Benchmarking Variant Effect Predictors through an Unbiased Framework

Marko Mrdelja
RBIF120 - Dr. Karol Estrada
June 24, 2024

Introduction

Variants of uncertain significance (VUS) are identified genetic variants “with insufficient or conflicting evidence supporting disease association”, such that they cannot be classified as pathogenic/likely pathogenic or benign/likely benign [1]. The American College of Medical Genetics and Genomics (ACMG) previously developed guidance for the interpretation of sequence variants, where a “five-tier classification system is typically used to interpret variation at the DNA sequence or chromosome level” [1,2]. Variants identified through genetic testing must be classified for pathogenicity to be used in diagnosis of disease and/or guiding treatment, where various types of evidence, such as population frequency and functional data, are integrated to inform classification [3]. The evidence required for accurate classification is often missing or conflicting, especially for rare variants, resulting in a VUS designation [3]. This dead-end can only “be overcome by the collection of additional evidence” [3]. VUSs are common in clinical genetic testing, hindering diagnosis and clinical management of diseases [4]. There is an explicit need “for better methods for interpreting missense variants, increased availability of clinical and experimental evidence for variant classification, and more diverse representation” in genomic databases [4].

The National Human Genomic Research Institute (NHGRI) set 10 bold predictions for 2030, one of them being that the diagnostic designation of “variant of uncertain significance” (VUS) would be rendered obsolete [3]. A cohort study by Chen et al investigating VUSs in hereditary disease genetic testing revealed that most VUSs were missense changes (86.6%) [4]. Missense variants are the most common type of coding genetic variant, representing a genetic alteration in which a single base pair substitution alters the genetic code and produces a different amino acid at a particular position [5]. The predominance of VUSs remains one of the biggest challenges to precision genomic medicine, reflecting the substantial value in the prediction of significance and comprehensive classification of all protein variants [3].

Efforts have been made to develop functional assays to classify these variants, but experimental characterization is often tedious and time-consuming [6]. In order to overcome limitations and classify VUSs in the absence of evidence, computational tools are being increasingly utilized to predict effects of variants. Computational variant effect predictors (VEPs) use information about a variant, such as multiple sequence alignments, evolutionary conservation, and protein structure, to distinguish between benign and pathogenic protein variants, and predict the functional impact and effect of such variants [3,6]. VEPs are developed and trained using different methodologies and data sets, resulting in significant variations of performance and variant classification. Some predictors can generate strong evidence for variant classification, both for pathogenic effects and benign effects, and prediction quality has improved significantly, as the pool of evidence has increased and more advanced approaches are applied [3,6].

The past few years have brought an increase in new computational tools that aim to predict the pathogenicity of protein coding variation. There are three VEPs in particular

which are high-performing, cutting-edge computational tools to make such predictions. PrimateAI leverages common variants in non-human primates to train a model that predicts pathogenicity of human variants [2]. PrimateAI is a deep neural network trained using common variants from population sequencing of non-human primate species that identifies pathogenic mutations in rare disease patients [7]. ESM1b is a 650-million-parameter protein language model designed to predict all ~450 million possible missense variant effects in the human genome [8]. AlphaMissense is a deep learning model which combines protein structure from AlphaFold, evolutionary conservation, and population frequency data to train a model that improves the classification of VUSs [9]. These VEPs each have different methodologies and features, but reportedly out-perform other VEPs, per their white papers.

The performance of prediction algorithms is typically evaluated utilizing datasets composed of variants of known clinical significance [6]. Benchmarking and performance evaluation has historically relied on either biased datasets such as ClinVar (which is heavily biased for pathogenic variants), or very limited datasets of Multiplexed assays of variant effects (MAVEs) [6,10,11]. VEPs are uniquely designed, developed, and trained with varying features and approaches, but this generates biases, discordance in prediction of effects, and inconsistency in performance [6,10,11]. Despite the relevance of these novel methods, none of them have used a large-scale unbiased method to benchmark their performance. Unbiased systematic benchmarking and comprehensive assessments of performance are imperative for accurate and reliable variant interpretation, especially in personalized medicine and clinical diagnostics [6,10,11].

The UK Biobank Pharma Proteomics Project (UKB-PPP) has released exome sequence and proteomics data for 49,736 individuals, encompassing 2,923 protein abundances using the OLINK panel [12]. This resource provides a unique opportunity to benchmark the three above mentioned tools using an unbiased framework. The objective of this research is to use the effect of missense variants in these 2,923 protein levels via exome-wide association study (ExWAS) as readouts to evaluate the performance of these three tools. In the research presented below, a statistical genetics analysis encompassing all synonymous and missense variants, including those with MAC less than 6, is conducted within the 2,923 genes coding for proteins in the OLINK panel. The mean effect on predicted loss of functions is used to define a threshold of pathogenicity for each gene, where effects of synonymous variants are used as negative controls. Using this framework and summarization of general patterns, benchmarking and systematic performance evaluation of PrimateAI, ESM1b, and AlphaMissense can be accomplished.

Methods

Data

All relevant data was accessed, stored, and analyzed through the Google Cloud Platform and Jupyter Notebooks. Of the 469,835 exome-sequenced UKB participants, a subset of 53,017 participants were included in the analysis. For details concerning

whole-exome sequencing, read alignment, and variant calling, please refer to the UK BioBank showcase [13]. Genetic data of all sequenced participants were provided in PLINK format for each chromosome (excluding Y), encompassing over 21 million variants exome-wide.

For the subset of participants, UKB made available normalized protein expression abundances [14]. For the 2,923 OLINK proteins, expression data was converted to z scores via rank-based inverse normal transformation (INT), and a profile was provided for each UKB participant. Additionally, covariate data for each of the subsetted subjects was provided in the form of a Hail MatrixTable. Hail is a Python-based genetic analysis platform, providing an efficient interface for exploring genomic data [15]. MatrixTables are a Hail-introduced distributed data structure, unifying variable input formats and supporting scalable queries [15].

ExWAS

The exome-wide association study (ExWAS) was performed using REGENIE, enabling systematic analysis and detection of associations of single-nucleotide variants (SNVs) in the human exome [16,17]. REGENIE is an efficient, statistically accurate method for parallel analysis of multiple phenotypes, enabling rapid genome-wide analysis [16,17].

In REGENIE step 1, a whole genome regression model is trained for each trait and 23 predictions for each phenotype are made [16,17]. PLINK 2.0 was utilized to reduce the pool of total variants to 12,900 high quality SNVs based on genotyping rate ≥ 90%, minor allele frequency (MAF) ≥ 5%, and R^2 correlation coefficient ≤ 0.05. The 12,900 variants were used for training the whole genome regression models. For REGENIE step 2, a larger set of variants on the same set of samples are tested for association, using a subset of ~1.8 million variants [16,17]. 39 subjects withdrew from the study and were thus removed from analysis. Due to time and technical restraints, only cis-coding-sequence (cis-CDS) associations were tested.

ExWAS performance was evaluated through Manhattan plots and quantile-quantile (QQ) plots of genomic association -log10 p-values for several randomly selected proteins, for all variants and those with MAC < 5. The plots were constructed using Python with the NumPy, pandas, Matplotlib, and Seaborn libraries. ExWAS analysis generated an estimated effect size, or BETA, for each single nucleotide variant. Variants with a BETA less than -1 are classified as loss of function variants (LoFs).

Quality Control

The OLINK platform employs two Versions: Explore and Expand. Expand-tested proteins had higher rates of missingness than Explore (mean 17.1% vs. mean 3.4%, respectively) (Figure 1).

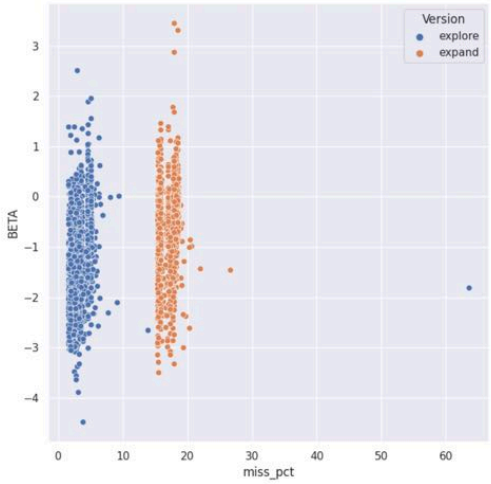

Figure 1. Missingness Distributions per Version.
Notebook: **Here**

Two proteins, CST1 and PCOLCE, had > 25% missingness and were not included in the missense analysis. In addition to Version, dilution factors associated with normalized protein expression abundances had significant confounding effects. The different groups of dilution factors (1:1, 1:10, 1:100, 1:1000, 1:100000) were investigated in both versions, comparing mean BETA and percent loss of function. 1:1 dilution proteins were less sensitive in detecting true variant effects, especially in the Expand version. 978 proteins that were 1:1 and in the Expand Version were removed from analysis. Of all proteins, 33 proteins were found to have mean BETA > 1 at the gene level, reflecting a gain of function variant, and were thus removed from analysis. Post QC, 1,202 genes are included in analysis.

dbNSFP

The database for nonsynonymous SNV functional predictions (dbNSFP4.7a) is a public database developed for functional prediction and annotation of all potential non-synonymous single-nucleotide variants (nsSNVs) in the human genome [18]. The database compiles functional annotation data, prediction scores from different algorithms, and other relevant information for both variants and protein-coding genes. This database sourced the variant pathogenicity predictions for PrimateAI, ESM1b, and AlphaMissense. Haploinsufficiency was predicted using HIPred, a machine learning approach that estimates the probability of a gene being haploinsufficient [20]. Known recessive status was obtained from MacArthur et al, which defined functional and evolutionary differences between LoF-tolerant and recessive disease genes, and provided a classification of recessive genes [21]. The downloadable dbNSFP4.7a.zip file provided a Java-based search tool for extracting data from the database based on specific inputs [18]. Of all available variants across 23 Chromosomes, 43 variants from Chromosome 10 were not present in this database and were excluded from analysis.

VEP Predictions

The PrimateAI deep learning network “accurately assigns high pathogenicity scores to residues at critical protein functional domains” [7]. The score ranges from 0 to 1, where a larger score reflects a greater likelihood of pathogenicity, and the authors suggest a threshold of 0.803 to separate damaging versus tolerable variants [7,18]. The ESM1b protein language model uses log-likelihood ratio (LLR) scores for predicting the pathogenic effects, ranging from -24.538 to 6.937. An LLR threshold of −7.5 was used to distinguish between pathogenic and benign variants, and a smaller score signifies a greater likelihood for a variant to be pathogenic [8,18]. The AlphaMissense deep learning model uses threshold score values to interpret a variant as “likely pathogenic,” “ambiguous,” or “likely benign” [9,18]. DbNSFP4.7 provides existing predictions of pathogenicity for each VEP, which were simplified and utilized for analysis. For PrimateAI and ESM1b, variants designated “Damaging/Deleterious” were considered “Pathogenic”. For AlphaMissense, Ambiguous and Benign designations were considered “NonPathogenic”.

Missense analysis

REGENIE outputs were merged with a provided annotation file, generating a combined file of 1,092,983 variants across 23 chromosomes. Missense variants for all chromosomes were extracted for analysis, of which there were 348,160 total variants. A supplemental data file from Dhindsa et al provided dilution factor and Version information at the protein level [19]. Investigation of dilution factors across allele frequencies was used for quality control and quality assurance. In order to ensure accurate merging and analysis of proteomic data, a unique ID was generated using variant genomic locus information and corresponding Ensembl ID.

For all missense variants, a pathogenicity classification logistic regression model was fit to the data with a pathogenicity prediction acting as the binary outcome variable. Descriptive tables were used to compare the number of pathogenic and non-pathogenic (benign) calls made by each VEP. Additionally, the variations in each VEP were investigated based on allele 1 frequency bins: Singleton (MAC = 1), Doubleton (MAC = 2), MAF < 0.01%, MAF 0.01-0.1%, MAF 0.1-1%, MAF 1-10%, and MAF > 10%. Comparison using allele frequency bins facilitated evaluation of allele frequency effects. After filtering out proteins identified by quality control, further analysis was performed using statistically significant variants, with a -log10 P-value greater than 7, including 5,150 variants. Singleton and Doubleton designations were removed from significant variant analysis due to low variant counts.

Statistically significant variant functional data was merged with data from dbNSFP. Of the 5,150 significant variants included, a set of significant variants were not annotated in dbNSFP, leaving 2,664 missense variants available for further analysis. For data files containing relevant variant information and predictions, this is a link for All Missense, and this is a link for only Significant Missense Variants. Logistic Regression described above was repeated for statistically significant variants. In addition to the defined

threshold of pathogenicity for each gene, general patterns are summarized by gene set analysis (haploinsufficiency, recessive genes, enrichment), protein families and domains, and allele frequencies. Using the data associated with the remaining genes, variant pathogenicity predictions were compared for each VEP for haploinsufficient versus haplosufficient genes, as well as recessive and non-recessive genes. Simple gene set enrichment analyses were performed using the 571 genes, identifying enriched and statistically significant GO terms and protein domains. Enrichment analysis was first performed for all variants, followed by analysis on rare variants, where variants with positive and negative BETA were evaluated separately.

Rare variants were identified by excluding any variants with an MAF > 0.1%. The number of rare variants with positive versus negative BETAs was quantified, in which 186 positive BETA variants and 957 negative BETA variants were identified. Gene/variant enrichment was performed using the EnrichR package in R [22]. EnrichR is a public, comprehensive domain tool for analyzing gene sets generated by genome-wide experiments, facilitating the investigation of a set of genes across a variety of databases, including GO, KEGG, and InterPro [22]. For this analysis, the following libraries were utilized: GO_Biological_Process_2023, GO_Cellular_Component_2023, GO_Molecular_Function_2023, InterPro_Domains_2019. Enrichment analysis was performed separately for variants with positive and negative BETA values, representing gain of function and loss of function, respectively. Gene set enrichment analysis facilitates the identification of significant, trait-associated genes and the prioritization of variants within the enriched genes/pathways.

Results

ExWAS Performance

Based on the Manhattan and QQ diagnostic plots, a small number of genuine significant signals are expected, but the analysis overall was successful with sufficient statistical power in detection of associations. One protein was randomly selected from each of Chromosomes 1 (Figure 2, a,b) and 22 (Figure 3 a,b), to generate diagnostic plots. ExWAS performed as expected. QQ plots in both Chromosomes reflect expected deviation from the null distribution, showing no inflation until the values pass a reasonable level of significance. The plots appear as anticipated, reflecting successful analysis.

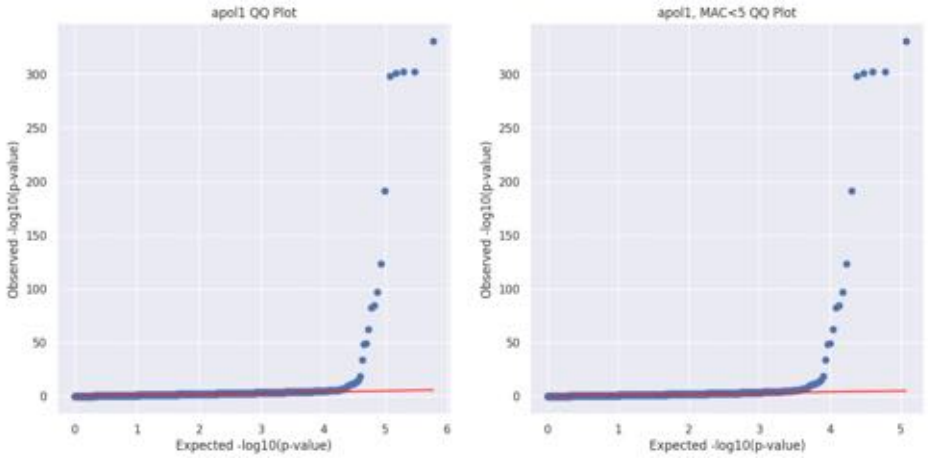

Figure 2a. Chromosome 1 Protein APOL1 QQ plots
Notebook: **Here**

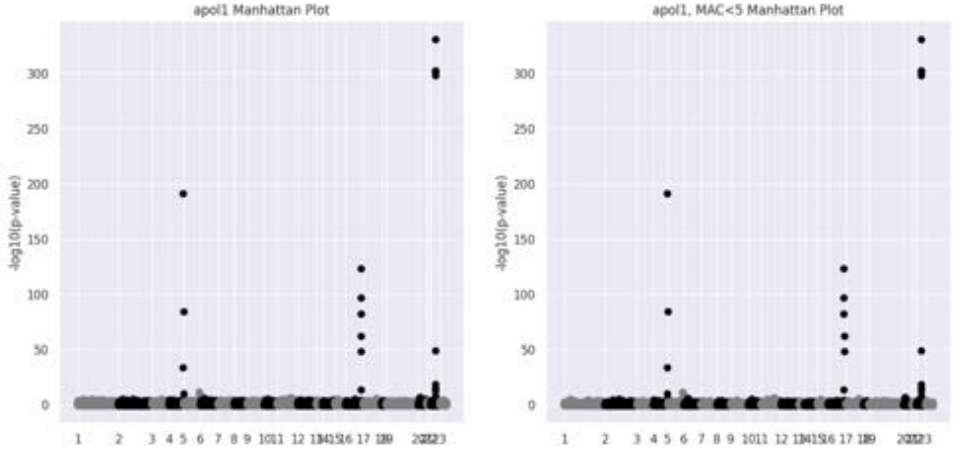

Figure 2b. APOL1 Manhattan plots
Notebook: **Here**

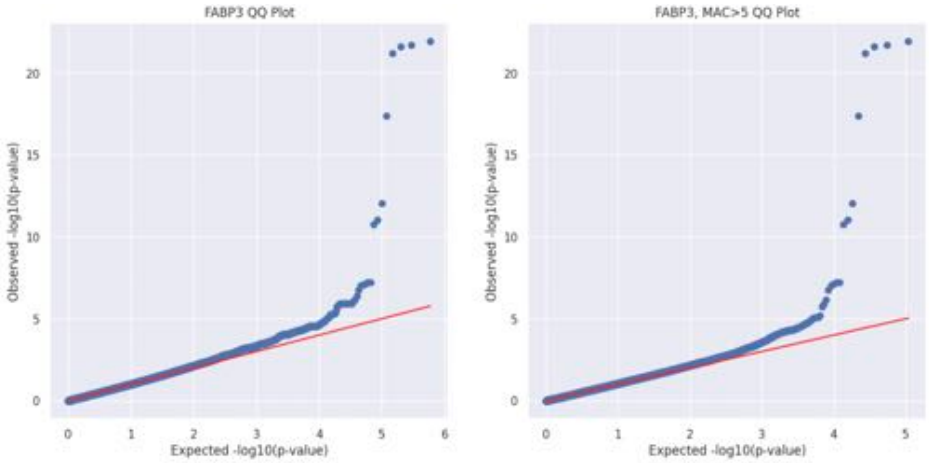

Figure 3a. Chromosome 22 Protein FABP3 QQ plots
Notebook: **Here**

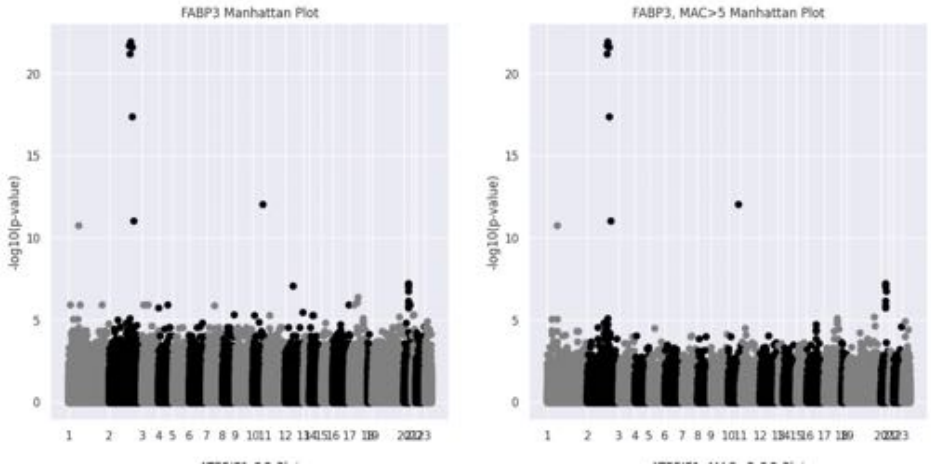

Figure 3b. FABP3 Manhattan plots
Notebook: **Here**

All Missense Variant Analysis Results

43 variants from Chromosome 10 were not present in dbNSFP, leaving a total of 348,117 missense variants for inclusion in the analysis. From this total, each VEP predicted varying numbers of pathogenic variants: PrimateAI - 31182; ESM1b - 112809; AlphaMissense - 55323. Variants with a BETA less than -1 can be classified as loss of function variants (LoFs). The mean effect on predicted loss of functions (mean BETA) is used to define a threshold of pathogenicity for each gene. For all missense variants, there is a significant disparity in which variants are predicted pathogenic and benign by PrimateAI (Table 1). There is also a significant difference in mean BETA between pathogenic and nonpathogenic calls by each VEP. Pathogenic calls by PrimateAI have a lower mean BETA than the other VEPs (-0.168 versus -0.317 and -0.409) (Table 1).

A data.frame: 6 × 8

	Count	Mean	SE	Min	25%	Median	75%	Max
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
PrimateAI: Pathogenic	31182	-0.1680561	1.0114009	-4.44400	-0.749608	-0.11204200	0.4466960	4.30417
PrimateAI: Nonpathogenic	309112	-0.1751454	0.9518925	-5.10702	-0.701931	-0.11422800	0.3739430	4.90521
ESM1b: Pathogenic	112809	-0.3166021	1.0641609	-5.10702	-0.933556	-0.23382400	0.3262300	4.90521
ESM1b: Nonpathogenic	203613	-0.1052860	0.8886649	-4.22075	-0.603890	-0.07013080	0.4007390	4.46768
AlphaMissense: Pathogenic	55323	-0.4090306	1.1558966	-5.10702	-1.114380	-0.32265400	0.3189655	4.65958
AlphaMissense: Nonpathogenic	292794	-0.1311721	0.9075045	-4.54508	-0.639858	-0.08752125	0.3878760	4.90521

Table 1. Descriptive Table of VEP predictions for all missense variants
Notebook: **Here**

Logistic Regression including all missense variants depicted unimpressive performance for all 3 VEPs. Due to significant differences in the scoring systems used by each VEP to predict pathogenicity, sensitivity was normalized to ~90% for a more accurate comparison, following a similar scheme as Gunning et al [23]. The pathogenicity prediction established with the BETA < -1 was used for comparison. The auROC for each VEP is slightly above 0.5, reflective of suboptimal performance but capability of class differentiation (PrimateAI: 0.508; ESM1b: 0.546; AlphaMissense: 0.564). Below are the summary details of the Logistic Regression for all VEPs (Figure 4,a-c).

```
Call:
glm(formula = PATHOGENICITY ~ primAI.pred.cat + A1FREQ, family = "binomial",
    data = allchr.merged.pred)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.6374  -0.6138  -0.6137  -0.6131   6.0758

Coefficients:
              Estimate Std. Error  z value Pr(>|z|)
(Intercept)  -1.573570   0.004748 -331.439  < 2e-16 ***
primAI.pred.cat   0.083201   0.015382    5.409 6.33e-08 ***
A1FREQ       -16.893162   1.314864  -12.848  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 318641  on 348116  degrees of freedom
Residual deviance: 317788  on 348114  degrees of freedom
AIC: 317794

Number of Fisher Scoring iterations: 8
```

Figure 4a. PrimateAI LR, all missense
Notebook: **Here**

```
Call:
glm(formula = PATHOGENICITY ~ esm.pred.cat + A1FREQ, family = "binomial",
    data = allchr.merged.pred)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.7294  -0.5561  -0.5561  -0.5558   5.8070

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   -1.788173   0.005933 -301.37   <2e-16 ***
esm.pred.cat   0.600046   0.009205   65.19   <2e-16 ***
A1FREQ       -15.080543   1.243092  -12.13   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 318641  on 348116  degrees of freedom
Residual deviance: 313652  on 348114  degrees of freedom
AIC: 313658

Number of Fisher Scoring iterations: 8
```

Figure 4b. ESM1b LR, all missense
Notebook: **Here**

```
Call:
glm(formula = PATHOGENICITY ~ am.pred.cat + A1FREQ, family = "binomial",
    data = allchr.merged.pred)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.8098  -0.5749  -0.5749  -0.5744   5.7779

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.716210   0.005184 -331.07   <2e-16 ***
am.pred.cat   0.769647   0.010800   71.26   <2e-16 ***
A1FREQ      -14.983942   1.230987  -12.17   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 318641  on 348116  degrees of freedom
Residual deviance: 313082  on 348114  degrees of freedom
AIC: 313088

Number of Fisher Scoring iterations: 8
```

Figure 4c. AlphaMissense LR, all missense
Notebook: **Here**

The summary data of each logistic regression reflects that allele frequency is a significant contributor to pathogenicity, and meta BETA is significantly correlated to allele frequency. Looking at the allele frequencies for all missense variants, mean BETA follows a trend of decreasing as allele frequency decreases (Table 2).

	Count	Mean	SE	Min	25%	Median	75%	Max
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
Singleton	185621	-0.18509453	1.1195100	-5.10702	-0.9021770	-0.15444100	0.56247800	4.90521
Doubleton	53667	-0.17606040	0.8713977	-3.98529	-0.6989090	-0.13637300	0.39094150	4.32256
MAF < 0.01%	69514	-0.16948717	0.7110506	-3.70177	-0.5355460	-0.11516800	0.26662800	4.35993
MAF 0.01-0.1%	28454	-0.15320570	0.5379724	-3.28311	-0.3375610	-0.07945600	0.12863200	4.03262
MAF 0.1-1%	6468	-0.11895171	0.4367265	-2.96496	-0.1987735	-0.03782900	0.05719295	2.96642
MAF 1-10%	2249	-0.10861732	0.3664190	-2.31966	-0.1648380	-0.01948640	0.02488410	2.35986
MAF > 10%	2144	-0.03471287	0.3284522	-1.78947	-0.1060755	-0.00314285	0.05417395	1.91470

Table 2. All missense variants, separated by allele frequency bins
Notebook: **Here**

After analyzing the data above, investigations into protein dilutions, OLINK panel version, and proportions of missing data identified problems with certain genes (see Methods - Quality Control). After applying quality control filters, 208,913 missense variants remained for further analysis. From this total, each VEP predicted varying numbers of pathogenic variants: PrimateAI - 15973; ESM1b - 69494; AlphaMissense - 33415. There is a significant disparity across VEPs in pathogenic versus nonpathogenic predictions (Figure 5).

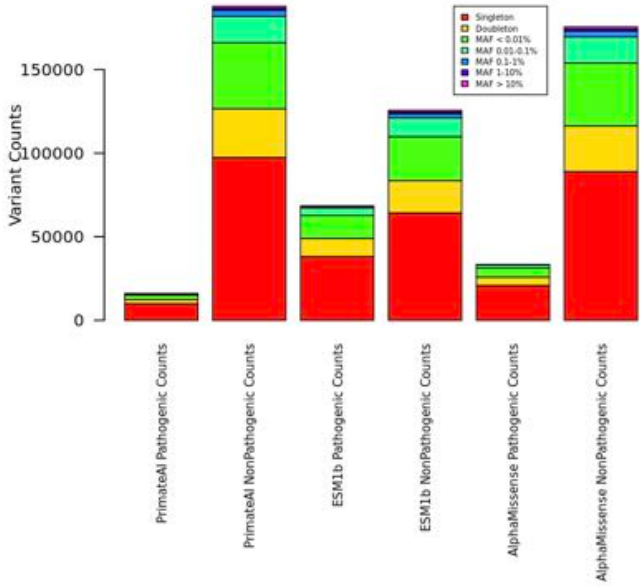

Figure 5. Prediction count comparisons per VEP, for all missense variants post QC
Notebook: **Here**

The trend of BETA decreases with decreasing allele frequency continues across all 3 VEPs (Table 3). The mean BETA for PrimateAI is noticeably different from the other VEPs. Pathogenic variants should, theoretically, have a lower mean BETA, reflecting a more significant loss of function. For PrimateAI, mean BETA for pathogenic and nonpathogenic predictions are very similar.

	Singleton	Doubleton	MAF < 0.01%	MAF 0.01-0.1%	MAF 0.1-1%	MAF 1-10%	MAF > 10%
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
PrimateAI Path Mean BETA	-0.2869152	-0.2412468	-0.2169854	-0.213508	-0.1272041	-0.1691529	-0.04495375
PrimateAI NonPath Mean BETA	-0.2575853	-0.2451313	-0.2341913	-0.2076731	-0.1573591	-0.1570438	-0.04281093
ESM1b Path Mean BETA	-0.4555181	-0.4263911	-0.4052827	-0.3827036	-0.3088751	-0.3179253	-0.2089483
ESM1b NonPath Mean BETA	-0.1433497	-0.1418993	-0.1452732	-0.1364483	-0.1095794	-0.1348865	-0.0443556
AlphaMissense Path Mean BETA	-0.5855128	-0.529366	-0.5304756	-0.5188807	-0.4065621	-0.3839851	-0.2037819
AlphaMissense NonPath Mean BETA	-0.1851766	-0.1951205	-0.1904161	-0.1773141	-0.1419657	-0.145326	-0.04517307

Table 3. Descriptive table for mean BETA of VEPs across allele frequencies
Notebook: **Here**

In the other VEPs, the mean BETA is higher for pathogenic predictions than nonpathogenic predictions, and both groups have a general pattern of decreasing mean BETA as allele frequency decreases (Table 3).

Significant Missense Analysis Results

The statistical significance of ExWAS results is reflected by -log10 p-value. The total pool of missense variants was subsetted using a -log10 p-value > 7. The analysis for all missense variants was replicated using only statistically significant variants. Due to a low count of variants, Singleton and Doubleton designations were not considered in subsequent analyses. There are significantly fewer pathogenic variants and significantly more nonpathogenic variants called by PrimateAI, as compared to the other VEPs, as observed in Figure 6.

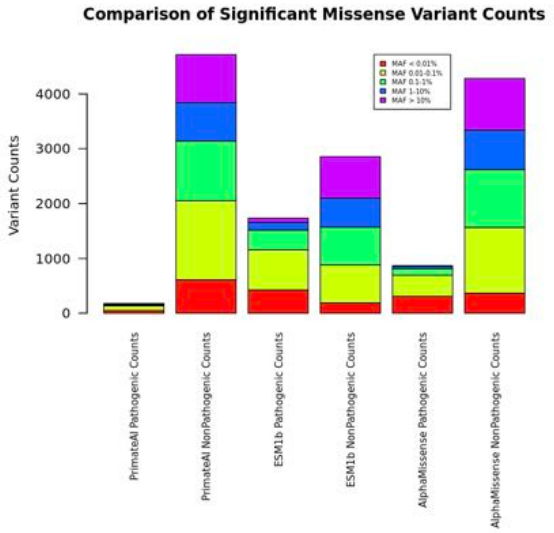

Figure 6. Comparison of Significant variant counts per VEP
Notebook: **Here**

Figure 7 illustrates the distribution of BETA values for VEPs, separated by prediction. The mean BETA for pathogenic and nonpathogenic are closer than expected, but there is a noticeable difference in AlphaMissense and ESM1b. The mean BETA for pathogenic AlphaMissense predictions is lower than the other VEPs. Mean BETA for nonpathogenic predictions by ESM1b is closer to 0 than other VEPs. The mean BETA for significant variants follows the trend of decreasing as allele frequency decreases across VEPs (Table 4).

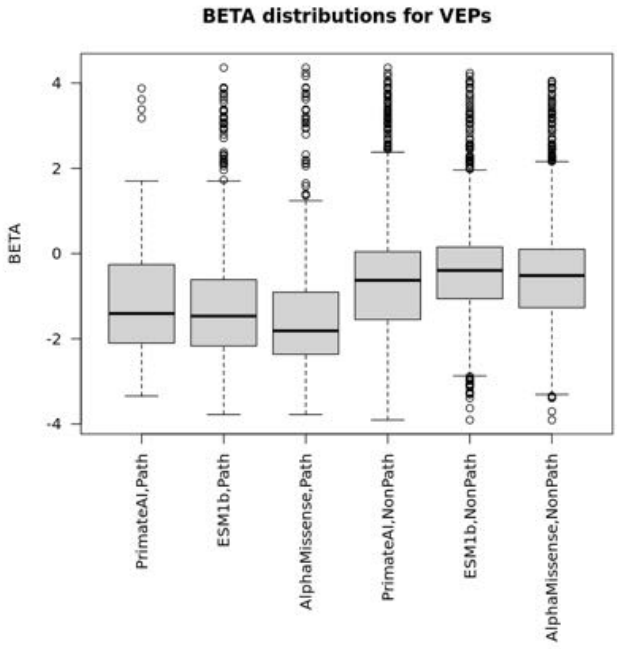

Figure 7. Distribution of BETA values for significant variants
Notebook: **Here**

	MAF < 0.01%	MAF 0.01-0.1%	MAF 0.1-1%	MAF 1-10%	MAF > 10%
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
PrimateAI Path Mean BETA	-1.642492	-1.191491	-0.7301853	-0.3254661	-0.05858
PrimateAI NonPath Mean BETA	-1.449974	-1.128525	-0.4622446	-0.2679857	-0.06541632
ESM1b Path Mean BETA	-1.798427	-1.398397	-0.6938958	-0.4898729	-0.2668411
ESM1b NonPath Mean BETA	-0.8602933	-0.8833435	-0.3559475	-0.2401004	-0.06729471
AlphaMissense Path Mean BETA	-1.724338	-1.533075	-0.8491859	-0.5667127	-0.3198649
AlphaMissense NonPath Mean BETA	-1.271266	-1.002788	-0.4221957	-0.247211	-0.06755533

Table 4. Descriptive Table of mean BETA values, across VEPs for significant variants only
Notebook: **Here**

The analysis of dilution factor effect was repeated as above for significant missense variants, post QC. For dilution factors, non-1:1 dilution genes have a significantly higher mean BETA than 1:1 dilution genes, the highest in rare variants,

	1:1 Counts	Non-1:1 Counts	1:1 Mean BETA	Non-1:1 Mean BETA
	<named list>	<named list>	<named list>	<named list>
MAF < 0.01%	166	509	-0.5634505	-1.776263
MAF 0.01-0.1%	480	1105	-0.9271864	-1.220869
MAF 0.1-1%	392	781	-0.3940537	-0.5013807
MAF 1-10%	267	496	-0.2836383	-0.256589
MAF > 10%	359	595	-0.0636262	-0.07459057

Table 5. Comparison of Dilution factors for Significant variants
Notebook: **Here**

decreasing as allele frequency decreases. There are significantly more non-1:1 variants than 1:1 variants (Table 5).

All following analysis is post quality control filtering, excluding genes in the 1:1 dilution factor/expand panel as well as the other genes as detailed in the methods (Quality Control). For the 2,664 statistically significant missense variants, haploinsufficiency status and known recessive gene status were utilized for comprehensive comparisons. 188 genes were predicted to be haploinsufficient. There is a moderate difference in mean BETA between haplosufficient and haploinsufficient genes, which follows the trend identified above. The number of missense variants per VEP is similar as above, but there is a noticeable difference in mean BETAs for ESM1b. For PrimateAI and AlphaMissense, the mean BETA for nonpathogenic predictions is higher than pathogenic predictions. This represents a significant issue in accurately classifying variants for haploinsufficient genes. Corresponding descriptive tables and plots can be found in the Jupyter notebooks (supplemental data: Here).

There were 42 genes known to be recessive, per dbNSFP. The mean BETA was higher for recessive genes, and both recessive and non-recessive gene groups follow the trend of decreasing with decreasing allele frequency. For recessive genes, all 3 VEPs have a similar BETA for pathogenic and nonpathogenic predictions. This depicts a lack of capability in accurately classification of variants of recessive genes. This should be considered carefully due to the low number of genes investigated. Corresponding descriptive tables and plots can be found in the Jupyter notebooks (supplemental data: Here).

Repeating Logistic Regression with significant variants demonstrated significantly higher auROC values, where PrimateAI had the highest auROC:
AlphaMissense: 0.736; ESM1b: 0.720; PrimateAI: 0.780.
Importantly, the model summary data indicates that the PrimateAI model had the weakest statistical significance (Figure 8a). Below are the summary details of the Logistic Regression for all VEPs (Figure 8,a-c).

```
Call:
glm(formula = PATHOGENICITY ~ primstatus + A1FREQ + A1FREQ *
    primstatus, family = "binomial", data = sig.dbgene.dat)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6655  -1.1410  -0.0369   1.1410   7.5966

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)         0.08987    0.04842   1.856 0.063459 .
primstatus          1.01620    0.29341   3.463 0.000533 ***
A1FREQ            -30.10401    2.78671 -10.803  < 2e-16 ***
primstatus:A1FREQ -295.54780  132.43568  -2.232 0.025639 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3550.3  on 2663  degrees of freedom
Residual deviance: 2882.9  on 2660  degrees of freedom
AIC: 2890.9

Number of Fisher Scoring iterations: 10
```

Figure 8a. PrimateAI LR, significant missense
Notebook: **Here**

```
Call:
glm(formula = PATHOGENICITY ~ esmstatus + A1FREQ + A1FREQ * esmstatus,
    family = "binomial", data = sig.dbgene.dat)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5742  -0.9655  -0.0652   0.8305   6.7658

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       -0.42335    0.06333  -6.685 2.31e-11 ***
esmstatus          1.32104    0.10208  12.941  < 2e-16 ***
A1FREQ           -23.36497    2.83889  -8.230  < 2e-16 ***
esmstatus:A1FREQ -13.96701    6.81960  -2.048   0.0406 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3550.3  on 2663  degrees of freedom
Residual deviance: 2718.1  on 2660  degrees of freedom
AIC: 2726.1

Number of Fisher Scoring iterations: 8
```

Figure 8b. ESM1b LR, significant missense
Notebook: **Here**

```
Call:
glm(formula = PATHOGENICITY ~ amstatus + A1FREQ + A1FREQ * amstatus,
    family = "binomial", data = sig.dbgene.dat)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.824  -1.040  -0.057   1.262   6.855

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       -0.19631    0.05405  -3.632 0.000281 ***
amstatus           1.65364    0.14017  11.797  < 2e-16 ***
A1FREQ           -24.23252    2.51230  -9.646  < 2e-16 ***
amstatus:A1FREQ -153.68442   35.28172  -4.356 1.33e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3550.3  on 2663  degrees of freedom
Residual deviance: 2730.9  on 2660  degrees of freedom
AIC: 2738.9

Number of Fisher Scoring iterations: 8
```

Figure 8c. AlphaMissense LR, significant missense
Notebook: **Here**

In order to identify significant, trait-associated genes and prioritize associated variants for use in comprehensive VEP comparison, enrichment analysis was performed. Enrichment analysis was performed on the gene set for statistically significant variants, highlighting the processes and pathways implicated by protein variants. This analysis did not account for BETA, and was thus repeated for variants with positive and negative BETA separately. Considering that evidence required for accurate pathogenicity classification is often missing and/or conflicting for rare and ultra-rare variants, the pool of statistically significant variants was subsetted for rare variants, including only variants with MAF ≤ 0.1% for the repeated enrichment analyses.

Below (Fig. 9-12) are enrichment plots for the above indicated libraries, separated as Figure a and b. Figure a is the original gene set of all significant variants. Figure b includes rare variants only. Only the negative BETA rare variant plots are included in this report, because a majority of terms for positive BETA rare variants were non-significant (p-value > 0.05). The terms for negative BETA variants are more statistically significant, and are associated with significantly more genes than positive BETA variants. This is consistent across libraries. Full data files can be found in the Jupyter notebooks.

For biological processes, across significant variants post QC, proteolysis is the most enriched term (GO:0006508) (Figure 9a). This indicates that proteolysis is the process most significantly impacted by this set of genes/variants. Rare variants with a positive BETA are associated with gain of function in regulatory processes, particularly positive regulation, with the most significant being Regulation of T Cell Proliferation. Rare variants with a negative BETA are associated with loss of function in regulatory processes as well, but the most significant term is proteolysis, reflecting a decreased ability of breaking down proteins (Figure 9b). Although there are many similarities and consistent terms between all variants and rare negative BETA variants, there are strongly significant terms not found in negative BETA, such as Cytokine-Mediated Signaling Pathway and Regulation of Inflammatory Response. This seems to indicate that loss of function variants impact certain regulatory biological functions more than others.

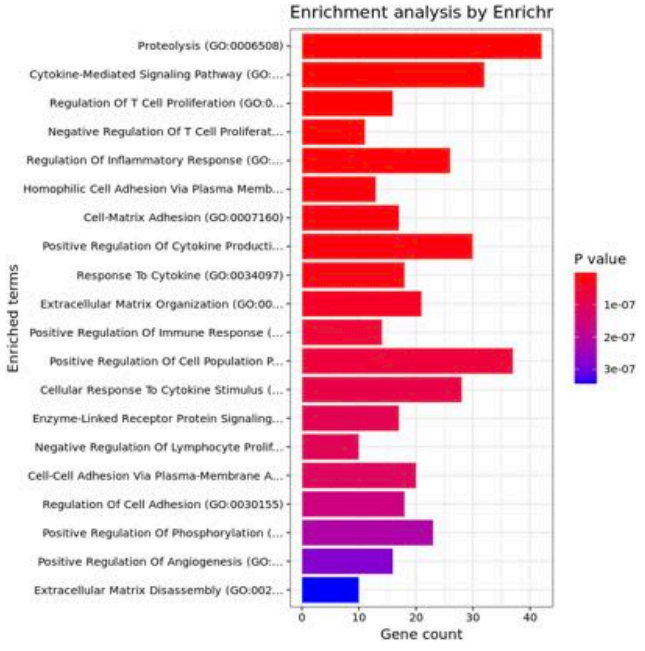

Figure 9a. GO Biological Process Enrichment plot, full gene set
Notebook: **Here**

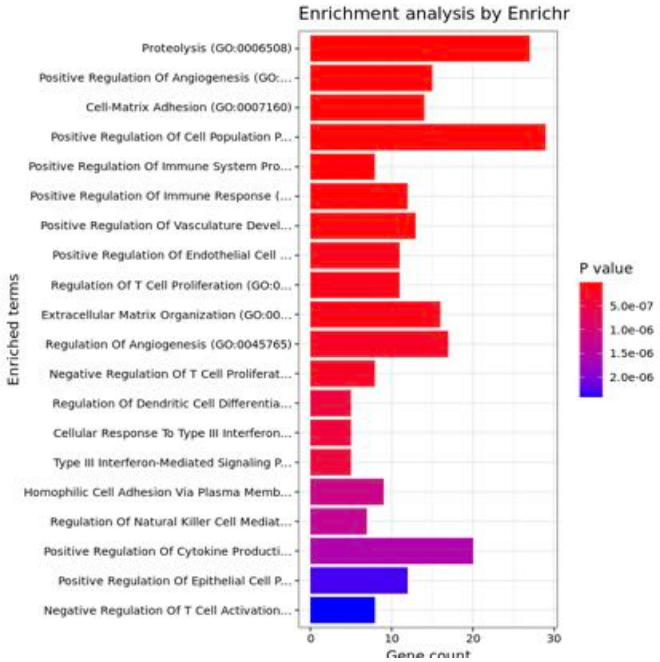

Figure 9b. GO Biological Process Enrichment plot, Rare negative BETA variants
Notebook: **Here**

For cellular components, the terms for all significant variants and rare variants are almost identical, with the primary difference being statistical significance of individual terms (Figure 10a,b). Collagen-containing Extracellular Matrix (GO:0062023) is the most significant term in both groups. Rare variants with a positive BETA do not have any statistically significant terms, as the lowest p-value is 0.2. Rare variants with negative BETA are mostly associated with the lumen and with granules, reflecting a diminished capacity of cell types in executing normal functions (Figure 10b). The terms Golgi Lumen and Lytic Vacuole are present in Figure 10b but not 10a, and Figure 10b is missing the terms Ficolin-1-Rich Granule Membrane and Perisynaptic Extracellular Matrix (Figure 10a,b).

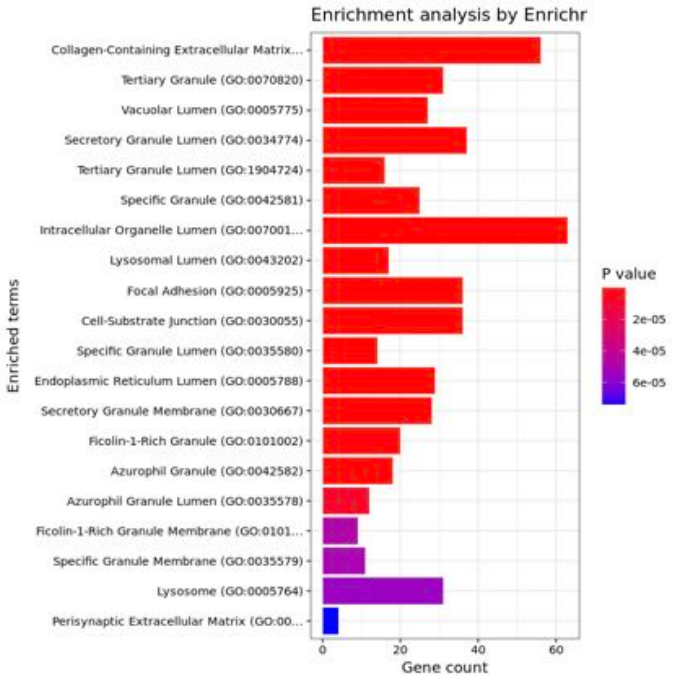

Figure 10a. GO Cellular Component Enrichment plot, full gene set
Notebook: **Here**

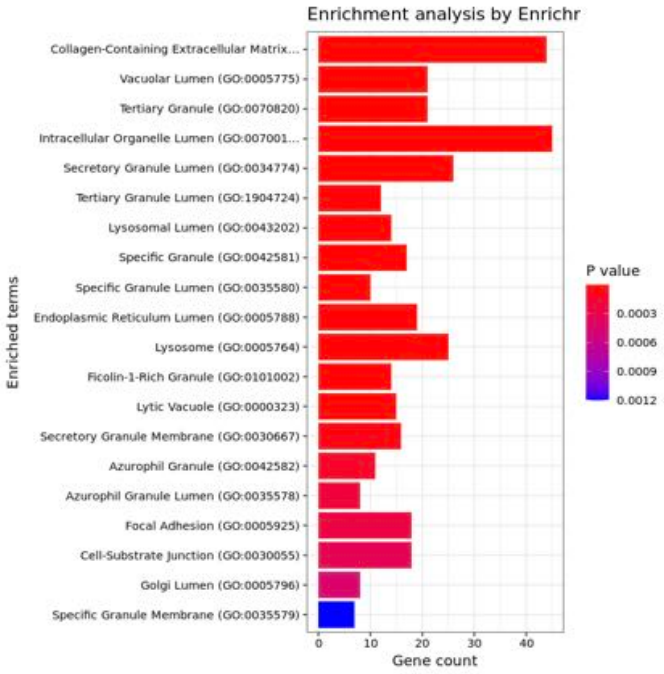

Figure 10b. GO Cellular Component Enrichment plot, Rare negative BETA variants
Notebook: **Here**

For molecular functions, both groups have the same most significant term - Cytokine Receptor Activity (GO:0004896) (Figure 11a,b). There are other shared terms, namely different types of peptidase activity. Rare variants with a positive BETA resemble those in cellular components and do not have any statistically significant terms, as the lowest p-value is 0.1. Rare variants with a negative BETA are associated with loss of function in cytokine receptor activity and various types of peptidase activity, including metallopeptidase, exopeptidase, and endopeptidase activity (Figure 11b). This correlates with the loss of function association with proteolysis identified in the GO biological processes library, as the most significant term was proteolysis, and peptidases are major functional catalyzing agents in proteolysis. There are several

terms present in Figure 11b which are missing in Figure 11a: Zinc Ion Binding, Protein Homodimerization Activity, Growth Factor Receptor Binding, Cytokine Receptor Binding (Figure 11a,b)

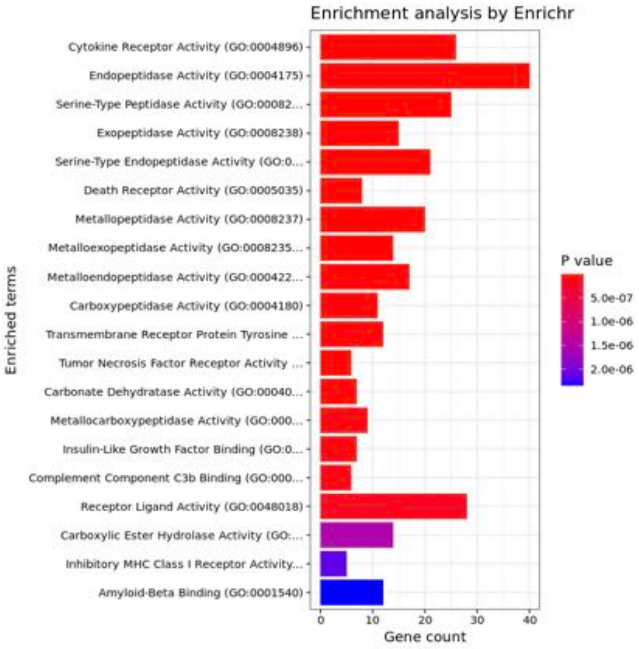

Figure 11a. GO Molecular Function Enrichment plot, full gene set
Notebook: **Here**

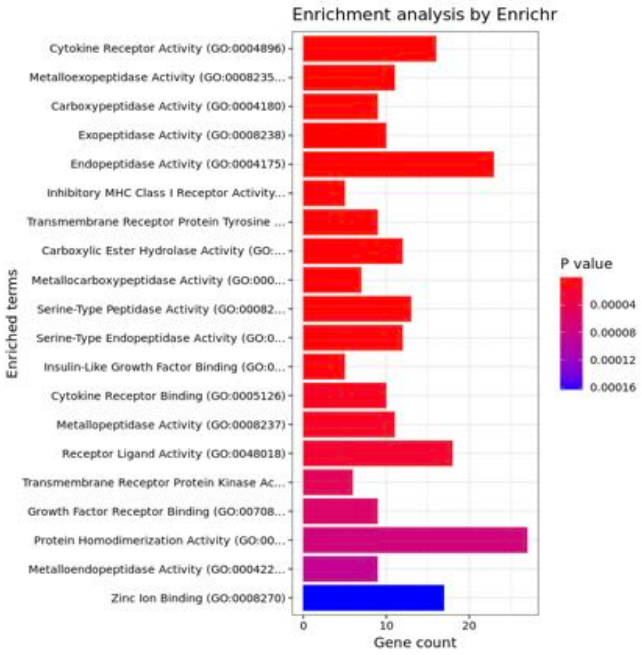

Figure 11b. GO Molecular Function Enrichment plot, Rare negative BETA variants
Notebook: **Here**

For domains, just as the libraries above, there are shared terms between the 2 groups. The most significant terms, found in both groups, related to Immunoglobulin domains (Figure 12a,b). There are more differences than expected, particularly as towards the bottom of the enrichment plots. This indicates that there are some protein domains more significantly impacted by loss of function variants than other domains. Rare variants with positive and negative BETA values shared some terms, but more terms and stronger statistical significance were identified with negative BETA values (Figure 12b). The most significant term for both was Immunoglobulin subtype, which correlates well with the information identified above. As protein domains are a major functional component of proteins, it is expected that variants have loss of function and gain of function in similar domains. The proteins in the OLINK platform are plasma proteins, meaning they can be found in the peripheral blood, and would have functions in immunity, coagulation, and other hematological processes. The domains for negative BETA variants correlate with the terms and associations identified in the other libraries. There are 3 terms present in Figure 12b which are missing from Figure 12a: GPS motif, ADAM cysteine-rich domain, von Willebrand factor type A.

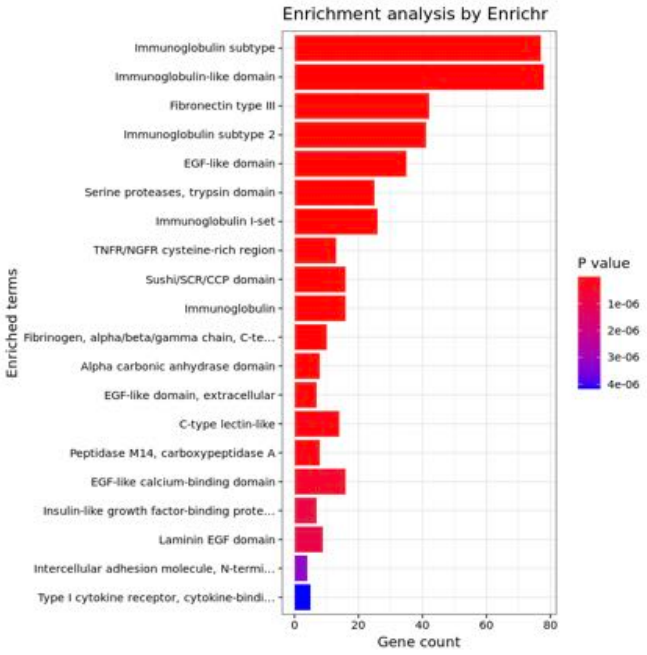

Figure 12a. InterPro Domains Enrichment plot, full gene set
Notebook: **Here**

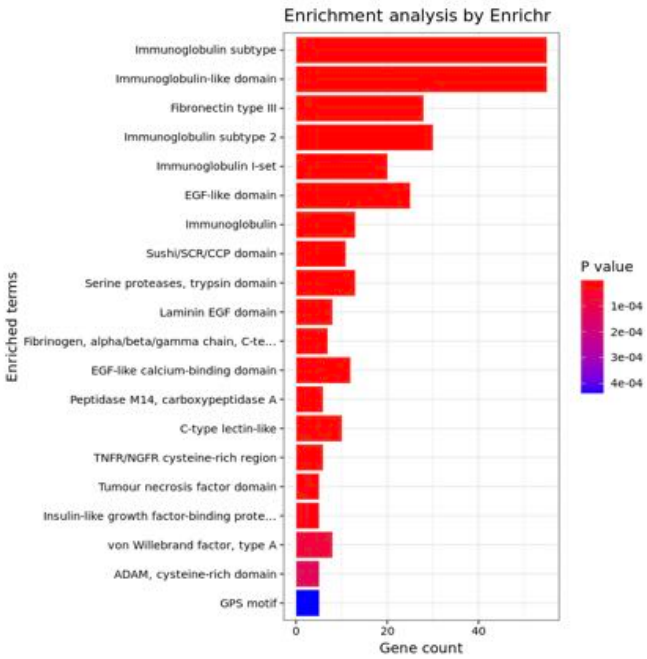

Figure 12b. InterPro Domains Enrichment plot, Rare negative BETA variants
Notebook: **Here**

For the most significant term identified in each library, genes associated with the term and corresponding variants were identified, and pathogenicity predictions made by each VEP were compared. Many of these variants were not previously identified, and could not be found in neither ClinVar nor ClinGen. For non-previously identified variants, there were discrepant results among VEP predictions. Since these predictions are made computationally and are based on an established and variable threshold, this discrepancy is expected. In order to truly indicate which VEP is correct, a methodical validation process, in the form of experimental characterization, to discover variant pathogenicity is required. A majority of the variants that were previously identified were either benign or VUS. For previous benign variants, all 3 VEPs correctly identified benign variants. For all previously identified VUSs, there were discrepant pathogenicity predictions between the 3 VEPs. There were a small number of previously identified pathogenic variants, and in a majority of these cases, all 3 VEPs correctly predicted pathogenicity. Notably, PrimateAI had the least success in accurately identifying pathogenicity. ESM1b and AlphaMissense performed similarly in this context.

Discussion

Genetic variants of uncertain significance are a hindrance in clinical genetic testing and precision medicine. Variants must be classified as pathogenic for clinical applications, and this process can be arduous, involving significant time and resource investment for successful classification. In order to circumvent this intensive process, computational variant effect predictors (VEPs) are actively being developed and employed to predict variant effect and thus pathogenicity. There are, however, a number of issues with the performance and application of VEPs, and there remains unresolved biases and inaccuracies [10,11]. Data circularity is an overwhelming problem in the training of and performance assessment of VEPs. Assessment of VEP performance is biased, as benchmarking is often performed using the same data used to train the tools [11]. In order to accurately classify VUSs and generate evidence necessary for application in clinical contexts, an unbiased framework for benchmarking is required.

Using the framework detailed in this research removes data circularity issues outlined by Livesey and Marsh, and provides detailed, comprehensive data for comparing VEP performance [10,11]. The UK Biobank Pharma Proteomics Project (UKB-PPP) exome sequence and proteomics data enabled benchmarking of three cutting-edge, high performing variant effect predictors without the use of biased data. A statistical genetics analysis and exome-wide association study provided the missense variant data necessary to evaluate the performance of the computational tools.

For this analysis, the source code of each VEP was unavailable, meaning new predictions could not be made. Original predictions of each VEP were available in public sources, and were cumulatively gathered in the dbNSFP database. This data, coupled with data regarding recessive genes and haploinsufficiency, facilitated this analysis. The effect of missense variants as readouts were investigated at several stages. There was an immediately noticeable difference in the number of pathogenic versus benign (non-pathogenic) variants called by each VEP. PrimateAI had predicted substantially fewer pathogenic variants than other VEPs, and was shown to be the weakest performing VEP throughout this analysis. There was a consistent trend of mean BETA decreasing with decreasing allele frequency. This indicates that the rarity of a missense variant is associated with stronger effect, notably stronger loss of function. The mean BETA for PrimateAI was very close in value, a strong contrast to the other VEPs which had a significant difference between pathogenic and nonpathogenic mean BETA value. Logistic Regression covering all missense variants confirmed that PrimateAI had the lowest performance with an auROC of 0.5078, making it barely capable of differentiating pathogenic classes. It is important to note that ESM1b and AlphaMissense performed equally poorly, although AlphaMissense had a slightly higher auROC than the other VEPs.

Statistical significance was determined using -log10 p-values from ExWAS, where p-values less than 7 were excluded from analysis. After filtering for quality control and statistical significance, patterns identified pertaining to all missense variants were

upheld. Mean BETA decreases as MAF decreases, reflecting a stronger loss of function effect with decreasing MAF. The mean BETA for PrimateAI is still close for pathogenic and nonpathogenic variants, indicating that PrimateAI performs poorly in differentiating pathogenic variants based on predicted effect. PrimateAI called significantly more nonpathogenic variants than any other VEP, meaning that it likely is missing true pathogenic variants. Looking at the distribution of BETA for each VEP, the mean effect for AlphaMissense predicted pathogenic variants is lower than the other VEPs, meaning the predicted pathogenic variants are more likely to be truly pathogenic. Logistic Regression composing of statistically significant variants outputted similar results as that of regression with all variants. All 3 VEPs struggled with accurate predictions when specifying haploinsufficient genes and known recessive genes. This should be interpreted with caution, as there is a lower number of genes and variants included in investigation. The patterns detailed through the analysis are still observable in both of these categories.

Enrichment analysis was utilized to investigate protein families and domains, as well as to identify significant, trait-associated genes and prioritize associated variants for use in comprehensive VEP comparison. Only rare variants with MAF < 0.1% were included, and variants with positive and negative BETA values were interrogated separately. The negative variants were strongly associated with loss of functions in regulatory processes, namely proteolysis and peptidase activity. Immunoglobulin domains were strongly impacted by missense variants, regardless of effect. This was expected, as the proteins investigated in this analysis are plasma proteins. For the most statistically significant term of each library, the variants of the pertinent genes were investigated for finalizing VEP comparisons. A majority of these variants were not previously identified. As a result, experimental characterization is required to determine which VEP is accurate. In almost every case, there was discrepancy between pathogenic and nonpathogenic predictions, and there was seemingly no consistent pattern. It appears that a major source of the discrepancy is the scoring by which each VEP uses to make predictions. All 3 VEPs consistently and accurately identified benign variants. A small number of variants were previously identified as pathogenic, but there was still some discrepancy between VEP predictions, namely by PrimateAI.

The results of this analysis have demonstrated the extent of discrepancy in predictions made by VEPs, even with an unbiased framework. Before making conclusions, there are some important limitations to address. First, there were issues with the preliminary association testing, causing delays in analysis and restricting timelines. This analysis was highly limited by time and technical constraints, resulting in the pursuance of simplified methodologies. In terms of the VEPs themselves, all VEPs use different scoring schemes to predict pathogenicity, and the thresholds are determined experimentally, based on internal performance metrics. This alone complicates any comparisons, because it amplifies the likelihood of a discrepancy between variant predictions. Specific to PrimateAI, this VEP has been updated significantly as compared to what is available in dbNSFP. The other VEPs are more up to date, but a crucial limitation of this analysis is the simplification of AlphaMissense. AlphaMissense uses 3 classes in its predictions, where a variant is either Pathogenic, Benign, or Ambiguous,

depending on the score. For this analysis, ambiguous predictions were considered likely benign. This reduces the observed performance capability, but due to time and technical restrictions, a simplification was necessary.

Conclusion

This analysis highlights the necessity of unbiased frameworks in benchmarking and performance evaluation of computational VEPs, due to underestimated discrepancies between pathogenicity predictions. Computational tools assist in interpreting missense variants and provide experimental evidence for variant classification. These tools will be pivotal in achieving the NHGRI goal of classifying all protein variants, but it is essential that they are accurate. If VEPs are to be used clinically, there must be consistency in predictions of pathogenicity, as conflicting evidence only complicates use of protein variants in the diagnosis and treatment of genetic disease. In terms of benchmarking, based on this analysis, PrimateAI is the weakest performing VEP. AlphaMissense and ESM1b perform similarly, with a slight advantage towards AlphaMissense.

References:

1. Joynt ACM, Axford MM, Chad L, Costain G. Understanding genetic variants of uncertain significance. Paediatr Child Health. 2021 Sep 13;27(1):10-11. doi: 10.1093/pch/pxab070. PMID: 35273666; PMCID: PMC8900699.
2. Richards, Sue, et al. "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 (2015): 405-423.
3. Fowler, D. M., & Rehm, H. L. (2023). Will variants of uncertain significance still exist in 2030?. The American Journal of Human Genetics.
4. Chen E, Facio FM, Aradhya KW, et al. Rates and Classification of Variants of Uncertain Significance in Hereditary Disease Genetic Testing. JAMA Netw Open. 2023;6(10):e2339571. doi:10.1001/jamanetworkopen.2023.39571
5. https://www.cancer.gov/publications/dictionaries/genetics-dictionary/def/missense-variant
6. Qorri, E.; Takács, B.; Gráf, A.; Enyedi, M.Z.; Pintér, L.; Kiss, E.; Haracska, L. A Comprehensive Evaluation of the Performance of Prediction Algorithms on Clinically Relevant Missense Variants. Int. J. Mol. Sci. 2022, 23, 7946. https://doi.org/10.3390/ijms23147946
7. Sundaram, L., Gao, H., Padigepati, S.R. et al. Predicting the clinical impact of human mutation with deep neural networks. Nat Genet 50, 1161–1170 (2018). https://doi.org/10.1038/s41588-018-0167-z
8. Brandes, N., Goldman, G., Wang, C.H. et al. Genome-wide prediction of disease variant effects with a deep protein language model. Nat Genet 55, 1512–1522 (2023). https://doi.org/10.1038/s41588-023-01465-0
9. Cheng, J., Novati, G., Pan, J., Bycroft, C., Žemgulytė, A., Applebaum, T., ... & Avsec, Ž. (2023). Accurate proteome-wide missense variant effect prediction with AlphaMissense. Science, 381(6664), eadg7492.
10. Livesey, B. J., & Marsh, J. A. (2020). Using deep mutational scanning to benchmark variant effect predictors and identify disease mutations. Molecular systems biology, 16(7), e9380.
11. Livesey, B. J., & Marsh, J. A. (2023). Updated benchmarking of variant effect predictors using deep mutational scanning. Molecular Systems Biology, e11474.
12. Sun, B. B., Chiou, J., Traylor, M., Benner, C., Hsu, Y. H., Richardson, T. G., ... & Whelan, C. D. (2023). Plasma proteomic associations with genetics and health in the UK Biobank. Nature, 1-10.
13. https://biobank.ndph.ox.ac.uk/showcase/label.cgi?id=170
14. https://biobank.ndph.ox.ac.uk/showcase/label.cgi?id=1839
15. https://hail.is/
16. https://rgcgithub.github.io/regenie/
17. Mbatchou, J., Barnard, L., Backman, J. et al. Computationally efficient whole-genome regression for quantitative and binary traits. Nat Genet 53, 1097–1103 (2021). https://doi.org/10.1038/s41588-021-00870-7

18. http://database.liulab.science/dbNSFP
19. Dhindsa, R.S., Burren, O.S., Sun, B.B. et al. Rare variant associations with plasma protein levels in the UK Biobank. Nature 622, 339–347 (2023). https://doi.org/10.1038/s41586-023-06547-x
20. Shihab HA, Rogers MF, Campbell C, Gaunt TR. HIPred: an integrative approach to predicting haploinsufficient genes. Bioinformatics. 2017 Jun 15;33(12):1751-1757. doi: 10.1093/bioinformatics/btx028. PMID: 28137713; PMCID: PMC5581952.
21. MacArthur DG, Balasubramanian S, Frankish A, Huang N, Morris J, Walter K, Jostins L, Habegger L, Pickrell JK, Montgomery SB, Albers CA, Zhang ZD, Conrad DF, Lunter G, Zheng H, Ayub Q, DePristo MA, Banks E, Hu M, Handsaker RE, Rosenfeld JA, Fromer M, Jin M, Mu XJ, Khurana E, Ye K, Kay M, Saunders GI, Suner MM, Hunt T, Barnes IH, Amid C, Carvalho-Silva DR, Bignell AH, Snow C, Yngvadottir B, Bumpstead S, Cooper DN, Xue Y, Romero IG; 1000 Genomes Project Consortium; Wang J, Li Y, Gibbs RA, McCarroll SA, Dermitzakis ET, Pritchard JK, Barrett JC, Harrow J, Hurles ME, Gerstein MB, Tyler-Smith C. A systematic survey of loss-of-function variants in human protein-coding genes. Science. 2012 Feb 17;335(6070):823-8. doi: 10.1126/science.1215040. Erratum in: Science. 2012 Apr 20;336(6079):296. PMID: 22344438; PMCID: PMC3299548.
22. Chen EY, Tan CM, Kou Y, Duan Q, Wang Z, Meirelles GV, Clark NR, Ma'ayan A. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. BMC Bioinformatics. 2013; 128 (14).
23. Gunning, Adam C., et al. "Assessing performance of pathogenicity predictors using clinically relevant variant datasets." Journal of medical genetics 58.8 (2021): 547-555.