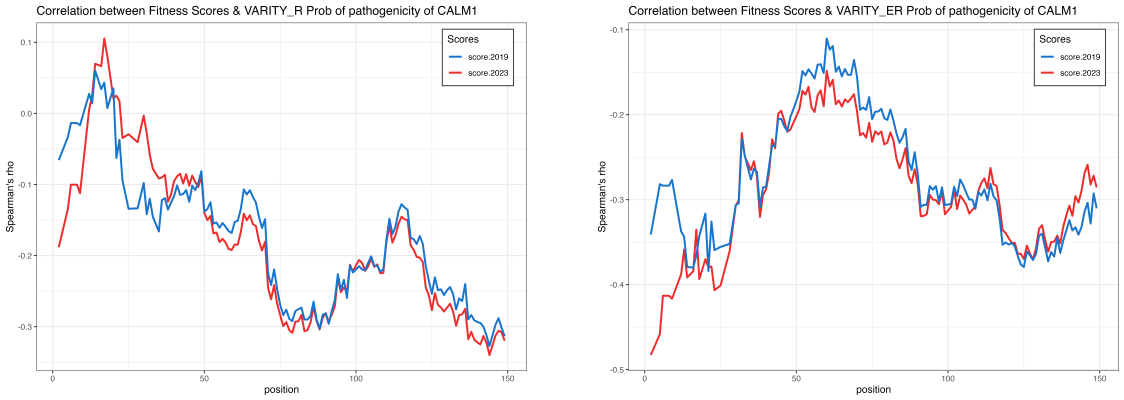
3.2.3 Comparison between TileSeqPro and the Legacy pipeline

Similar to the standard error of SUMO1 fitness scores, the standard error of CALM1 fitness scores generated by TileseqPro pipelines have larger standard error than the Legacy version (see supplemental Fig. 7). We also observed that the standard error of the nonsense variants' fitness scores is the highest amongst others.

The fitness score from the two pipeline versions agrees with each other with a correlation of $\rho = 0.92$ (see supplemental Fig. 8). Nonsense variants have lower correlation ($\rho = 0.59$), compared to missense and synonymous variants ($\rho = 0.91$), likely because nonsense variants have fewer of available amino acid change data and larger standard error. The moving window correlation graph indicates a good correlation from amino acid position 25 to the end, while the correlation for the first 20 amino acid are low because few amino acid changes in that region "survived" after the harsh filtering.

3.2.4 Moving Window Correlation between Fitness Scores and VARITY Scores

We analyzed the moving window correlation of VARITY scores and fitness scores (Fig. 8). As expected, there is an anti-correlation between these scores. Surprisingly however, the trend of the correlation between fitness scores vs. VARITY_R and VARITY_ER are not similar. The Anti-correlation between fitness scores and VARITY_R are better than with VARITY_ER. Besides, The anti-correlation between TileseqPro score versus VARITY scores and Legacy scores vs. VARITY scores are similar, despite the region around the first twenty amino acids.



(a) Correlation between VARITY_R Scores and Fitness Scores

(b) Correlation between VARITY_ER Scores and Fitness Scores

Figure 8: Moving window correlation between VARITY Scores and Fitness Scores. Blue line represents the old scores, and red line represents the updated scores.

3.2.5 Conclusion

Though we threw out lots of codon changes with low marginal frequencies when re-calculating the map with TileseqPro, the newer version of CALM1 map performed better in terms of gaining more precision. There is still a high correlation between both versions of CALM1 fitness scores, and the anti-correlation between VARITY scores and both version of fitness scores look similar.

3.3 MTHFR

We also re-processed maps for MTHFR[12]. The original maps were measured in two different genetic backgrounds (WT and A222V) and at four different concentrations of folinic acid (12 μg/ml (f12AV), 25 μg/ml (f25AV), 100 μg/ml (f100AV) to 200 μg/ml (f200AV)). Here we completed reprocessing of the four maps in the A222V background. The MTHFR variant effect maps currently deposited on MaveDB were first calculated in 2019, and last updated by 2020.

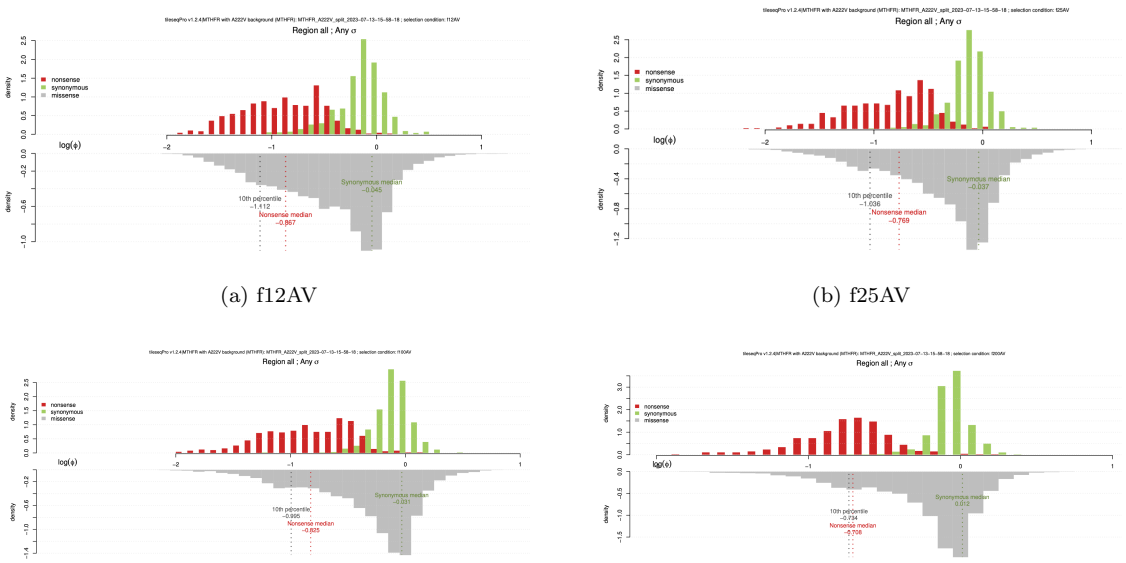
3.3.1 Observations from QC results

The MTHFR map was subdivided into four separate mutagenesis regions, which show an extrapolated average number of 0.941, 0.769, 0.864, and 0.791 amino acid changes per clone, respectively (see supplemental Fig. 10). Accordingly, the overall coverage presented by the coverage heatmap (see supplemental Fig. 9) appears low, with lots of low-frequency variants, especially in Tiles 10, 18, and 19.

Inspecting the distribution of nonsense and synonymous variant enrichment ratios ($\log(\phi)$) relative to marginal frequency thresholds at different folate concentrations (see supplemental Fig. 11), we observed that the separation between enrichment values improved at a frequency of $10^{-4.6}$. Given the sequencing depth of 2M reads per sample, we chose a count threshold at 50 corresponding to this observation. After setting filters using this setting, around half of variants were filtered out by the frequency and bottleneck filter.

Following the filtering steps, there is a relatively clear separation between the enrichment values of nonsense and synonymous MTHFR variants (Fig. 9) at four folate concentrations, with the 200 $\mu g/ml$ folinate condition showing the best separation. The

mode of nonsense variants is located around $\log(\phi) = -0.8$, and the mode for synonymous variants is located around -0.03 for region all. However, the modes vary a lot by regions: $\log(\phi) = -0.85$ for nonsense variants and $\log(\phi) = -0.023$ for synonymous variants in Region 1; $\log(\phi) = -1.17$ for nonsense variants and $\log(\phi) = 0.084$ for synonymous variants in Region 2; $\log(\phi) = -1.2$ for nonsense variants and $\log(\phi) = -0.23$ for synonymous variants in Region 3; and $\log(\phi) = -0.542$ and $\log(\phi) = -0.04$ for nonsense and synonymous variants in Region 4, respectively. The missense variants have only one mode near 0, indicating most missense variants of MTHFR have relatively mild fitness effects.



(a) f12AV

(b) f25AV

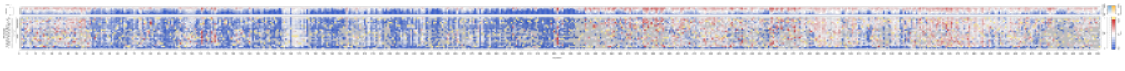
(c) f100AV

(d) f200AV

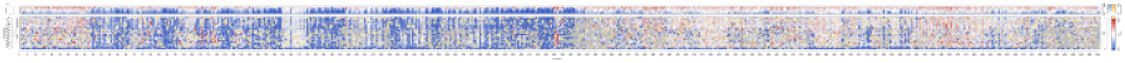
Figure 9: enrichment value distribution of MTHFR variants in A222V background, separated by different folate concentrations.

A comparison between the MTHFR variant effect maps calculated by TileseqPro and the Legacy pipelines is shown in Figure 10 and 11. The comparison is complicated by the fact that the maps for f12AV, f25AV, and f100AV available on MaveDB were previously "flipped and floored", i.e. any variants with scores above 1 were inverted by a $f(x) = \frac{1}{x}$ operation and scores below zero were set to exactly zero. This is visible

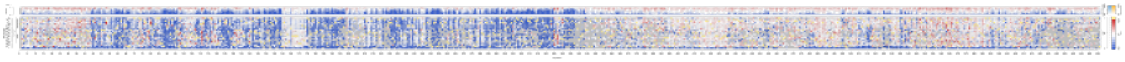
especially in the serine-rich region and regulatory domain, where the maps calculated by TileseqPro pipelines show hypercomplementing variants which are not visible in the MaveDB map.



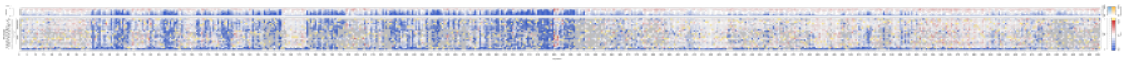
(a) f12AV



(b) f25AV

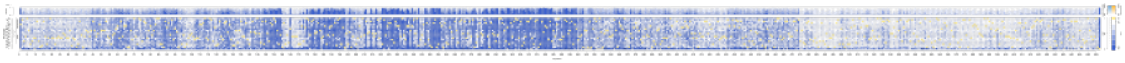


(c) f100AV

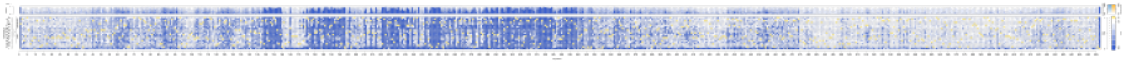


(d) f200AV

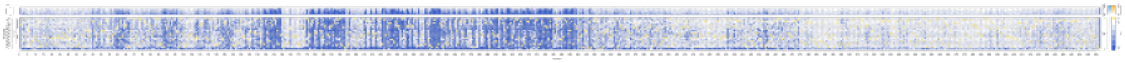
Figure 10: Mavevis heatmaps for MTHFR variants in A222V background calculated by TileseqPro pipelines, separated by different folate concentrations.



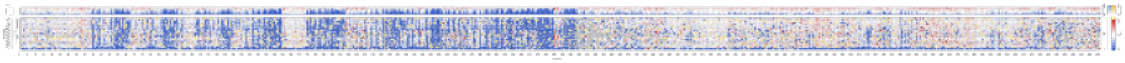
(a) f12AV



(b) f25AV



(c) f100AV

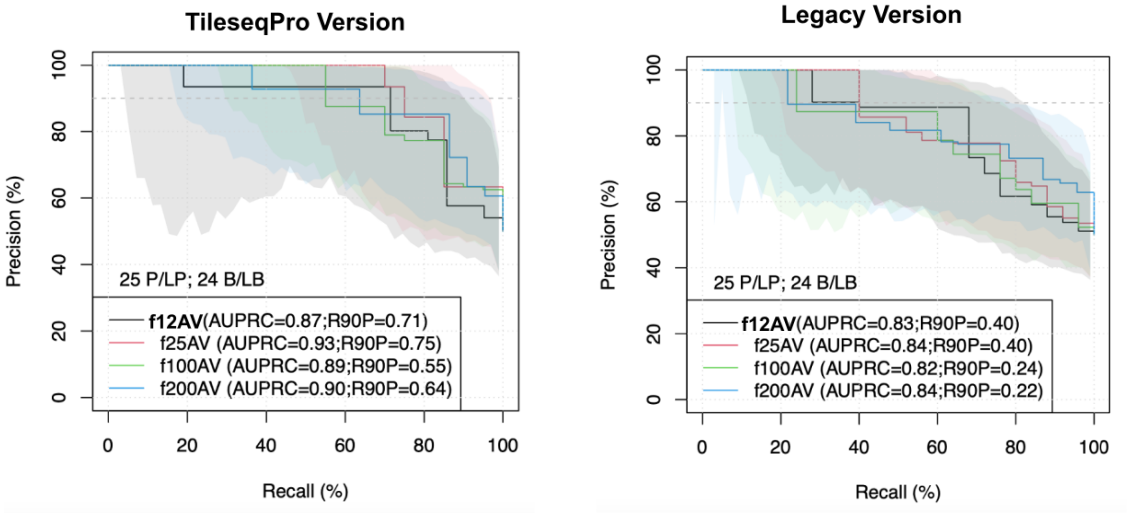


(d) f200AV

Figure 11: Mavevis heatmaps for MTHFR variants in A222V background calculated by Legacy pipelines, separated by different folate concentrations. Fitness scores for MTHFR with folate concentration f12AV, f25AV, and f100AV are flipped and floored

3.3.2 Precision-recall Curve and Log-likelihood Ratio for Pathogenicity

Comparing the performance of our updated TileseqPro version MTHFR maps and the Legacy version by the Precision-recall Curve 12, the TileseqPro version outperformed the Legacy version, with greater area under PRC and better recall at 90% precision. Inspecting the TileseqPro version PRC, and comparing between MTHFR maps at different folate concentrations, the map at 25 μg/ml (f25AV) performed the best with area under PRC of 0.93, and a recall of 75% at 90% precision.



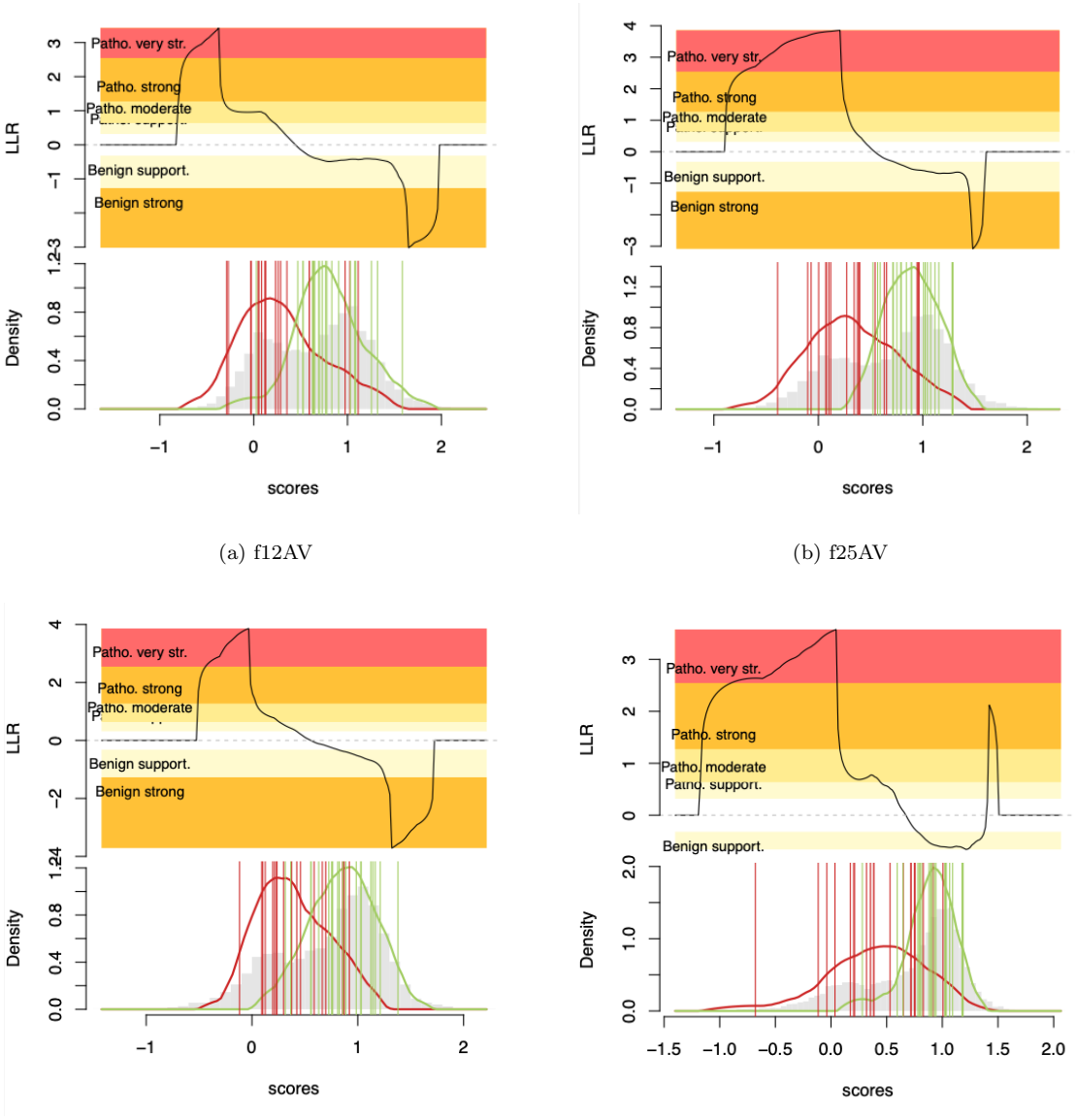
(a) 2023 version PRC

(b) 2021 version PRC

Figure 12: PRC for 2023 and 2021 version of MTHFR maps in A222V background, compared between different folate concentrations(map: f12AV, f25AV, f100AV, f200AV) using the manually curated reference sets from Weile et al.[12]

The transformation to log-likelihood ratio for TileseqPro MTHFR maps at different folate concentrations are compared in Figure 13. The overall shapes of these LLR curves are similar despite some minor differences, where LLR is positive when the fitness scores range from -0.5 to 0.5, indicate a tendency towards pathogenicity. Whereas LLR is negative when the fitness score is approximately between 0.8 to 1.5. There is still issues with the transformation function, since in the f200AV map the negative LLR spikes at

fitness scores above 1.5 with no real reference data.



(a) f12AV

(b) f25AV

(c) f100AV

(d) f200AV

Figure 13: LLR of pathogenicity of MTHFR variants in A222V background, separated by different folate concentrations.

3.3.3 Conclusion

The TileseqPro version of MTHFR maps outperformed our Legacy version in terms of getting more precision in classifying variants. However, the fitness scores calculated by the Legacy and TileseqPro pipelines need to be compared again after finding the non-flipped and floored scores.

4 Discussion

In this BCB330 Project, we successfully re-processed the variant effect maps for several genes using the TileseqPro pipelines. We evaluated their performance by inspecting QC results, comparing fitness scores with the Legacy pipelines, finding the moving window correlation to computational predictor VARITY, and drawing the PRC. We also inferred the fitness score ranges that corresponding to "benign" and "pathogenic" by calculating Log-likelihood of pathogenicity.

The TileseqPro and the Legacy pipeline performed similarly in SUMO1 map, but TileseqPro outperformed the Legacy pipelines in CALM1 map however at the expense of losing much of its original coverage, since we filtered out a lot more low quality data than before. To address this problem, machine-learning methods could be applied to impute those "missing spots" in the next step. For the MTHFR map in A222V background, TileseqPro pipelines performed better in getting more precision in prediction, but the fitness scores calculated by the TileseqPro and Legacy pipelines need to be compared after re-scaling them.

Other future goals for this project are as follows: First, we will continue to re-processing more of the existing maps on MaveDB, compare the results with the older version, and give recommendation on optimizing the pipeline implementations. Second, we will try curating benchmark reference sets for disease genes, when there is no good reference available for them. Third, we will continue to document the results for the maps that we already re-processed to a GitHub wiki page. Furthermore, we will also compare the new and old fitness scores of disease genes to other computational predictors such as ESM1v[16] and PROVEAN[17].

5 Supplementary Materials

Supplementary Materials can be found on GitHub at Supplementary

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] Jochen Weile and Frederick P. Roth. Multiplexed assays of variant effects contribute to a growing genotype–phenotype atlas. *Human Genetics*, 137(9):665–678, September 2018.

[3] 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.

[4] 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 Nussbaum, 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning: with Applications in R*. Springer Texts in Statistics. Springer US, New York, NY, 2021.

[10] Melissa J Landrum, Jennifer M Lee, Mark Benson, Garth R Brown, Chen Chao, Shanmuga Chitipiralla, Baoshan Gu, Jennifer Hart, Douglas Hoffman, Wonhee Jang, Karen Karapetyan, Kenneth Katz, Chunlei Liu, Zenith Maddipatla, Adriana Malheiro, Kurt McDaniel, Michael Ovetsky, George Riley, George Zhou, J Bradley Holmes, Brandi L Kattman, and Donna R Maglott. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Research*, 46(D1):D1062–D1067, 11 2017.

[11] Siwei Chen, Laurent C. Francioli, Julia K. Goodrich, Ryan L. Collins, Masahiro

Kanai, Qingbo Wang, Jessica Alföldi, Nicholas A. Watts, Christopher Vittal, Laura D. Gauthier, Timothy Poterba, Michael W. Wilson, Yekaterina Tarasova, William Phu, Mary T. Yohannes, Zan Koenig, Yossi Farjoun, Eric Banks, Stacey Donnelly, Stacey Gabriel, Namrata Gupta, Steven Ferriera, Charlotte Tolonen, Sam Novod, Louis Bergelson, David Roazen, Valentin Ruano-Rubio, Miguel Covarrubias, Christopher Llanwarne, Nikelle Petrillo, Gordon Wade, Thibault Jeandet, Ruchi Munshi, Kathleen Tibbetts, gnomAD Project Consortium, Anne O'Donnell-Luria, Matthew Solomonson, Cotton Seed, Alicia R. Martin, Michael E. Talkowski, Heidi L. Rehm, Mark J. Daly, Grace Tiao, Benjamin M. Neale, Daniel G. MacArthur, and Konrad J. Karczewski. A genome-wide mutational constraint map quantified from variation in 76,156 human genomes. *bioRxiv*, 2022.

[12] 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.

[13] 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.

[14] Daniel Esposito, Jochen Weile, Jay Shendure, Lea M. Starita, Anthony T. Papenfuss, Frederick P. Roth, Douglas M. Fowler, and Alan F. Rubin. MaveDB: an

open-source platform to distribute and interpret data from multiplexed assays of variant effect. *Genome Biology*, 20(1):223, December 2019.

[15] Peter Bayer, Andreas Arndt, Susanne Metzger, Rohit Mahajan, Frauke Melchior, Rainer Jaenicke, and Jörg Becker. Structure determination of the small ubiquitin-related modifier sumo-1. *Journal of Molecular Biology*, 280(2):275–286, 1998.

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

[17] 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.