DeepSVP: Integration of genotype and phenotype for structural variant prioritization using deep learning (Supplementary materials)

Azza Althagafi[1,2], Lamia Alsubaie[3], Nagarajan Kathiresan[4], Katsuhiko Mineta[1], Taghrid Aloraini[3], Fuad Almutairi[5,6], Majid Alfadhel[5,6,7], Takashi Gojobori[8], Ahmad Alfares[3,7,9], and Robert Hoehndorf[1,*]

[1] Computer, Electrical and Mathematical Sciences & Engineering Division, Computational Bioscience Research Center (CBRC), King Abdullah University of Science and Technology, 4700 King Abdullah University of Science and Technology, Thuwal 23955-6900, Saudi Arabia, [2] Computer Science Department, College of Computers and Information Technology, Taif University, Taif 26571, Saudi Arabia., [3] Department of Pathology and Laboratory Medicine, King Abdulaziz Medical City, Riyadh, Saudi Arabia. , [4] Supercomputing Core Lab, King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Saudi Arabia., [5] Division of Genetics, Department of Pediatrics, King Abdulaziz Medical City, Riyadh, Saudi Arabia., [6] King Saud bin Abdulaziz University for Health Sciences, King Abdulaziz Medical City, Riyadh, Saudi Arabia., [7] King Abdullah International Medical Research Center, Riyadh, Saudi Arabia., [8] Biological and Environmental Science and Engineering Division, Computational Bioscience Research Center, King Abdullah University of Science and Technology (KAUST), Thuwal, 23955-6900, Saudi Arabia., [9] Department of Pediatrics, College of Medicine, Qassim University, Qassim, Saudi Arabia., * robert.hoehndorf@kaust.edu.sa

1 Data sources and ontologies

The training dataset of variants contains 14,197 (10,401 deletions, 3,796 duplications) pathogenic or likely pathogenic structural variants and 4,477 variants associated with one or more diseases (3,737 deletions, 586 duplications), as well as 25,890 (13,742 deletions, 12,148 duplications) benign or likely benign structural variants.

For each pathogenic structural variant, we defined variant–disease pairs with associated diseases from Online Mendelian Inheritance in Men (OMIM) database [1]. There are 3,805 structural variants linked with one or more than one OMIM disease; if a variant is associated with n OMIM diseases, we generate n variant–disease pairs. As a result, we obtained 5,907 causative pathogenic variant–disease pairs. As negative training pairs we selected both benign and pathogenic variants and associate them with a randomly selected disease; we include pathogenic variants in the negative training pairs to simulate the case where variants may be pathogenic but not associated with the phenotypes observed in a patient (i.e., with a different phenotypes). After this step we obtained 36,041 negative (not causative) variant–disease pairs .

We downloaded phenotypes from the HPO database on 16 July 2020 and obtain 169,281 phenotype associations for 4,315 human genes. We downloaded phenotype associations from MGI on 20 March 2020 and obtained 228,214 associations between 13,529 mouse genes and classes in MP. We identify the human ortholog for mouse genes using the `HMD_HumanPhenotype.rpt` orthology file from MGI; the file contains 10,951 human orthologs for the 13,529 mouse genes resulting in 168,550 associations between human genes and MP classes. We obtain the GO annotations for human gene products from the GO Annotation database [2] on 20 March 2020 for 18,495 gene products with 495,719 annotations in total. We filtered out all the GO annotations with the evidence code indicating that the annotation was inferred from electronic annotation (IEA), or no biological data is available (ND). We map the UniProt accessions of the gene products to Entrez gene identifiers using the mappings provided by the Entrez database [3], and obtain 17,786 genes which have GO annotations for their gene product resulting in 208,630 associations between genes and GO classes. For the anatomical location of gene expression, we downloaded the GTEx Tissue Expression Profiles from the Gene Expression Atlas [4] which identifies gene expression across 53 tissues; 20,538 genes have an expression above the 4.0 threshold which was previously determined to be useful for predicting disease associations [5] resulting in 585,765 associations between genes and UBERON classes. We represent the tissues with their UBERON classes, excluding the tissues *transformed skin fibroblast* and *EBV-transformed lymphocyte* as these are not found in UBERON. For the cell type, we downloaded single-cell RNAseq data from the Tabula Muris project [6] in which genes are annotated with the CL. From this dataset, we obtain 6,559 human genes which have CL annotations, and 17,149 associations between genes and one or more classes from CL.

We use the combined PhenomeNET ontology [7] downloaded on 6 October, 2020, from the AberOWL ontology repository, as our phenotype ontology. PhenomeNET combines the phenotypes of human and other model organism as well as UBERON, GO and CL, and allows them to be compared.

2 Estimating variant pathogenicity by supervised prediction

2.1 Phenotype prediction model

We apply DL2Vec to generate the feature representation for the patient phenotypes and genes. DL2Vec learns the "representation" for phenotypes and genes based on their annotations to ontology classes. The inputs to DL2Vec are associations of entities with ontology classes and the outputs are vectors (embeddings) of these entities. DL2Vec utilizes the axioms in ontologies to construct a graph representing phenotypes and their interrelations. DeepSVP incorporates biological background knowledge about the relation between phenotypes resulting from a loss of function in mouse/human genes, gene functions as defined using the GO, as well as the celltype and anatomical site of gene expression.

The phenotype model takes two vectors v_1 and v_2 as input, representing the embedding for the patient's phenotypes and the embedding for a gene, respectively. The embeddings are used as input for two neural network models ν_1 and ν_2. We then calculate the inner product for $\nu_1(v_1)$ and $\nu_2(v_2)$ and apply a sigmoid activation function to generate a prediction score, between the embedding for the phenotypes v_1, and gene v_2. We use binary cross-entropy as a loss function to train our model defined as:

$$Loss = -\frac{1}{N}\sum_{i=1}^{N} Y_i \cdot \log(P(Y_i)) + (1 - Y_i) \cdot \log(1 - P(Y_i)) \tag{1}$$

where N correspond to the number of training samples, Y_i is the true value for sample i, and $P(Y_i)$ is the predicted value for sample i.

Each neural network ν_1 and ν_2 consists of two hidden layers, in which the first layer with 256 units, and the second layer with 50 units. After each layer, we use dropout [8] with a rate of 20%, followed by a Leaky Rectified Linear Unit (LeakyReLU) [9] activation function. We use the Adam optimizer [10] to optimize the model parameters. We develop five different models using DL2Vec embeddings based on different feature types: functions of gene products (GO), mouse model phenotypes (MP), human phenotype (HP), celltype (CL), and site of expression

(UBERON). For each set of phenotypes (characterizing a disorder, or the clinical phenotypes observed in an individual), and for each prediction model, we rank each gene based on the DL2Vec prediction score, from smallest to highest, and represent the association between them by their m-quantile in this distribution [11] (Figure 5 shows the distribution of the normalized quantile scores). This normalization and ranking aims to make prediction scores comparable across sets of phenotypes [12]. We use the quantile as one of the features of the combined prediction models.

2.2 Combined prediction model

The combined prediction model uses the variant features and the phenotype-based scores produced by the DL2Vec-based predictions; the model is an artificial deep neural network model that uses genomic features derived from a variant as input together with the prediction score generated from the phenotype prediction model. We trained a separate model for each ontology dataset and aggregation type, either the maximum or average features scores for the genes within the variant region. The features used by the combined model are listed in the Supplementary Table 2.

Given a structural variant τ affecting regions that contain genes $G_1, ..., G_n$, we obtain the phenotype prediction score $\phi(G_i)$ for the genes $G_1, ..., G_n$ using each of the phenotype-based prediction models. We transform these scores into a feature for the variant τ using either the maximum or average of all the gene scores, i.e., either $\phi_{max}(\tau) = \max_{1 \leq i \leq n} \phi(G_i)$ or $\phi_{avg}(\tau) = \frac{1}{n} \sum_{i=1}^{n} \phi(G_i)$. We normalize all the features using z-score normalization, in which the values for and feature F are scaled based on the mean, and standard deviation of F. The value v_i of F is normalized to v_i' by $v_i' = \frac{v_i - \mu_F}{\sigma_F}$ where v_i' is z-score normalized one values, v_i is the i-th value for the feature A, μ_F is the mean, and σ_F the standard deviation, for feature F. We use the same mean and standard deviation to normalize the testing set.

We use 22 features for variants (8 features for the variant and 9 derived from the genes overlapping the variant) as well as 5 features from ontology embeddings. Some features are missing for some variants; to account for missing values, we use imputation. We imputed missing val-

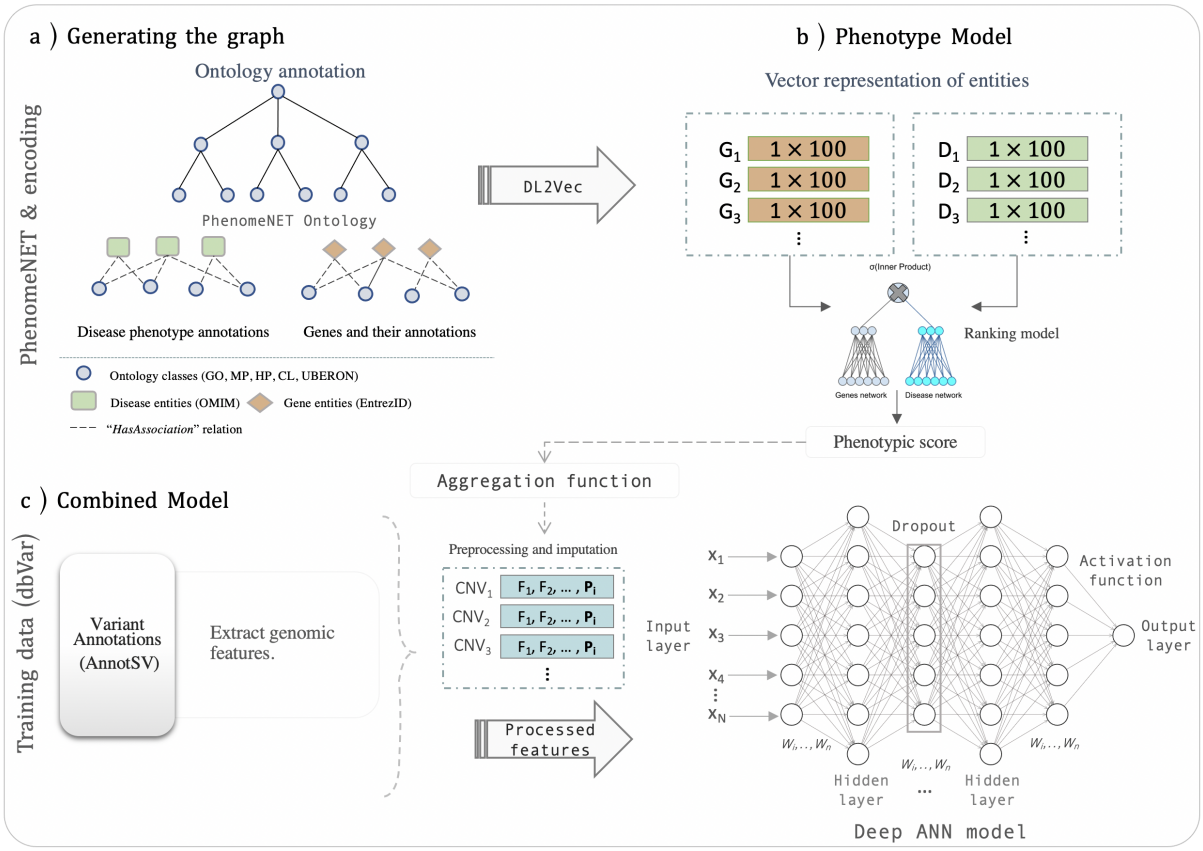

Figure 1: Workflow for DeepSVP training. (**a**) Generate the graph using the ontology annotations. (**b**) Generate the embeddings after generating the walks on the graph, then run the DL2vec prediction model to rank the genes for every disease using the OMIM disease and genes embeddings. (**c**) Training the combined model, by first Collecting the genomic features derived from the training data and the DL2vec prediction score for the genes within the variants and the associated diseases. Abbreviations: `G`: Genes, `D`: Disease, `F`: Features, `P`: Phenotypic score, `VCF`: Variant Call Format.

ues by assigning them a zero value and additionally created indicator variables with a value set to 1 if the corresponding variant is missing and 0 otherwise. We use one-hot encoding to represent the categorical features with an "undefined" category for missing categorical annotations. We provide an analysis of feature importance and the correlation between features in the Supplementary Materials Section 3.1.

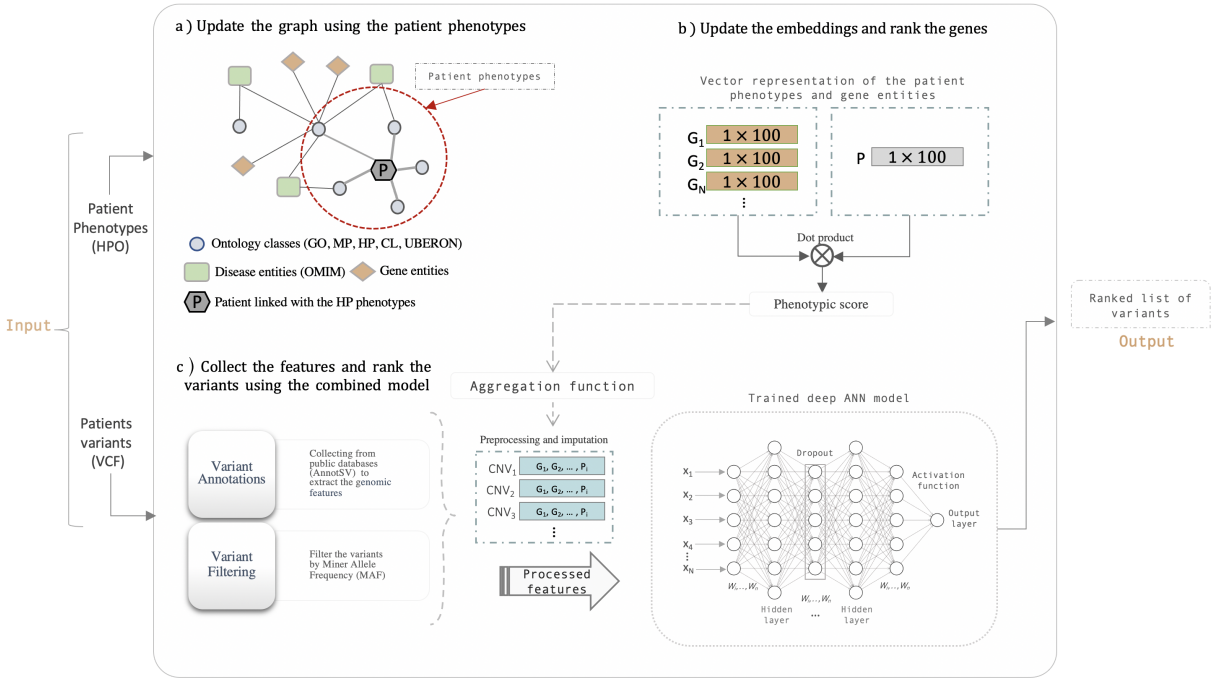

Figure 2: Workflow for DeepSVP inference. (**a**) Update the graph using the patient phenotypes by adding a node for the patient and edges to the set of phenotypes observed in the patient. (**b**) Update the embeddings after generating the walks on the updated graph starting by the patient node, then run the DL2vec prediction model to rank the genes for the patient phenotypes using the patient and genes embeddings. (**c**) Collect the genomic features derived from the patient VCF and rank the variants using the DeepSVP combined prediction model, the rank determines how likely the variant is causative of the phenotypes observed in the patient. Abbreviations: `G`: Genes, `P`: Patient, `F`: Features, `P`: Phenotypic score.

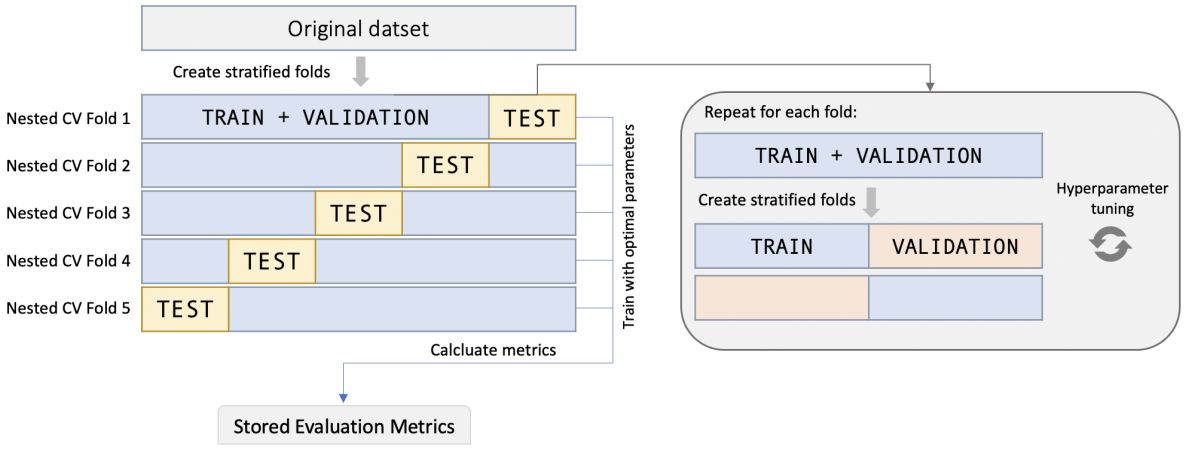

Figure 3: Workflow diagram of the model's training using nested Cross-validation

2.3 Training and testing

We tuned for the following set of hyperparameters for the model: learning-rate $r \in [1 \times 10^{-6}, 1 \times 10^{-2}]$ with logarithmic transformation, and dropout $r \in [0.1, 0.5]$; number of layers $l \in [2, 6]$, and the number of nodes for each of the dense layers $n \in [50, 512]$; activation functions $a \in \{\text{relu}, \text{sigmoid}, \text{selu}\}$ between the layers. Following our experiments, the optimal parameters for each model are summarize in Table 1. We used a nested 5×2 fold cross validation for training (Figure 3).

		Learning rate	Dense layers	Dense nodes	dropout	Activation
DeepSVP models using maximum score	GO	0.0006	6	50	0.1000	Relu
	MP	0.0002	3	146	0.1555	
	HP	0.0007	3	347	0.1000	
	CL	0.0001	2	212	0.1000	
	UBERON	0.0003	5	212	0.2033	
	Union	0.0002	6	120	0.1000	
DeepSVP models using average score	GO	0.0023	4	344	0.3327	
	MP	0.0100	4	484	0.1000	
	HP	0.0001	5	320	0.1498	
	CL	0.0015	6	201	0.1000	
	UBERON	0.0100	4	512	0.1000	
	Union	0.0002	3	512	0.1110	

Table 1: Optimized models parameters values

3 Variant-based features

We use AnnotSV v2.3 [13], which uses data from multiple external databases to annotate and rank SVs based on the overlapping regions of the variants with known pathogenic variants in dbVar [14], the Database of Genomic Variants (DGV) [15], and disease-associated genes from OMIM. For each variant, AnnotSV generates annotations based on the variant length and the genes with which the variant overlaps (choosing among Refseq [16] gene annotations). Furthermore, AnnotSV reports the list of promoters with which the variant overlaps. From the annotations provided by AnnotSV, we use the variant length, variant type, GC content around the variant's breakpoints (GCcontent_left, GCcontent_right), and the number of promoters and genes affected by the variant as features.

We further use AnnotSV to obtain information about genes with which a structural variant overlaps: the length of the Coding DNA Sequence (CDS), transcript length (tx length), haploinsufficiency ranks collected from the Deciphering Developmental Disorders (DDD) study [17], haploinsufficiency (HI_DDDpercent) and triplosensitivity estimates from ClinGen [18], gene intolerance annotations from ExAC [19] including six annotations for synonymous variants (synZExAC), missense variants (misZExAC), loss of function variants (pLIExAC), deletion (delZExAC), duplications (dupZExAC), and CNV intolerance (cnvZExAC). Table 2 summarizes all the features used in our predictive model.

While not used as a feature of our prediction model, we also use AnnotSV to identify the allele frequency of variants using the 1,000 genomes allele frequency [20] and allele frequency from gnomAD [21]. We use this information to filter out common variants before applying our prediction.

3.1 Correlation between features

To test the redundancy of different scores of structural variants corresponding to the susceptibility of the phenotypes for the disease, we analyze the features by evaluating the pairwise correlation between them (Figure 4). According to their correlation coefficients, the features

were broadly clustered into four major groups using the maximum and average score. As expected, measures of the structure of variants such as SV length and the number of genes or promoters were highly correlated. We further assess the importance of the features using two methods; the first using Extremely Randomized Trees ensemble learning Classifier (Extra Trees Classifier) (Figure 6), and the second using the Shapiro-Wilk algorithm [22] (Figure 7). Both methods explore the linear relationships across features; Extra Trees Classifier aggregates the results of multiple decision trees and outputs the ranked features based on the information gain, while Shapiro-Wilk assesses the normality of the distribution of examples with respect to the feature. We noticed that the features rank using both methods are similar; however, both do not capture potential nonlinear relationships among features, so we included all the ranked features in our model.

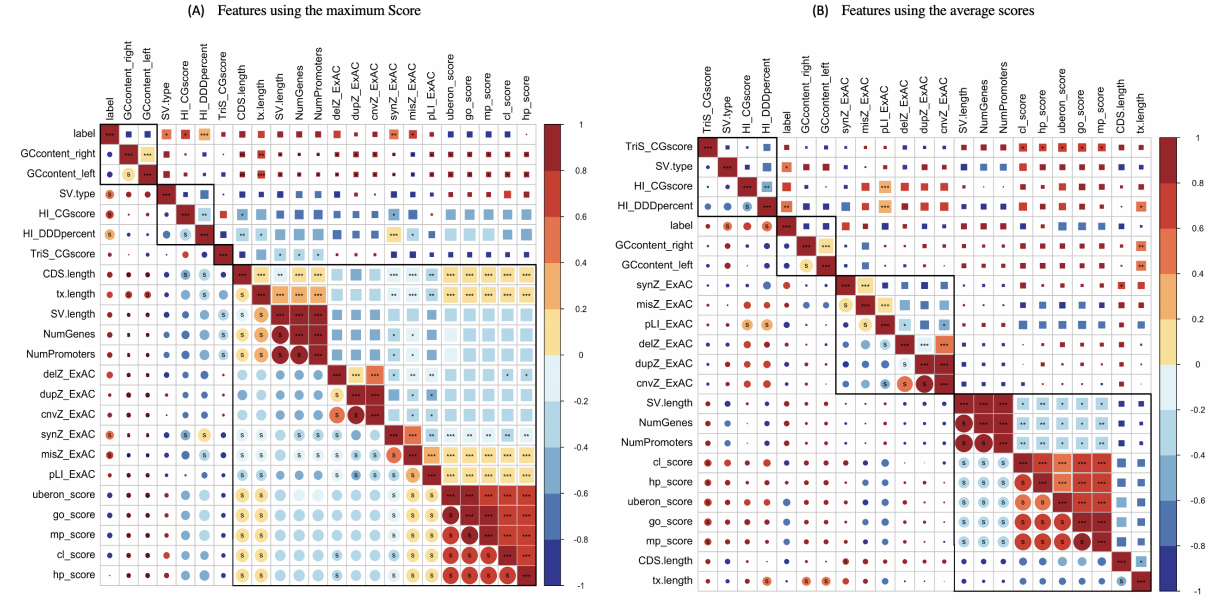

Figure 4: Correlation between the combined model features generated with *corrplot* package (version 0.84) in R, using *corrplot* function. 22 features of 42,202 SVs, (8 features for the variant, and 14 derived from the genes overlapping the variant) were obtained from the disease phenotypes and AnnotSV using various databases. The pairwise correlation was computed on all the features. Figure **A** shows results using the maximum score, and Figure **B** the average score. The features are ordered, and different clusters are highlighted based on the hierarchical clustering. Significant correlations ($P < 0.05$) are indicated by a letter 's' in the lower triangle. The color and size of circles represent the correlation strength (correlation coefficient). Statistical significance is indicated with asterisks (*P <0.05;** P <0.01;*** P <0.001).

	Feature	Description
Structural Variant (SV)	SV length	Length of SV.
	SV type	Type of the SV (DEL, DUP) Categorical features.
	GCcontent_left*	Breakpoints annotations, GC content around the left SV breakpoint ($\pm$ 100bp).
	GCcontent_right*	Breakpoints annotations, GC content around the right SV breakpoint ($\pm$ 100bp).
	Number of genes*	Number of genes within the SVs.
	Number of promoters*	Number of promoters within the SVs.
	HI_CGscore	HaploInsufficiency Score (categorical features).
	TriS_CGscore	TriploSensitivity Score (categorical feature).
Genes-based annotations	DL2vec_Score*	Predict associations between genes and sets of phenotypes different ontolgies (GO, HP, CL, MP, and UBERON).
	CDS length*	Length of the CoDing Sequence (CDS) (bp) overlapping with the SV.
	tx length*	Length of transcript (bp) overlapping with the SV.
	synZ_ExAC*	Gene intolerance to synonymous variation.
	misZ_AxAC*	Gene intolerance to missense variation.
	pLI_ExAC*	The probability that a gene is intolerant to a loss of function variation.
	dupZ_ExAC*	Gene duplication intolerance.
	delZ_ExAC*	Gene deletion intolerance.
	cnvZ_ExAC*	Gne CNV intolerance.
	HI_DDDpercent*	Haploinsufficiency ranks, where in a single functional copy of a gene is insufficient to maintain normal function.

Table 2: Annotation features for model training and prediction. An asterisk (*) indicates that a boolean indicator variable was created in order to handle undefined values for that feature.

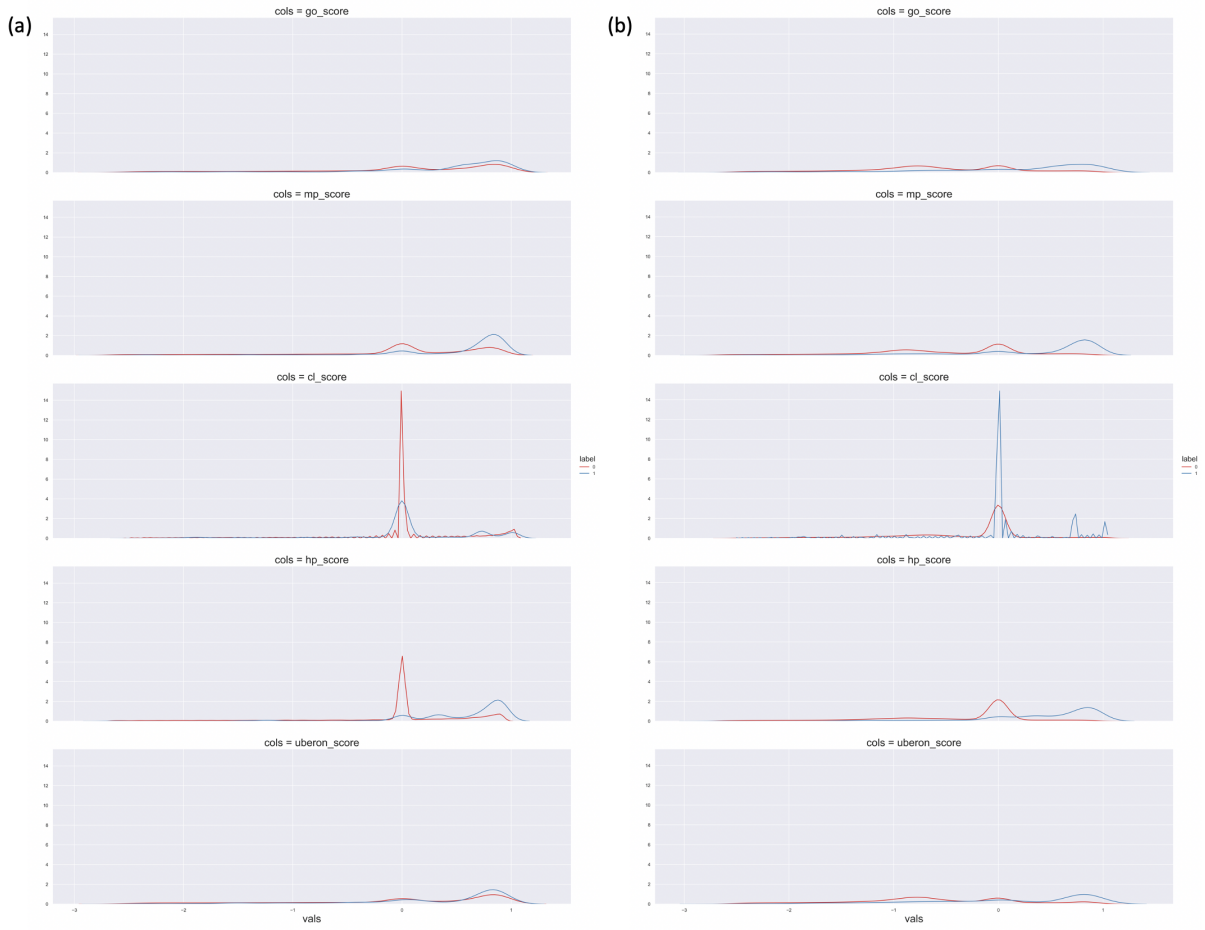

Figure 5: Distribution of the phenotype features, **(a)** using the maximum features scores, and **(b)** using the average score.

Aggregation method for genes within CNV	DeepSVP Models	F1-Score	ROCAUC	PRAUC	DOR
Maximum score	GO	0.8245 (± 0.0211)	0.9090 (± 0.0120)	0.9103 (± 0.0080)	26.1573 (± 5.5995)
	MP	0.8273 (± 0.0180)	0.9122 (± 0.0128)	0.9160 (± 0.0102)	26.9852 (± 6.2736)
	HP	0.8748 (± 0.0576)	0.9509 (± 0.0277)	0.9540 (± 0.0260)	85.3049 (± 50.2101)
	CL	0.8093 (± 0.0486)	0.9031 (± 0.0326)	0.9042 (± 0.0316)	24.3906 (± 11.8240)
	UBERON	0.7937 (± 0.0523)	0.8936 (± 0.0241)	0.8959 (± 0.0256)	19.8162 (± 7.5107)
	Union	0.8822 (± 0.0234)	0.9509 (± 0.0154)	0.9552 (± 0.0119)	74.3240 (± 23.1871)
Average score	GO	0.8241 (± 0.0281)	0.9112 (± 0.0140)	0.9138 (± 0.0099)	27.2239 (± 6.4774)
	MP	0.8372 (± 0.0299)	0.9259 (± 0.0155)	0.9266 (± 0.0154)	38.1741 (± 12.3755)
	HP	0.8628 (± 0.0267)	0.9400 (± 0.0164)	0.9443 (± 0.0138)	53.2996 (± 21.4353)
	CL	0.8327 (± 0.0489)	0.9184 (± 0.0302)	0.9183 (± 0.0318)	31.0700 (± 12.8467)
	UBERON	0.8258 (± 0.0308)	0.9084 (± 0.0286)	0.9071 (± 0.0310)	28.4376 (± 10.9040)
	Union	**0.9142 (± 0.0218)**	**0.9693 (± 0.0129)**	**0.9700 (± 0.0122)**	**144.0755 (± 67.6037)**

Table 3: Summary of the evaluation for predicting causative variants in our testing data set using nested 5 folds cross-validation. We reported the mean for the different evaluation metrics along with the standard deviation.

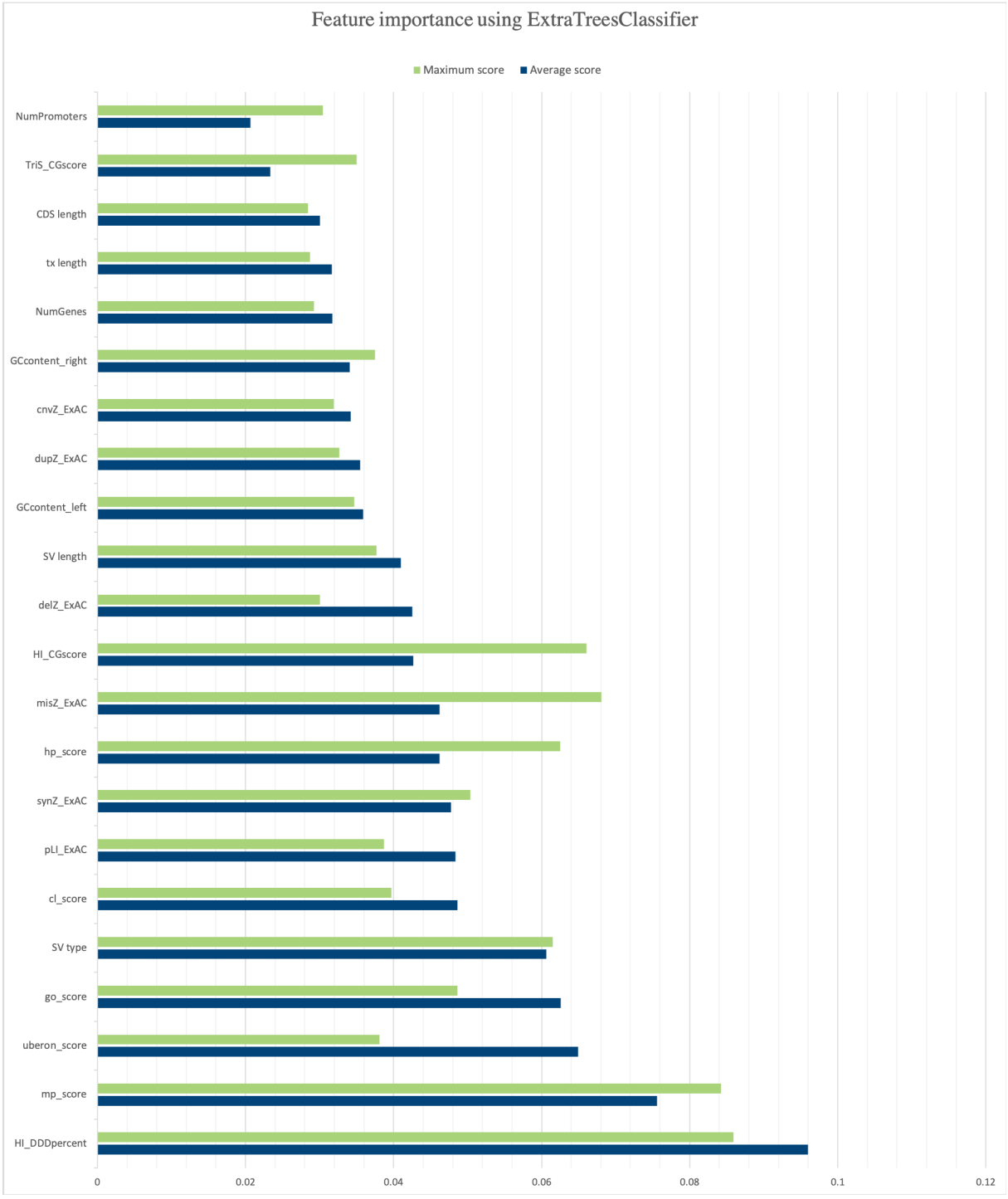

Figure 6: Feature importance using ensemble learning technique Extremely Randomized Trees Classifier (Extra Trees Classifier) that aggregates the results of multiple decision trees and output the ranked features based on the information gain.

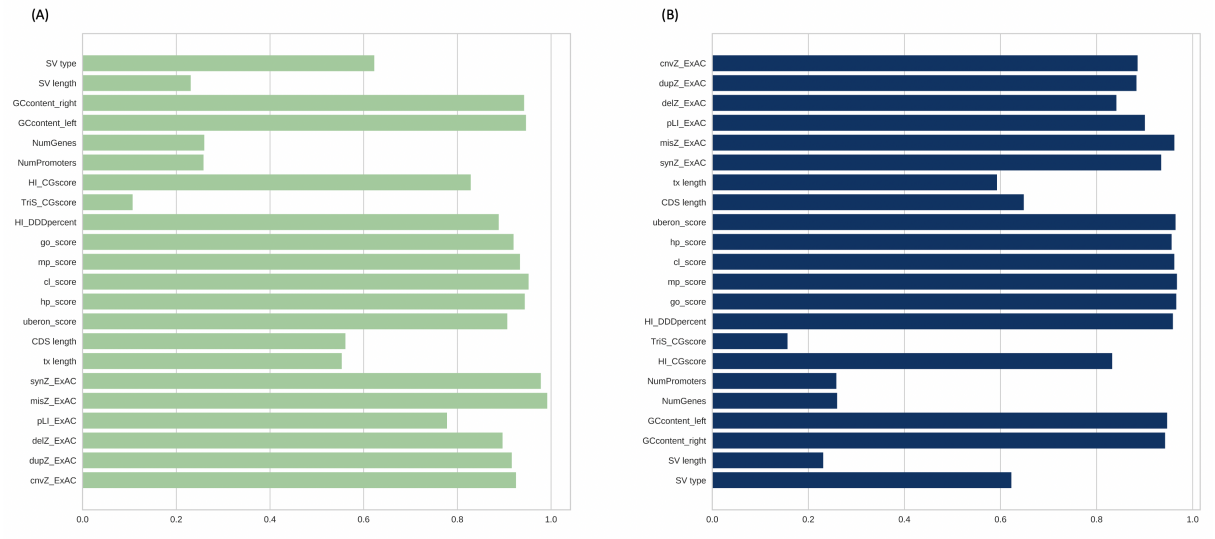

Figure 7: A one-dimensional ranking of features utilizing the Shapiro-Wilk algorithm generated using *Yellowbrick* Python package (version 1.0). The Shapiro algorithm takes into account a single feature at a time to assess the normality of the distribution of instances with respect to the feature. A barplot showing the relative ranks of each feature **(A)** using the maximum features scores, and **(B)** using the average score.

		Recall@1	Recall@10	Recall@30	ROCAUC	PRAUC
DeepSVP models using average score	**GO**	451 (0.3001)	628 (0.4178)	787 (0.5236)	0.9650	0.3441
	MP	633 (0.4212)	1023 (0.6806)	1232 (0.8197)	0.9851	0.5099
	HP	515 (0.3426)	941 (0.6261)	1188 (0.7904)	0.9837	0.4347
	CL	137 (0.0912)	543 (0.3613)	885 (0.5888)	0.9742	0.1657
	UBERON	234 (0.1557)	600 (0.3992)	1106 (0.7359)	0.9758	0.2280
	Union	**679 (0.4518)**	**1060 (0.7053)**	1250 (0.8317)	0.9859	**0.5441**
DeepSVP models using maximum score	**GO**	328 (0.2182)	521 (0.3466)	726 (0.4830)	0.9557	0.2678
	MP	237 (0.1577)	626 (0.4165)	878 (0.5842)	0.9602	0.2461
	HP	410 (0.2728)	1040 (0.6919)	**1340 (0.8916)**	**0.9936**	0.4111
	CL	275 (0.1830)	823 (0.5476)	1146 (0.7625)	0.9801	0.2575
	UBERON	247 (0.1643)	648 (0.4311)	1055 (0.7019)	0.9737	0.2377
	Union	321 (0.2136)	953 (0.6341)	1130 (0.7518)	0.9758	0.3474
SV pathogenicity prediction/ranking	**StrVCTVRE**	72 (0.0479)	223 (0.1484)	405 (0.2695)	0.9178	0.0952
	CADD-SV	38 (0.0253)	620 (0.4125)	1020 (0.6786)	0.9816	0.1262
	AnnotSV	19 (0.0126)	229 (0.1524)	700 (0.4657)	0.9605	0.2203

Table 4: Summary of the evaluation for predicting causative variants in the synthetic whole-genome benchmark dataset derived from dbVar, using 90% of the phenotypes. AnnotSV are computed based on ranking variants from "pathogenic" to"benign". We break the ties uniformly for all the methods randomly and report the absolute number of variants recovered at each rank together with the recall, as well as areas under the ROC curve (using micro-averages per synthetic genome) and precision-recall curve. Best performing results (using maximum or average score) for each measure are indicated in bold.

		Recall@1	Recall@10	Recall@30	ROCAUC	PRAUC
DeepSVP models using average score	**GO**	435 (0.2894)	619 (0.4118)	799 (0.5316)	0.9643	0.3320
	MP	639 (0.4251)	1047 (0.6966)	1226 (0.8157)	0.9850	0.5153
	HP	529 (0.3520)	962 (0.6401)	1200 (0.7984)	0.9816	0.4445
	CL	157 (0.1045)	543 (0.3613)	882 (0.5868)	0.9739	0.1758
	UBERON	241 (0.1603)	602 (0.4005)	1102 (0.7332)	0.9756	0.2315
	Union	**676 (0.4498)**	1056 (0.7026)	1250 (0.8317)	0.9859	**0.5428**
DeepSVP models using maximum score	**GO**	330 (0.2196)	520 (0.3460)	716 (0.4764)	0.9555	0.2688
	MP	233 (0.1550)	639 (0.4251)	884 (0.5882)	0.9607	0.2487
	HP	431 (0.2868)	**1059 (0.7046)**	**1338 (0.8902)**	**0.9936**	0.4283
	CL	273 (0.1816)	829 (0.5516)	1151 (0.7658)	0.9800	0.2568
	UBERON	242 (0.1610)	645 (0.4291)	1062 (0.7066)	0.9735	0.2342
	Union	323 (0.2149)	953 (0.6341)	1128 (0.7505)	0.9756	0.3476
SV pathogenicity prediction/ranking	**StrVCTVRE**	72 (0.0479)	223 (0.1484)	405 (0.2695)	0.9178	0.0952
	CADD-SV	38 (0.0253)	620 (0.4125)	1020 (0.6786)	0.9816	0.1262
	AnnotSV	19 (0.0126)	229 (0.1524)	700 (0.4657)	0.9605	0.2203

Table 5: Summary of the evaluation for predicting causative variants in the synthetic whole-genome benchmark dataset derived from dbVar, using 80% of the phenotypes. AnnotSV are computed based on ranking variants from "pathogenic" to"benign". We break the ties uniformly for all the methods randomly and report the absolute number of variants recovered at each rank together with the recall, as well as areas under the ROC curve (using micro-averages per synthetic genome) and precision-recall curve. Best performing results (using maximum or average score) for each measure are indicated in bold.

		Recall@1	Recall@10	Recall@30	ROCAUC	PRAUC
DeepSVP models using average score	**GO**	458 (0.3047)	628 (0.4178)	792 (0.5269)	0.9643	0.3453
	MP	647 (0.4305)	1029 (0.6846)	1223 (0.8137)	0.9848	0.5180
	HP	529 (0.3520)	942 (0.6267)	1204 (0.8011)	0.9803	0.4433
	CL	158 (0.1051)	543 (0.3613)	885 (0.5888)	0.9742	0.1748
	UBERON	245 (0.1630)	599 (0.3985)	1103 (0.7339)	0.9759	0.2338
	Union	**676 (0.4498)**	**1058 (0.7039)**	1244 (0.8277)	0.9858	**0.5429**
DeepSVP models using maximum score	**GO**	335 (0.2229)	531 (0.3533)	713 (0.4744)	0.9550	0.2730
	MP	243 (0.1617)	631 (0.4198)	882 (0.5868)	0.9597	0.2519
	HP	471 (0.3134)	1045 (0.6953)	**1331 (0.8856)**	**0.9930**	0.4389
	CL	273 (0.1816)	822 (0.5469)	1140 (0.7585)	0.9801	0.2565
	UBERON	245 (0.1630)	651 (0.4331)	1049 (0.6979)	0.9733	0.2362
	Union	325 (0.2162)	947 (0.6301)	1128 (0.7505)	0.9754	0.3480
SV pathogenicity prediction/ranking	**StrVCTVRE**	72 (0.0479)	223 (0.1484)	405 (0.2695)	0.9178	0.0952
	CADD-SV	38 (0.0253)	620 (0.4125)	1020 (0.6786)	0.9816	0.1262
	AnnotSV	19 (0.0126)	229 (0.1524)	700 (0.4657)	0.9605	0.2203

Table 6: Summary of the evaluation for predicting causative variants in the synthetic whole-genome benchmark dataset derived from dbVar, using 70% of the phenotypes. AnnotSV are computed based on ranking variants from "pathogenic" to "benign". We break the ties uniformly for all the methods randomly and report the absolute number of variants recovered at each rank together with the recall, as well as areas under the ROC curve (using micro-averages per synthetic genome) and precision-recall curve. Best performing results (using maximum or average score) for each measure are indicated in bold.

		Recall@1	Recall@10	Recall@30	ROCAUC	PRAUC
DeepSVP models using average score	**GO**	458 (0.3047)	630 (0.4192)	788 (0.5243)	0.9644	0.3461
	MP	644 (0.4285)	1026 (0.6826)	1223 (0.8137)	0.9850	0.5158
	HP	519 (0.3453)	942 (0.6267)	1179 (0.7844)	0.9792	0.4335
	CL	137 (0.0912)	534 (0.3553)	878 (0.5842)	0.9742	0.1651
	UBERON	238 (0.1583)	600 (0.3992)	1107 (0.7365)	0.9755	0.2309
	Union	**678 (0.4511)**	**1062 (0.7066)**	1256 (0.8357)	0.9858	**0.5438**
DeepSVP models using maximum score	**GO**	324 (0.2156)	532 (0.3540)	719 (0.4784)	0.9550	0.2682
	MP	227 (0.1510)	629 (0.4185)	880 (0.5855)	0.9598	0.2461
	HP	407 (0.2708)	1008 (0.6707)	**1313 (0.8736)**	**0.9926**	0.4040
	CL	272 (0.1810)	817 (0.5436)	1144 (0.7611)	0.9801	0.2563
	UBERON	254 (0.1690)	652 (0.4338)	1059 (0.7046)	0.9734	0.2401
	Union	323 (0.2149)	945 (0.6287)	1132 (0.7532)	0.9755	0.3480
SV pathogenicity prediction/ranking	**StrVCTVRE**	72 (0.0479)	223 (0.1484)	405 (0.2695)	0.9178	0.0952
	CADD-SV	38 (0.0253)	620 (0.4125)	1020 (0.6786)	0.9816	0.1262
	AnnotSV	19 (0.0126)	229 (0.1524)	700 (0.4657)	0.9605	0.2203

Table 7: Summary of the evaluation for predicting causative variants in the synthetic whole-genome benchmark dataset derived from dbVar, using 50% of the phenotypes. AnnotSV are computed based on ranking variants from "pathogenic" to"benign". We break the ties uniformly for all the methods randomly and report the absolute number of variants recovered at each rank together with the recall, as well as areas under the ROC curve (using micro-averages per synthetic genome) and precision-recall curve. Best performing results (using maximum or average score) for each measure are indicated in bold.

	Rank using maximum scores	Rank using the average score
DeepSVP (GO)	9	9
DeepSV (MP)	8	9
DeepSVP (HP)	**1**	**1**
DeepSVP (CL)	7	27
DeepSVP (UBERON)	6	35
DeepSVP (Union)	**1**	4
StrVCTVRE	4	
CADD-SV	4	
AnnotSV	5	

Table 8: Ranking of disease-associated variant in a Saudi family using different DeepSVP models, and other methods. The ranks of the DeepSVP models are determined based on ranking a total of 47 variants. StrVCTVRE ranks only 6 out of 47 variants. AnnotSV predicted 11 variants as pathogenic out of 47 variants.

4 Structural variant calling

We prepared 150-bp paired-end libraries using the TruSeq Nano DNA Sample Preparation kit (Illumina, USA). Sequencing was performed using an Illumina HiSeq 4000 at the Bioscience core laboratory, KAUST, with approximately 30X coverage. Following sequencing, we aligned reads to human genome build hg38 using the BWA MEM algorithm [23] and following the GATK standard workflows. We trim adapters using Trimmomatic (version 0.38), use BWA (version 0.7.17) for alignment, and samtools (version 1.8) [24] to remove duplicates and sort bam files. We use Manta (version 1.6) [25] to call structural variants. In total, we identified 8,723, 9,003, 9,608, 7,631, and 8,367 variants for the mother, father, first affected, second affected, and unaffected, respectively. We assume an autosomal recessive mode of inheritance or *de novo* variants, and use the pedigree to filter variants. We further filter common variants (minor allele frequency greater than 0.01) using gnomAD (version 2.1.1) [21] and the SVs from the 1000 genomes project. After filtering by pedigree, 148 structural variants remain; removing common variants reduced the number of variants to 47. We use the DeepSVP combined model with *maximum* as aggregation operation for the genes within the variant to prioritize disease-associated SVs.

References

[1] J. Amberger, C. Bocchini, and A. Hamosh, "A new face and new challenges for online mendelian inheritance in man (omim®)," *Human mutation*, vol. 32, no. 5, pp. 564–567, 2011.

[2] R. P. Huntley, T. Sawford, P. Mutowo-Meullenet, A. Shypitsyna, C. Bonilla, M. J. Martin, and C. O'Donovan, "The GOA database: Gene Ontology annotation updates for 2015," *Nucleic Acids Research*, vol. 43, no. D1, pp. D1057–D1063, 11 2014. [Online]. Available: https://doi.org/10.1093/nar/gku1113

[3] D. Maglott, J. Ostell, K. D. Pruitt, and T. Tatusova, "Entrez gene: gene-centered information at ncbi," *Nucleic acids research*, vol. 39, no. suppl_1, pp. D52–D57, 2010.

[4] I. Papatheodorou, P. Moreno, J. Manning, A. M.-P. Fuentes, N. George, S. Fexova, N. A. Fonseca, A. Füllgrabe, M. Green, N. Huang *et al.*, "Expression atlas update: from tissues to single cells," *Nucleic Acids Research*, vol. 48, no. D1, pp. D77–D83, 2020.

[5] J. Chen, A. Althagafi, and R. Hoehndorf, "Predicting candidate genes from phenotypes, functions and anatomical site of expression," *Bioinformatics*, 10 2020, btaa879. [Online]. Available: https://doi.org/10.1093/bioinformatics/btaa879

[6] Tabula Muris Consortium *et al.*, "Single-cell transcriptomics of 20 mouse organs creates a tabula muris." *Nature*, vol. 562, no. 7727, p. 367, 2018.

[7] M. Á. Rodríguez-García, G. V. Gkoutos, P. N. Schofield, and R. Hoehndorf, "Integrating phenotype ontologies with phenomenet," *Journal of biomedical semantics*, vol. 8, no. 1, p. 58, 2017.

[8] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014. [Online]. Available: http://dl.acm.org/citation.cfm?id=2670313

[9] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *Proc. icml*, vol. 30, no. 1, 2013, p. 3.

[10] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[11] J. Breckling and R. Chambers, "M-quantiles," *Biometrika*, vol. 75, no. 4, pp. 761–771, 1988.

[12] A. J. Cornish, A. David, and M. J. E. Sternberg, "PhenoRank: reducing study bias in gene prioritization through simulation," *Bioinformatics*, vol. 34, no. 12, pp. 2087–2095, Jan. 2018. [Online]. Available: https://doi.org/10.1093/bioinformatics/bty028

[13] V. Geoffroy, Y. Herenger, A. Kress, C. Stoetzel, A. Piton, H. Dollfus, and J. Muller, "AnnotSV: an integrated tool for structural variations annotation," *Bioinformatics*, vol. 34, no. 20, pp. 3572–3574, 2018.

[14] M. Griffith and O. L. Griffith, "dbVar (Database of Genomic Structural Variation)," *Dictionary of Bioinformatics and Computational Biology*, 2004.

[15] J. R. MacDonald, R. Ziman, R. K. Yuen, L. Feuk, and S. W. Scherer, "The database of genomic variants: a curated collection of structural variation in the human genome," *Nucleic acids research*, vol. 42, no. D1, pp. D986–D992, 2014.

[16] N. A. O'Leary, M. W. Wright, J. R. Brister, S. Ciufo, D. Haddad, R. McVeigh, B. Rajput, B. Robbertse, B. Smith-White, D. Ako-Adjei *et al.*, "Reference sequence (refseq) database at ncbi: current status, taxonomic expansion, and functional annotation," *Nucleic acids research*, vol. 44, no. D1, pp. D733–D745, 2016.

[17] H. V. Firth and C. F. Wright, "The deciphering developmental disorders (DDD) study," *Developmental medicine and child neurology*, vol. 53, no. 8, p. 702, 2011.

[18] H. L. Rehm, J. S. Berg, L. D. Brooks, C. D. Bustamante, J. P. Evans, M. J. Landrum, D. H.

Ledbetter, D. R. Maglott, C. L. Martin, R. L. Nussbaum *et al.*, "Clingen—the clinical genome resource," *New England Journal of Medicine*, vol. 372, no. 23, pp. 2235–2242, 2015.

[19] K. J. Karczewski, B. Weisburd, B. Thomas, M. Solomonson, D. M. Ruderfer, D. Kavanagh, T. Hamamsy, M. Lek, K. E. Samocha, B. B. Cummings *et al.*, "The exac browser: displaying reference data information from over 60 000 exomes," *Nucleic acids research*, vol. 45, no. D1, pp. D840–D845, 2017.

[20] P. H. Sudmant, T. Rausch, E. J. Gardner, R. E. Handsaker, A. Abyzov, J. Huddleston, Y. Zhang, K. Ye, G. Jun, M. H.-Y. Fritz *et al.*, "An integrated map of structural variation in 2,504 human genomes," *Nature*, vol. 526, no. 7571, pp. 75–81, 2015.

[21] R. L. Collins, , H. Brand, K. J. Karczewski, X. Zhao, J. Alföldi, L. C. Francioli, A. V. Khera, C. Lowther, L. D. Gauthier, H. Wang, N. A. Watts, M. Solomonson, A. O'Donnell-Luria, A. Baumann, R. Munshi, M. Walker, C. W. Whelan, Y. Huang, T. Brookings, T. Sharpe, M. R. Stone, E. Valkanas, J. Fu, G. Tiao, K. M. Laricchia, V. Ruano-Rubio, C. Stevens, N. Gupta, C. Cusick, L. Margolin, K. D. Taylor, H. J. Lin, S. S. Rich, W. S. Post, Y.-D. I. Chen, J. I. Rotter, C. Nusbaum, A. Philippakis, E. Lander, S. Gabriel, B. M. Neale, S. Kathiresan, M. J. Daly, E. Banks, D. G. MacArthur, and M. E. T. and, "A structural variation reference for medical and population genetics," *Nature*, vol. 581, no. 7809, pp. 444–451, May 2020. [Online]. Available: https://doi.org/10.1038/s41586-020-2287-8

[22] S. S. Shapiro and M. B. Wilk, "An analysis of variance test for normality (complete samples)," *Biometrika*, vol. 52, no. 3/4, pp. 591–611, 1965.

[23] H. Li, "Aligning sequence reads, clone sequences and assembly contigs with bwa-mem," *arXiv preprint arXiv:1303.3997*, 2013.

[24] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin, "The sequence alignment/map format and samtools," *Bioinformatics*,

vol. 25, no. 16, pp. 2078–2079, 2009.

[25] X. Chen, O. Schulz-Trieglaff, R. Shaw, B. Barnes, F. Schlesinger, M. Källberg, A. J. Cox, S. Kruglyak, and C. T. Saunders, "Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications," *Bioinformatics*, vol. 32, no. 8, pp. 1220–1222, 2016.