Peer-reviewed version available at *Genes* **2020**, *11*, 1076; doi:10.3390/genes11091076

Article

Annotation of Human Exome Gene Variants with Consensus Pathogenicity

Victor Jaravine[1,*], James Balmford[1], Patrick Metzger[2,3], Melanie Börris[2,3], Harald Binder[1] and Martin Böker[1]

[1] Institute of Medical Biometry and Statistics; Faculty of Medicine and Medical Center, University of Freiburg, Germany; victor.zharavin@imbi.uni-freiburg.de;
[2] Institute of Molecular Medicine and Cell Research; Faculty of Medicine and Medical Center, University of Freiburg, Germany;
[3] Institute of Medical Bioinformatics and Systems Medicine; Faculty of Medicine and Medical Center, University of Freiburg, Germany; melanie.boerries@uniklinik-freiburg.de;
* Correspondence: Correspondence: victor.zharavin@imbi.uni-freiburg.de;

Received: date; Accepted: date; Published: date

Abstract: Pathogenicity is unknown for the majority of human gene variants. For prioritization of sequenced somatic and germline mutation variants, *in silico* approaches can be utilized. In this study, 84 million non-synonymous Single Nucleotide Variants (SNVs) in the human coding genome were annotated using consensus Variant Effect Prediction (cVEP) method. An algorithm, implemented as a stacked ensemble of supervised learners, performed combination of the 39 functional, conservation mutation impact scores from dbNSFP4.0. Adding gene indispensability score, accounting for differences in the pathogenicities of the variants in the essential and the mutation-tolerant genes, improved the predictions. For each SNV the consensus combination gives either a continuous-value pathogenicity score, or a categorical score in five classes: pathogenic, likely pathogenic, uncertain significance, likely benign, benign. The provided class database is aimed for direct use in clinical practice. The trained prediction models were 5-fold cross-validated on the evidence-based categorical annotations from the ClinVar database. The rankings of the scores based on their ability to predict pathogenicity were obtained. A two-step strategy using the rankings, scores and class annotations is suggested for filtering and prioritization of the human exome mutations in clinical and biological applications of NGS technology.

Keywords: Variant of Unknown Significance (VUS); Single-Nucleotide Variant (SNV); Variant Effect Prediction (VEP); Stacked Ensemble of Supervised Deep Learners (SESDL); Next Generation Sequencing (NGS); Alternative Allele Frequency (AAF)

1. Introduction

With the growing availability of NGS technology in medical diagnostics, quantitatively exemplified by 68,000 genetic tests currently offered in clinics according to Genetests [1], the amount of experimentally sequenced data on human genetic variation in both healthy and patient populations is rapidly increasing. Accurate and exhaustive variant annotation is important for every application of NGS technology, including development of therapies, selection of effective individualized therapy, and comparing multiple samples in clinical studies, to mention a few; however, currently there are major challenges associated with sequencing data analysis and interpretation [2]. While most of the prevalent or common variants (about one million of currently sequenced variants) have been annotated as benign (neutral) based on their high frequency of occurrence in the healthy population [3], the remaining majority are ultra-rare Variants of Unknown Significance (VUS). In total there are almost 100 million possible human Single Nucleotide Variants (SNV) in the gene coding 'exome' space, of which around 83 million are non-synonymous nsSNVs

and around 15 million are short splice-site ssSNVs. However, only a tiny fraction (ca 0.1% of the total number) of variants have definitive clinical annotations. The current inability to annotate and prioritize the vast number of sequenced VUSs by variant effect significance is an obstacle to the reliable and wider use of NGS in diagnostics and treatment of cancer and other diseases.

As an alternative to the scarce annotations based on clinical and biological evidence alone and a plethora of predicted real-valued pathogenicity scores, in the following we are proposing an approach for combining both these types of data into two consensus scores: a continuous-value score and a categorical classification score. These two types of scores have their pros and cons, which are compensated in their combination. While there is a number of continuous ensemble scores that are already commonly used, they possess a key disadvantage of lacking determined thresholds for different genes, leading to uncertainty or errors in differentiating between pathogenic and benign variants based on the score value alone. Such thresholds need to be determined for each of the continuous-value pathogenicity prediction scores, and are in most cases not known. Furthermore, it might be not correct to use a single pathogenicity score threshold for all genes, as for example, variants in the essential genes typically require a higher threshold compared to the mutation-tolerant genes. To remedy this situation, here we use a gene indispensability score as one of the inputs in training to account for such differences due to varying gene pathogenicities. In contrast, categorical classes, such as 'Pathogenic', 'Uncertain_significance' or 'Likely_benign', are easier interpretable and can be used directly in clinical practice, without the need for any thresholds determination. To the best of our knowledge, no studies have developed an ensemble method able to classify variants into pathogenicity classes across the entire coding genome, and with high enough accuracy. In this study we have achieved accuracy of classification of around 1%. We propose to use the predicted pathogenicity classes in the same way as the currently used evidence-based classification data, where such are missing, i.e. for the large majority of variants.

The classification (termed consensus variant effect prediction, or cVEP) of gene variants by pathogenicity uses a stacked ensemble of supervised learners machine learning approach. The classification scheme is in accordance with the standards and guidelines of the American College of Medical Genetics and Genomics (ACMG), and the Association of Molecular Pathology (AMP) [4]. The variants are classified into five categories: pathogenic, likely pathogenic, uncertain significance, likely benign, benign. This evidence-based data, used for our input, is rather accurate, since ACMG/AMP expanded guideline requires to have certainty from experimental clinical/biological evidence of better than 90% [5] for inclusion into the above classes. It is worth mentioning another disadvantage of the majority of existing continuous-value pathogenicity predictions – they use only a subset of the evidence-based classes for training, namely only 'benign' and 'pathogenic' class data points are used to train model predictions into the 0 to 1 range, and simply do not use data from other classes.

Recently, there has been a rapid increase in the use of variant allele frequency as a proxy for differentiating between disease-causing (or pathogenic) and benign (or neutral) variants, based on the assumption that when a variant has the allele frequency of above 0.1% it is very likely to be benign. For example, Minor Allele Frequency (MAF) and Alternative Allele Frequency (AAF) annotations, which are being obtained in large population sequencing projects such as ExAC [6] and GnomAD [7], are routinely used in various clinical and diagnostic applications in relation to Mendelian and other diseases to distinguish between common and rare variants. MAF is the frequency of the second-most occurring allele, while AAF describes frequencies of all possible non-synonymous mutations at a gene locus. Alternatively, one can compute the population maximal frequency (e.g. AAFpopmax), which is calculated as the maximum of AAF values across individual sub-populations. It has been shown [6] that use of AAFpopmax produces considerably fewer variants after filtering with the same threshold, compared to that using AAF.

While it is possible to a certain extent to use the AAF value alone for benign/pathogenic classification, it has clear limitations. Firstly, since the histogram of the known AAF values follows a gamma distribution, it is intrinsically difficult to choose where to put a threshold. Consequently, it is problematic to achieve a clear separation to discriminate between benign and pathogenic variants

using a single AAF threshold for all genes. The problem is alleviated recently by proposed use of different AAF thresholds for each gene [8], with the threshold values selected manually to aim for the best discrimination in the ClinVar database. Secondly, despite the proliferation of use of variant frequency, these annotations are currently available for only a small proportion of the total number of possible variants (~6%). Consequently, a large majority of germline and somatic rare variants (i.e. below 0.1% AAF) discovered in WES samples currently have unknown a priori functional significance.

Whole Exome Sequencing (WES) of a tumor sample usually reveals a few mutation variants that are important for cancer initiation and survival, called 'drivers', hundreds of 'passenger' variants associated with cancer evolution and differentiation, and hundreds of thousands of neutral variants that have nearly no impact [9]. The small number of causative variants can be effective therapy targets, but they first need to be distinguished from the bulk of VUSs and common variants by filtering and prioritization using annotations. To remedy this situation, many *in silico* methods for variant effect prediction (VEP) have been developed based on gradually improving machine learning methods. Generally, rare variants are typically less evolutionarily conserved and have higher predicted pathogenicity than frequent ones [10]. This relationship is utilized in a number of VEP tools (e.g., GERP++, SiPhy, bStatistic) [11] that are based on nucleotide conservation in alignments of genetic sequences in mammalian species. Another group of tools, which aim to predict the functional impact of mutations (e.g. MutationTaster, Eigen, fitCons) are trained on slowly growing number of pathogenic/benign annotations derived from clinical and biological studies. In addition, a group of VEP methods based on structural approaches has shown better performance in finding cancer-causing driver variants, based on predictions of structural stability, ligand binding or protein-protein network interactions. Some of methods in this group are based on structural clustering and classify a mutation as a driver if it is close to a cluster of known driver mutations [12]. As an example of such a tool, CADD [13] uses mapping of mutation variants to protein 3D structure for prediction of effects of mutations.

Ensemble methods, which use combinations of predictive scores of several types, have shown better performance than any one individual tool. In The Cancer Genomic Atlas (TCGA), a dataset comprised of 10,000 tumor samples across 33 cancer types, it was shown [14] that improved prediction results were obtained by combining sequence conservation and population data-based tools with structural approaches. In this study [14] *in vitro* functional tests were able to confirm the predictions of four structural tools in 78% of cases, a considerably better rate than when using methods that do not use structural information. Recent advances in ensemble methods include combining a larger number of tools, greater methodological diversity, using more training data, and using better machine learning methods such as random forests or gradient boosting machines, which overcome over-fitting and other computational problems such as highly unbalanced datasets.

Despite these advances, prediction errors are common even with the best available tools, particularly for rare variants[15]. For example, recent quantitative comparison of performance of the current tools on rare cancer variants has shown [15] that they have adequate sensitivity of 84-95%, but low specificity in the range of 25-83%, meaning a high probability of errors in prediction. The level of accuracy that is currently achievable would translate to thousands of false annotations in a patient tumor sample with a high mutation rate, which is unacceptable. As a consequence, there remains a lack of consensus regarding which method is best for predicting variant effects.

In the study, first, we illustrate how variant class discrimination improves with increasing number of dimensions of input data. Then, we describe a meta-learning consensus method which combines a selection of 39 of the best-performing functional impact effect, sequence conservation and pathogenicity scores, and one gene-level score. We then train and cross-validate the consensus ensemble model on a dataset of known pathogenicity annotations derived from the ClinVar database. We provide rankings for the ability of these scores to predict pathogenicity, both individually and in combination, derived from the feature importance in the base models. We apply the ensemble model to predict pathogenicity scores and classes for the human exonic variants in all chromosomes. Finally,

to demonstrate the potential of the method for variant effect prediction in clinical applications, we provide data on its application to several highly pathogenic genes.

2. Materials and Methods

2.1. Machine Learning algorithm

A Stacked Ensemble of Supervised Deep Learners (SESDL) is a machine learning algorithm which combines several prediction algorithms, so-called base-learners, using a process called stacking (also known as meta- or super-learning). Unlike bagging and boosting approaches, also used for combining base-learners, the goal in stacking is to combine a diverse set of learners together. It has been shown [16] that a super-learner ensemble represents an asymptotically optimal system for learning. Among stacking algorithms super-learning is distinguished by the direct implementation of cross-validation for tuning parameter selection in the algorithm.

As a basis for our approach, we use the H2O tool [17], which is an open source, in-memory, distributed, fast, and scalable machine learning and predictive analytics tool that allows users to build machine learning models for big data analysis, implemented as a package in the statistical environment R. The 'meta-learning' algorithm, implemented here using the *h2o.stackedEnsemble* function of the H2O tool (version 3.0), finds an optimal combination of the following three base learners: generalized linear model (GLM) [18], gradient boosting machine (GBM) [19] and distributed random forest (DRF) [20].

The popular GLM method is a well-known generalization of linear regression for response variables. More recently developed DRF and GBM methods both build collections of decision trees. A decision tree captures interactions among different features, with the interactions order controlled by a tree depth parameter – higher order for higher depths. There are significant differences in these algorithms: briefly, DRF averages multiple decision trees, created on different random samples of rows and columns; GBM builds the model in a stage-wise fashion. The algorithms are non-linear, robust for noisy data, provide importance of each predictor in the models. Here the base-models use 'multinomial' distribution for class prediction; the types of distributions used for continuous combination score are given in the Supplementary Materials.

The key parameters (max_depth, ntrees, max_after_balance_size, learn_rate/sample_rate, min_rows) for DRF, GBM and GLM, were chosen using the grid search in the parameter space to produce fits with the lowest logarithmic score and with its achieved stability in the five-fold cross-validations. The logarithmic score [21] method is connected to Shannon entropy and Kullback–Leibler divergence; here it is denoted by the term *logloss* (Eq.1), same as was used in H2O documentation. It measures the performance of a multinomial classification model, with the goal of the machine learning model being to minimize it. A perfect model would have a *logloss* of 0. *Logloss* increases as the predicted probability diverges from the actual label. For the multi-class classification, the sum of *logloss* values for each class prediction in the observation is taken:

$$Logloss = - \sum_c y_{c,o} \log(p_{c,o}) \quad (1)$$

Where $y_{c,o}$ is a binary indicator (0 or 1) of whether class label c is the correct classification for observation o; $p_{c,o}$ - the model's predicted probability that observation o is of class c.

Since there were a different number of points for each class, the learning was performed with balancing classes. The type of algorithm used as the meta-learner was GLM with nonnegative weights. When the algorithms were learning categorical values, instead of the R^2 the McFadden's pseudo R-squared, also known as likelihood-ratio index, was used. It is a statistical measure of how close the predicted and input data are in the fits. The log likelihood of the intercept model is treated as a total sum of squares, and the log likelihood of the full model is treated as the sum of squared errors. The ratio of the likelihoods, scaled to 0-1 range, gives the level of improvement over the intercept model offered by the full model. When comparing two models on the same data, McFadden's ratio is higher for the model with the greater likelihood.

2.2. Datasets

The GnomAD [7] dataset extends the ExAC [6] dataset from 60,706 to 125,748 unrelated individuals of a diverse set of ethnicities including European, African, Latino, South Asian, East Asian, and Other. It contains allele frequencies of the sequenced genetic (exome and whole genome) variants.

ClinVar [22] is a freely accessible, public archive of the relationships among human variations and phenotypes, with supporting evidence. Its interpretations of the clinical significance of germline and somatic variants for reported conditions are based on "ground-truth" - clinical and biological evidence. The archive located at the National Center for Biotechnology Information (NCBI) is maintained according to variant classification standards and guidelines of the American College of Medical Genetic and Genomics (ACMG) and the Association for Molecular Pathology (AMP) [5]. The ClinVar dataset is the subset of the GnomAD dataset with non-empty ClinVar annotations, including a total of 227,232 nsSNVs.

The dbNSFP [23] is a database integrating annotations from multiple sources, including allele frequencies from ExAC, GnomAD, clinical annotations from ClinVar, and many state-of-the-art functional and conservation annotations for the comprehensive collection of human SNVs [23]. Its current academic version (named "a"-branch), including a total of 84,013,490 non-synonymous SNVs (nsSNVs), compiles prediction scores from 29 functional prediction algorithms (*SIFT, SIFT4G, Polyphen2-HDIV, Polyphen2-HVAR, LRT, MutationTaster2, MutationAssessor, FATHMM, MetaSVM, MetaLR, CADD, VEST4, PROVEAN, FATHMM-MKL coding, FATHMM-XF coding, fitCons, LINSIGHT, DANN, GenoCanyon, Eigen, Eigen-PC, M-CAP, REVEL, MutPred, MVP, MPC, PrimateAI, GEOGEN2, ALoFT*), 9 conservation scores (*bStatistic, phyloP100way_vertebrate, phyloP30way_mammal, phyloP17way_primate, phastCons100way_vertebrate, phastCons30way_mammal, phastCons17way_primate, GERP++ and SiPhy*) and a number of other functional annotations from various sources (Table S2).

Each of the curated training, prediction and validation datasets contained the following annotations from the dbNSFP4.0 database [23]: 40 functional annotation scores (Table 1, the scores' sources are listed in Table S1); the gene-level score *Gene_indispensability_score*, a probability prediction of whether the gene is essential (from 0 to 1); and the two AAF and two AAFpopmax fields for the exome and control samples: *gnomAD_exomes_AF* is the alternative allele frequency in all of the gnomAD exome samples (125,748 samples); *gnomAD_exomes_POPMAX_AF* is the maximum allele frequency across populations (excluding samples of Ashkenazi, Finnish, and indeterminate ancestry); *gnomAD_exomes_controls_AF* is the alternative allele frequency in the controls subset of the gnomAD exome samples (54,704 samples); *gnomAD_exomes_controls_POPMAX_AF* is the maximum allele frequency across populations (excluding samples of Ashkenazi, Finnish, and indeterminate ancestry) in the controls subset.

The gene indispensability score was included because it was expected to lead to a significant improvement in predictive performance, particularly for rare variants and VUSs. It is well known that a moderately deleterious variant in a non-essential gene can often exhibit itself as non-pathogenic, as e.g. such can be located in s mutation-tolerant gene, or in a truncated dis-functional gene having a duplicate functional gene; whereas such a variant in an essential gene is usually very pathogenic, since such genes are involved in several cellular networks, leading to a major cell function disruption. The original indispensability model [24] builds a 'multinet' from several networks and then derives an 'Indispensability' score from the degree-centralities and other parameters in various types of networks: protein-protein interactions (PPI), phosphorylation, signaling, metabolic, genetic, regulatory, essentiality, LoF-tolerance, number of networks in which a gene is involved, number of protein interaction interfaces, dN/dS ratios, paralogs, etc. The score value is high e.g. for variants in oncogenes and tumor suppressors, while it is very low in duplicated and mutation-tolerant genes.

Table 1. List of scores used (the sources are provided in the Suppl. Materials), 1-31 functional scores, 32-40 conservation scores, 41 gene indispensability score. .

*X**	*SCORE NAME*		
1	*SIFT_converted_rankscore*	**21**	*DANN_rankscore*
2	*SIFT4G_converted_rankscore*	**22**	*fathmm-MKL_coding_rankscore*
3	*Polyphen2_HDIV_rankscore*	**23**	*fathmm-XF_coding_rankscore*
4	*Polyphen2_HVAR_rankscore*	**24**	*Eigen-raw_coding_rankscore*
5	*LRT_converted_rankscore*	**25**	*Eigen-PC-raw_coding_rankscore*
6	*MutationTaster_converted_rankscore*	**26**	*GenoCanyon_score_rankscore*
7	*MutationAssessor_rankscore*	**27**	*integrated_fitCons_rankscore*
8	*FATHMM_converted_rankscore*	**28**	*GM12878_fitCons_rankscore*
9	*PROVEAN_converted_rankscore*	**29**	*H1-hESC_fitCons_rankscore*
10	*VEST4_rankscore*	**30**	*HUVEC_fitCons_rankscore*
11	*MetaSVM_rankscore*	**31**	*Mean_rankscore*
12	*MetaLR_rankscore*	**32**	*GERP++_RS_rankscore*
13	*M-CAP_rankscore*	**33**	*phyloP100way_vertebrate_rankscore*
14	*REVEL_rankscore*	**34**	*phyloP30way_mammalian_rankscore*
15	*MutPred_rankscore*	**35**	*phyloP17way_primate_rankscore*
16	*MVP_rankscore*	**36**	*phastCons100way_vertebrate_rankscore*
17	*MPC_rankscore*	**37**	*phastCons30way_mammalian_rankscore*
18	*PrimateAI_rankscore*	**38**	*phastCons17way_primate_rankscore*
19	*DEOGEN2_rankscore*	**39**	*SiPhy_29way_logOdds_rankscore*
20	*CADD_raw_rankscore*	**40**	*bStatistic_rankscore*
		41	*Gene_indispensability_score*

2.3. Training, prediction and validation procedures

Training and prediction method used the 40 scores (Table 1) and the gene indispensability score derived from the dbNSFP4 database. The *LINSIGHT_rankscor*e ("x31") was missing for most of the variants in GnomAD, it was omitted and was replaced by us with a mean of 39 scores available for each variant, termed "Mean_rankscore".

The combined "GnomAD-AAF" dataset consisted of 5,502,403 variants, each with *Gnomad_exome_AF* annotations. This dataset was further subdivided into 'Control', with non-zero *Gnomad_exome_control_AF* annotations (3,451,486 variants obtained from 54,704 samples from individuals not associated with any disease), and 'Disease', which included the remaining 2,050,917 variants (i.e. annotated with *Gnomad_exome_AF* but not with *Gnomad_exome_control_AF*). There were two AAF versions used: *Gnomad_exome_AF* and *Gnomad_exome_AFpopmax* annotations.

All nsSNV variants listed in the DBNSFP4.0 database were classified into established categories of ClinVar annotations. The training/cross-validation dataset comprised the GnomAD variants with ClinVar annotations (133,514 variants) across five classes: 'Benign' (35,659 variants, denoted by "2" label in calculations), 'Likely_benign' (21,931 variants, labelled as "1", including Benign/Likely_benign), 'Uncertain_significance' (18,716 variants, labelled as "0"), 'Likely_pathogenic' (21,510 variants, labelled as "-1", including 'Pathogenic/Likely_pathogenic'), and 'Pathogenic' (35,698 variants, labelled as "-2"). Note that to improve the imbalance between the classes the original ClinVar categories Benign/Likely_benign and Likely_benign were merged into the 'Likely_benign' class, and similarly, ClinVar's Pathogenic/Likely_pathogenic and

Likely_pathogenic were merged into the 'Likely_pathogenic' class; the 'Uncertain_significance' class was down-sampled seven times randomly. The 2D histograms for the resulting five ClinVar categories were plotted as heat maps. The histogram plot illustrates the distribution of variants across the 41 score dimensions separately for the classes.

The models were then applied to predict cVEP classes for the prediction dataset comprised of all nsSNVs contained in the dbNSFP4.0 database, i.e. 84,013,490 human nsSNVs in the human coding genome (exome) split by the chromosome regions. This database is the resulting output provided via an URL link. In addition, we illustrate the dependence of the cVEP classifications on the gene indispensability, and differences in individual genes of two pathogenic types.

3. Results

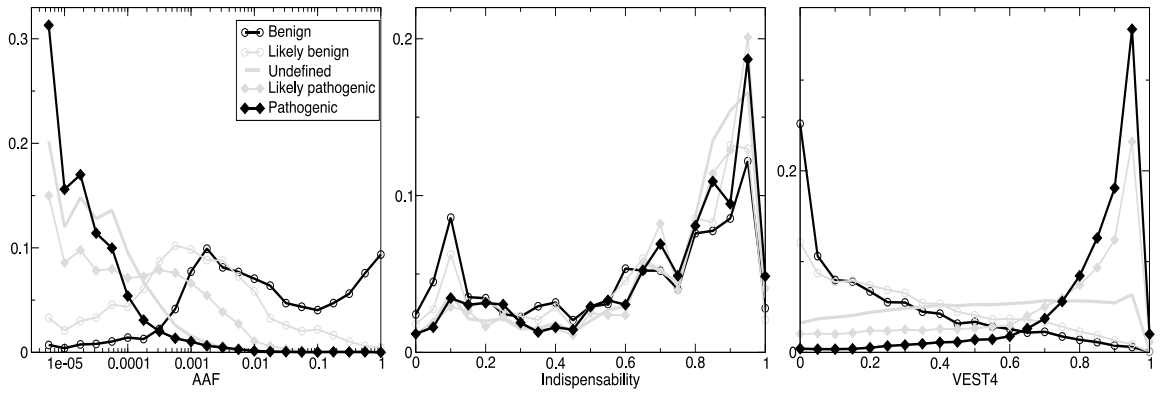
3.1. Visualization in 1- and 2-dimensional space

For the purpose of illustrating the usefulness of combining different scores, we show how the variants belonging to different pathogenicity classes can be progressively better discriminated, going from 1- to 2-dimensional space using one or two selected features, respectively. Subsequently, we show how the classification in nD space (41 features) is learnt by our proposed ensemble machine learning approach.

3.1.1. One-dimensional discrimination of ClinVar variants by using AAF, Indispensability and VEST4 annotations

Figure 1 displays three histograms of the five classes of variants from the GnomAD database with available ClinVar annotations, in relation to the fit features: AAF, Indispensability and VEST4, where the latter is the best performing functional annotation (see Section 3.2.4). For AAF there is a threshold clearly identifiable on the plot between the benign and the pathogenic classes at 0.0003, but the corresponding "Likely" classes are more overlapped. The histograms in the middle panel are very similar, but there are twice as many variants for pathogenic class in relation to the benign class for low indispensability seen in the histogram bin at 0.1, and the opposite is true at 0.95. This finding translates to there being twice as many benign variants for the mutation-tolerant genes (i.e. having low indispensability score), and twice as many pathogenic variants for essential genes (i.e. with high indispensability). There is a very clear separation in the right panel, where pathogenic and benign (as well as the respective "Likely" classes) can be distinguished by a threshold at 0.64 VEST4 score. Here there is no differentiation for the "Uncertain_significance" class ('Undefined' label).

Figure 1. The histograms of the GnomAD variants with ClinVar annotations in five classes are plotted



in one dimension for AAF (left), indispensability (middle) and VEST4 score (right). The counts in each are normalized to the total count in all five classes. There are 20 bins (0.05 width) for the scores, and

24 bins for AAF (0.25 width on the log10 scale). For AAF the x-axis is plotted on the log10-scale, e.g. 10-2 means that the variant can be found on average in 1 every 100 individuals. The five classes are: 'Benign' (thick open black circles), 'Likely_benign' (thin open grey circles), 'Uncertain_significance' (thick grey), 'Likely_pathogenic' (thin grey closed diamonds) and 'Pathogenic' (thick closed black diamonds).

3.1.2. 2D discrimination of variants by pathogenicity using the population allele frequencies

Figure 2 illustrates the distribution of the prevalence of all AAF and AAF_{popmax} values among the nsSNV variants, sequenced in the GnomAD (Figure. 2 left panel) and the ClinVar (Figure 2 right panel) databases. A variant is above the diagonal on the plots when $AAF_{popmax} > AAF$, signifying that the variant has higher prevalence in some sub-population in comparison to the entire population. The 'control' group variants (cyan colored dots) span the entire range of AAF above the diagonal; rare variants are found both above and below the diagonal to a similar extent. Conversely, 'disease' variants, i.e. the GnomAD exome variants that are not control variants, are mostly very rare (AAF range 10^{-4}–10^{-6}) and are situated above the diagonal on the plot. Similar to the 'control' group, the ClinVar annotations in the *'Benign'* and *'Likely_benign'* groups span the entire range of frequencies, while similarly to the 'disease' group, the *'Likely_pathogenic'* and *'Pathogenic'* groups are also clustered in the very rare range and are situated above the diagonal. The *'Conflicting'* group (ClinVar term *'Conflicting_pathogenicity_interpretations'*) has mostly intermediate frequency values.

Figure 2. The variants are plotted as colored dots in two dimensions on the log_{10}-scale: x-axis - AAF, and y-axis - AAF_{popmax}, e.g. 10^{-2} means that the variant can be found on average in 1 every 100 individuals. In the left panel, the variants from the GnomAD database are plotted, separated into two groups – 'control' (cyan, 54,704 samples with *Gnomad_exome_control_AF*>0) and 'disease' (brown, 125,748 samples excluding 'control' samples with *Gnomad_exome_AF*>0). In the right panel, the variants from GnomAD database with available AAF/AAF_{popmax} annotations and with the following ClinVar annotations are plotted by colored dots in five groups: *'Benign'* (green), *'Likely_benign'* (cyan), *'Conflicting'* (grey), *'Likely_pathogenic'* (yellow) and *'Pathogenic'* (red).

Overall, the majority of common, rare and very rare variants are above the diagonal on both plots, while a large proportion of rare variants are below the diagonal. It is noteworthy that both the 'Disease' (brown dots, Figure 2, left) and 'Pathogenic' variants (red circles, Figure 2, right) lie mostly above the diagonal. Consequently, the measure of difference ($AAF_{popmax} - AAF > 0$) is potentially useful (in combination with $AAF<10^{-4}$) as an additional measure to distinguish very rare pathogenic variants from very rare benign variants. This highlights the need for combining multiple scores by a flexible machine learning approach that, e.g., can automatically leverage such difference information.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 July 2020 doi:10.20944/preprints202007.0735.v1

Peer-reviewed version available at Genes 2020, 11, 1076; doi:10.3390/genes11091076

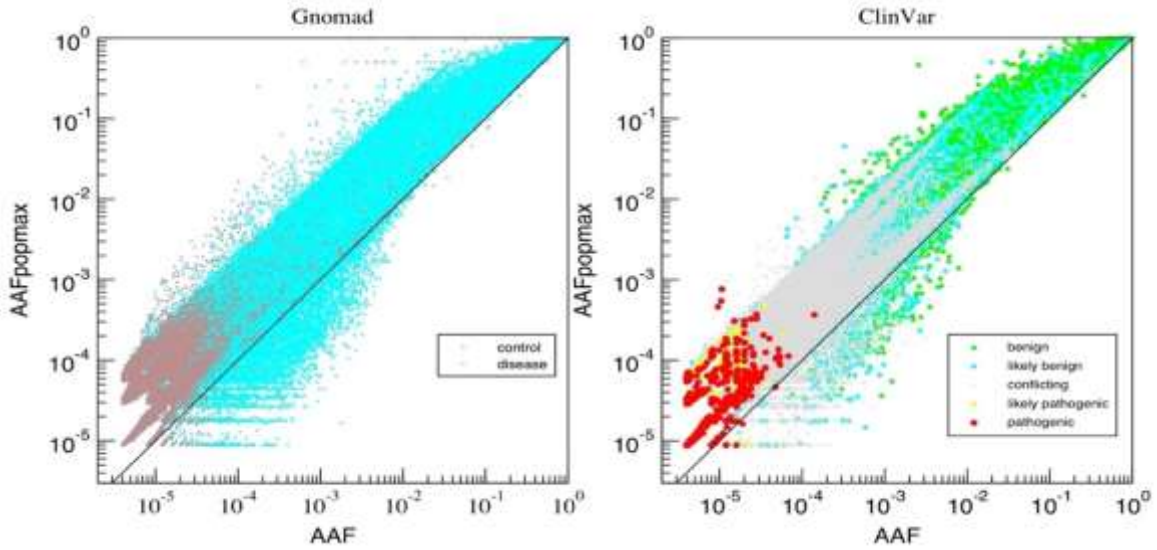


Figure 2. The variants are plotted as coloured dots in two dimensions on the $\log_{10}$-scale: x-axis - AAF, and y-axis - AAF_{popmax}, e.g. 10^{-2} means that the variant can be found on average in 1 every 100 individuals. In the left panel, the variants from the GnomAD database are plotted, separated into two groups – 'control' (cyan, 54,704 samples with *Gnomad_exome_control_AF*>0) and 'disease' (brown, 125,748 samples excluding 'control' samples with *Gnomad_exome_AF*>0). In the right panel, the variants from GnomAD database with available AAF/AAF_{popmax} annotations and with the following ClinVar annotations are plotted by coloured dots in five groups: *'Benign'* (green), *'Likely_benign'* (cyan), *'Conflicting'* (grey), *'Likely_pathogenic'* (yellow) and *'Pathogenic'* (red).

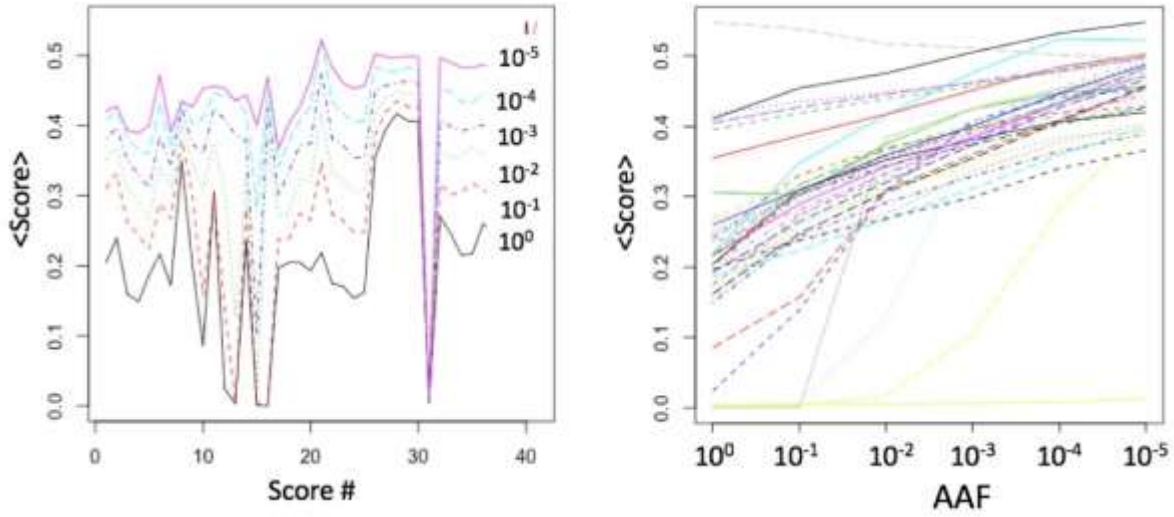


Figure 3. Comparison of the averages of the 40 functional impact scores by the six AAF range groups. Approx. 5 million exome variants from the GnomAD sequencing project are annotated using dbNSFP4, containing variants from approx. 100K individuals. The left panel shows the score means (<Score>) as a function of the score numbers; sorting corresponds to the sequential order of columns in the dbNSFP database, as provided in Table 1. The right panel shows the same transposed data, with the average score values (<Score>) shown as a function of AAF ranges, $\log_{10}$-scale.

Each functional prediction tool produced a converted rank-score for each variant, with the values ranging from 0 (benign) to 1 (pathogenic). To illustrate the average score values in relation to AAF, Figure 3 shows the means of the 40 scores for the GnomAD dataset, computed for the six equally separated AAF ranges on the $\log_{10}$-scale. As the ranges regress from common to very rare (from bottom to top on the left panel), the score means increase, for example from around 0.2 for the range [1.0-0.1] to 0.5 for the range [1e-5 – 1e-6]. The right panel shows that for most scores, there is a close to linear relationship between the score mean values and the AAF value (lower bound of range), while some scores show step-function dependence.

3.1.4. 2D discrimination of variants using the pathogenicity scores

In Figure 4 all variants with ClinVar annotations are plotted by colored squares in four groups. The pathogenic variants (red, orange) are clustered in the top-right corner on both plots, while benign variants (dark-green, green) have a considerably greater spread across the panels, with noticeably better clustering in the bottom-left corner for the M-CAP vs VEST4 2D (left panel), in comparison to the REVEL vs VEST4 (right panel). The 2D variant discrimination is more resolved compared to 1D, but there is a fairly large amount of overlap between the classes.

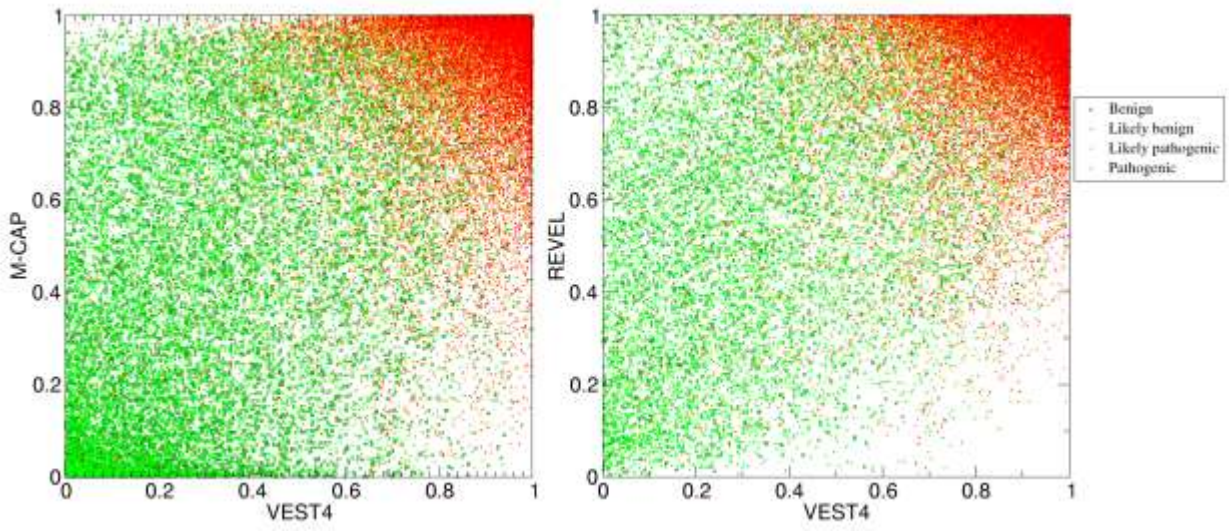


Figure 4. The ClinVar annotated variants are plotted in two dimensions: x-axis – VEST4 scores, and y-axis – M-CAP (left panel) or REVEL (right panel) scores. The classes: *'Benign'* (dark green), *'Likely_benign'* (green), *'Likely_pathogenic'* (orange) and *'Pathogenic'* (red).

3.2. Training and validation using ClinVar Annotations

3.2.1. Ranking of predictive capacity of individual pathogenicity features

To obtain a ranking of the features based on cross-correlation between the score values and the predicted class values, we used the ClinVar dataset as described in the *Methods* section, with training using only one base classifier "GLM" applied to predict classes when inputting only one feature, i.e. each of the scores individually. The rankings sorted by McFadden's pseudo R-squared, which quantifies explained variation, are given in Table 2. The most highly correlated score was "x10" (VEST4) with R^2 =0.82, closely followed by "x20", "x24" with R^2 of 0.81. The other R^2 values were in the range of 0.75 – 0.81.

Peer-reviewed version available at Genes 2020, 11, 1076; doi:10.3390/genes11091076

Table 2. Ranking of features sorted by pseudo-R^2.

"x10" 0.818 VEST4	"x22" 0.790 fathmm-MKL	"x32" 0.775 phyloP100_vertebrate
„x20" 0.811 CADD	"x11" 0.789 MetaSVM	"x8" 0.771 fathmm
„x24" 0.811 Eigen	"x17" 0.787 MPC	"x37" 0.770
„x31" 0.808 Mean	"x9" 0.787 Provean	phastCons30mammalian
„x6" 0.806 MutationTaster	"x3" 0.786 Polypen2_HDIV	"x34" 0.768 phyloP30_mammalian
"x25" 0.805 Eigen-PC	"x7" 0.786 MutationAssessor	"x38" 0.767 phastCons17_primate
"x15" 0.799 MutPred	"x1" 0.785 SIFT	"x26" 0.765 GenoCanyon
"x14" 0.798 REVEL	"x33" 0.784 phyloP100_vertebrate	"x41" 0.763 Gene_indispensability
„x13" 0.798 M-CAP	"x2" 0.781 SIFT4G	"x35" 0.762 phyloP17way_primate
„x12" 0.797 MetaLR	"x5" 0.781 LRT	"x27" 0.751 integrated_fitCons
„x19" 0.793 Deogen2	"x39" 0.781	"x29" 0.750 H1-hESC_fitCons
"x16" 0.793 MVP	SiPhy_29way_logOdds	"x40" 0.750 bStatistic
"x18" 0.792 PrimateAI	„x36" 0.780	"x30" 0.750 HUVEC_fitCons
"x4" 0.791 Polyphen2_HVAR	phastCons100vertebrate	"x28" 0.750 GM12878_fitCons
	„x21" 0.779 DANN	
	„x23" 0.779 fathmm-XF	

3.2.2. Histogram of distribution of 41 score values for the ClinVar dataset

Figure 5 shows 2D histograms of the values of 41 scores for the nsSNVs, divided into five ClinVar-derived pathogenicity classes (see *Methods*). To aid interpretation the scores were sorted in descending order along the x-axis according to Table 2, i.e. to the individual feature correlation with the predicted classes. Consequently, it is apparent that the score values in the left half of the plot are considerably more correlated and similar than in the right half, where the histogram patterns are much more random or less correlated to the predicted class. In particular, in the left half the class of pure *'Benign'* has high histogram (normalized) counts for score values that are close to 0.0 (i.e. towards the bottom of the class panel), and low counts for the medium and high score values. The class of *'Likely_benign'* variants has a very similar pattern to the pure '*Benign*' class. In contrast, the *'Likely_pathogenic'* and *'Pathogenic'* classes have high variant prevalence for score values close to 1.0 (towards the top of the class panel). The histogram pattern in the '*Uncertain_significance*' class in the left half is quite indistinct, while in the right half the pattern is similar to the *'Benign'* class. In the right half of each panel the histogram patterns tend to be somewhat similar, and thus poorly correlated with the predicted class.

Figure 5. 2D histogram of the score values of the variants by the five ClinVar classes. The x-axis displays the 41 features, sorted from left-to-right according to decreasing r^2 as shown in Table 2. The y-axis on the left side denotes the five ClinVar classes, and the Y-axis on the right corresponds to the histogram bins for the values of each feature (on the x-axis), i.e. 10 equally 0.1-spaced bins between 0 and 1. The colour from white (0) to blue (0.7) corresponds to histogram counts in each of the bins, normalized to sum to 1 for the variant counts for each class and each score.

3.2.3. Training and cross-validation of the ensemble model

The prediction method uses for training the annotations features (listed in Table 1) derived for the GnomAD variants having ClinVar annotations, as described in the *Methods* section. The various accuracy measures for the three base- and ensemble-learners can be seen (Table S2-S4). The summary metrics (Table 3a) shows that the *GLM* has the worst performance, *GBM* intermediate and *DRF* the highest among the base-learners. The majority of the metrics are significantly better for *DRF* and *Ensemble* compared to the first two base-learners. The performance measures of five-fold cross-

validation (80% training, 20% testing, with data points selected randomly using the 'modulo' algorithm from the *h2o* package) for the three base-learners ("GLM", "GBM", "DRF") and the Super-learner ("Ensemble") are given in Tables S2-5, with the main test metrics summarized in Table 3a. The similar results of the performance measures and the small standard deviations in the five cross-validation runs (Tables S2-5) point to the robustness and of the prediction of cVEP classes.

Table 3. a. Summary of test-set metrics for cVEP class prediction.

Metric	*GLM*	*GBM*	*DRF*	*Ensemble*
mse	0.39	0.21	0.12	0.10
rmse	0.62	0.46	0.34	0.32
logloss	1.06	0.65	0.39	0.36
mpce	0.53	0.29	0.074	0.084
Hit ratio 1	0.55	0.74	0.939	0.925
Hit ratio 2	0.83	0.93	0.997	0.988

Mse, mean standard error; rmse, root mean square error; logloss, logarithmic loss; mpce, Mean Per-Class Error; Hit ratio 1, within class; Hit ratio 2, within 2 classes.

Table 3. b. Summary of test-set metrics for real-value prediction.

Metric	*GLM*	*GBM*	*DRF*	*Ensemble*
mse	0.79	0.50	0.55	0.24
rmse	0.89	0.71	0.74	0.49
mae	0.62	0.52	0.55	0.36
mrd	0.79	0.50	0.55	0.23
R_c^2	0.82	0.89	0.957	0.952

Mse, mean standard error; rmse, root mean square error; mae, mean average error; mrd, mean residual deviation; R_c, Pearson's correlation coefficient.

The resulting *ensemble* performance: *mean-per-class-error* is 8.4%; 7.4% of total number of misclassifications according to *'Confusion Matrix'* (Table S4), i.e. when the predictions fall into the other than correct classes. This accuracy corresponds to *"Hit ratio 1"* of 93% for predicting the correct class exactly, and a "Hit ratio 2" of 98.8% for predicting either the correct or the nearest class. Since 'next' is along the class sequence ("-2" "Pathogenic" , "-1" "Likely_pathogenic" , "0" "Uncertain_significance", "1" "Likely_benign", and "2" "Benign"), misclassification into next-class is acceptable, while three- or more class misclassifications are not (*Hit ratios* 3-5). The errors are attributable to ambiguity in the classes themselves, e.g. mixing between *likely* and *pure* classes. As can be seen from (Table S4) most of the misclassifications occur between "Pathogenic" and "Likely_pathogenic", "Likely_benign" and "Benign". Besides, the inaccuracies are due to the errors in the input scores, and due to missing data for some scores in the sets (in the GnomAD database the number of scores available for each variant range between minimum of 23 and maximum of 40).

Table 3b shows the accuracies for real-valued pathogenicity prediction. Here the training/testing datasets were exactly the same, except the predictive column was treated as real-values, i.e. -2, -1, 0, 1, 2 (instead of treating these as symbolic classes "-2", "-1", .., in the above). The majority of the training parameters were the same, except for the distribution types.

3.2.4. Ranking by importance of the features in the consensus models

Ranking the scores in terms of their correlations with predicted class (Table 2, Figure 5) suggested a high level of redundancy among the features, since all of the correlations were similarly high. Alternatively, ranking by feature importance in the models more distinctly shows the most important and unique contributions made by each feature.

It is important that the models have different most contributing features, signifying good diversity among the base learners. As shown in Figure 6, for the GLM model the most contributing feature by magnitude is *mean_score* ("x31"), followed by much smaller contributions from *Eigen-raw* ("x24") and *Eigen-PC* ("x25"). For GBM the top-5 most important are *VEST4, Meta-LR, Eigen-raw, CADD, MutPred* (x10,x12,x24,x13,x15). For DRF the top-5 contributing features are *VEST4, CADD, M-CAP, MutPred, MVP* (x10,x13,x20,x15,x16).

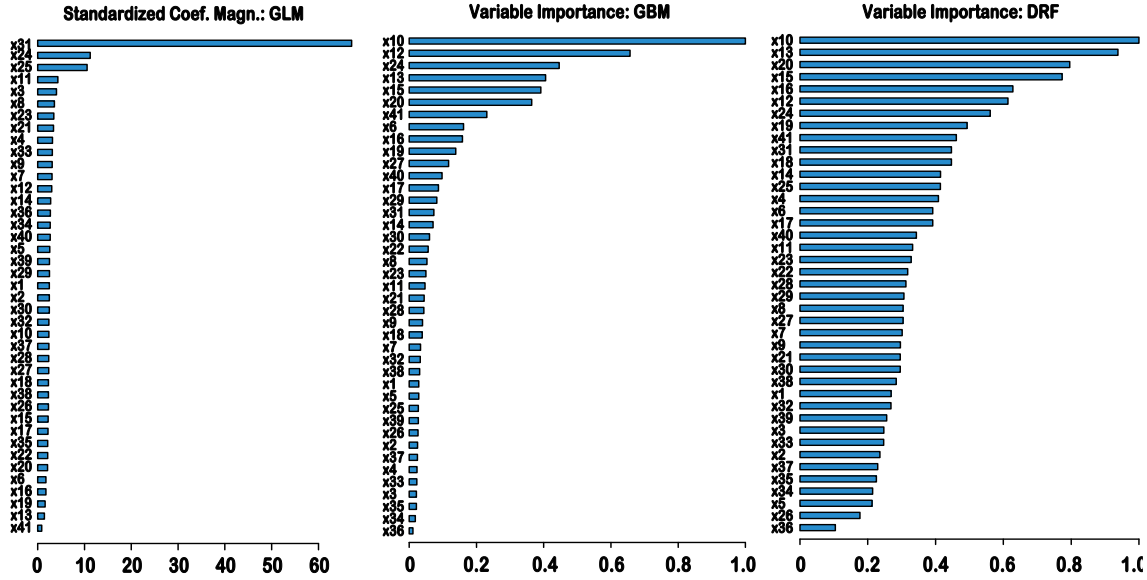


Figure 6. Ranking of the features by importance in the three base-learner models: GLM (left panel), GBM (middle panel) and DRF (right panel). The ordering from top to bottom corresponds to decreasing importance in the respective models.

3.3. Prediction of Consensus Pathogenicity

3.3.1. cVEP values for all nsSNVs by chromosomes

Table 4 shows the proportions of predicted cVEP classes for all nsSNV variants grouped by the human chromosomes 1-22,X,Y, and M, and cVEP classes across all chromosomes combined. The percent values are similar for the majority of chromosomes. There are around 34-41% and 11-22% in the 'Likely_benign' and 'Benign' classes, around 16-18% and 4-5% in the 'Pathogenic' and 'Likely_pathogenic' classes respectively, and with a quarter in the 'Uncertain_significance' class. The situation is clearly different for the Y and M chromosomes: Y has much fewer pathogenic and much more benign variants. It is especially notable that M has nearly no variants in the 'Likely' and 'Uncertain' classes, with all variants either 'Pathogenic' (40%) or 'Benign' (55%). It is worth noting that the "control" dataset has considerably higher percentages for the benign classes and about half as many variants in the pathogenic classes, in comparison to the values for most individual chromosomes and to all chromosomes combined. Since the distribution of variants by AAF follows a gamma-distribution and the majority of variants are very-rare/ultra-rare, the class percentages of Table 4 must be determined by how many of such variants are in each class, however, we do not have the AAF values for 94% of all variants to confirm this hypothesis.

Table 4. Proportions of cVEP classes of the nsSNVs grouped by the chromosomes.

Chr	Pathogenic	Likely_pathogenic	Uncertain	Likely_benign	Benign
1	0.170	0.043	0.252	0.374	0.162
2	0.171	0.051	0.308	0.350	0.120
3	0.186	0.049	0.273	0.381	0.111
4	0.177	0.039	0.256	0.389	0.138
5	0.185	0.046	0.273	0.377	0.120
6	0.170	0.039	0.241	0.392	0.159
7	0.170	0.041	0.248	0.373	0.169
8	0.175	0.041	0.283	0.367	0.133
9	0.166	0.046	0.254	0.369	0.166
10	0.179	0.043	0.263	0.370	0.145
11	0.163	0.042	0.252	0.388	0.155
12	0.174	0.056	0.278	0.381	0.112
13	0.175	0.047	0.302	0.341	0.136
14	0.162	0.056	0.265	0.386	0.132
15	0.175	0.047	0.271	0.361	0.146
16	0.168	0.043	0.260	0.373	0.156
17	0.181	0.054	0.271	0.355	0.138
18	0.178	0.041	0.258	0.388	0.135
19	0.136	0.034	0.189	0.418	0.223
20	0.183	0.049	0.274	0.375	0.119
21	0.169	0.036	0.260	0.385	0.150
22	0.178	0.044	0.261	0.381	0.136
X	0.188	0.061	0.183	0.391	0.177
Y	0.070	0.001	0.018	0.330	0.581
M	0.404	0.000	0.000	0.042	0.553
All chromosomes	0.172	0.046	0.258	0.377	0.148
GnomAD,“control”	0.100	0.020	0.237	0.433	0.211

3.2. Variation in cVEP by gene indispensability score

To assess differences in predicted variant pathogenicities as a function of the gene indispensability score *I*, the variants were selected from the entire GnomAD dataset in three ranges (see Table 5).

From the table it is clear that the mutation-tolerant group has a considerably higher proportion of predicted benign or likely benign variants at around 80%. In contrast, the essential group has a smaller proportion of predicted benign variants, with most classified as uncertain or likely benign. 66% of the intermediate group’s variants were predicted to be as benign or likely benign.

Table 5. Proportions of predicted pathogenicity classes for three gene indispensability ranges.

Gene indispensability ranges	*Pathogenic*	*Likely_pathogenic*	*Uncertain*	*Likely_benign*	*Benign*
Low: mutation-tolerant genes 0.0<*I*<0.3	0.105	0.010	0.091	0.489	0.305
intermediate by *I* genes 0.3<*I*<0.7	0.122	0.017	0.197	0.499	0.165
High: essential genes 0.95<*I*<1.0	0.104	0.038	0.439	0.335	0.085

3.3.3. Application of cVEP to pathogenic genes

The proportion of variants predicted to be pathogenic in a selection of known pathogenic genes (Table 6) was, as expected, higher than in the total GnomAD dataset. The highest proportion of pathogenic or likely pathogenic variants for cardiac disorder genes can be found for the MYH7 (61%) and FBN1 (52%) genes. The proportion of variants with predicted uncertain significance is relatively large in this group, with the highest of 74% for DSP and 69% for PKP2. For recessive genes the pathogenic proportion is highest for HBB (70%) and GJB2 (62%).

Table 6. Proportions of pathogenicity classes for the selected pathogenic genes.

Genes	*Pathogenic*	*Likely_pathogenic*	*Uncertain*	*Likely_benign*	*Benign*
Cardiac disorder genes					
fbn1	0.156	0.362	0.430	0.051	0.001
myh7	0.103	0.504	0.374	0.020	0.000
tnni3	0.187	0.203	0.484	0.085	0.042
tnnt2	0.152	0.259	0.532	0.041	0.017
mybpc3	0.166	0.098	0.566	0.163	0.007
dsg2	0.110	0.021	0.580	0.274	0.015
pkp2	0.127	0.048	0.686	0.133	0.006
dsp	0.099	0.036	0.743	0.118	0.004
dsc2	0.160	0.023	0.452	0.339	0.026
Recessive genes					
hbb	0.694	0.013	0.163	0.095	0.035
gjb2	0.333	0.284	0.349	0.031	0.004
cftr	0.320	0.160	0.473	0.042	0.005
mefv	0.124	0.009	0.227	0.563	0.077

4. Discussion

The approach presented here enables the two-step prioritization of variants, whereby initial classification into the cVEP classes is followed by filtering in the selected class, e.g. 'pathogenic' class, using the continuous-value consensus score, or any of the best performing scores. The study ranks the features by importance, which can be used as a guidance when selecting scores for the variant filtering. Many NGS applications, especially in the clinical domain, would benefit from directly using a predicted pathogenicity classification, which is compatible with the clinically and experimentally determined ClinVar database classes, and defined according the ACMG/AMP standards and guidelines for use in clinical practice. The presented results demonstrate that prediction of class is of high accuracy, with error rate being of around 1% for cVEP 2-class classification. Moreover,

Peer-reviewed version available at Genes 2020, 11, 1076; doi:10.3390/genes11091076

prediction performance is likely to improve over time with more data available for training and with the development of better machine learning tools. Clearly, it would be preferable to use actual clinical, experimental in vitro or in vivo test results to determine variant pathogenicity instead of using a prediction score. However, such data are currently unavailable for more than 99% of all potential nsSNV mutations. This situation is likely to improve in the future, e.g. it might become possible to obtain direct measures of pathogenicity using high-throughput in vitro cell screening assays of effects of synthetic mutations, e.g. using the CRISPR approach [25], leading to an increase in the percentage of known annotations. However, there are types of variants, such as null mutations, which appear to be too pathogenic to exist in a living cell or will be below a screen detectability level.

The rankings by feature importance in the three base-models provides three differing selections of the best scores for variant filtering. The GBM approach sequentially selects features one-by-one using such algorithm: a next selected feature is orthogonal to the preceding in the sequence features and contributes the most to classification. Thus, the GBM list of top-ranking features has much less redundancy compared to the DRF list, where selection of the most-contributing features is done without an orthogonality requirement. It is interesting that the GLM list top-ranks the 'Mean' feature, which is computed as a simple average of the features available for each variant. The mean score should indeed have theoretically lower by factor of 40 error level compared to each individual score feature used alone, due to averaging of noise, which corresponds to its high rank in the linear regression model. However, this mean-rankscore is less important for DRF and GBM base-approaches. The most contributing to the resulting ensemble or consensus scores are VEST4, CADD and MutPred, which are intersection of the top-5 most-important features for GBM and DRF. These or any the top-5 features in the base-models are proposed for use in variant filtering.

A novel score, gene indispensability was included into predictors for several reasons. On the one hand, there is a considerable number of potentially deleterious (i.e. having high pathogenicity scores) variants located in the mutation-tolerant genes. Clearly, such variants, if having high allele frequencies, could be assumed neutral based on that measure, or could be associated with no negative effects in the cells, and as a result, were attributed into the 'Benign' class in the ClinVar database. On the other hand, moderately deleterious variants (i.e. having medium-level pathogenicity scores) in the essential genes can be potentially associated with very adverse effects in cells, and were included into the 'Pathogenic' ClinVar class. A relatively large number of such cases could be a major source of errors, when training is based on pathogenicity scores alone. Thus, it is not surprising that the indispensability score had contributed significantly to the improvement in accuracy, since this score ("x31") was listed 10th in the DRF ranking by importance. In our initial classification tests (data not shown) the addition of this score has reduced the amount of total classification errors more than five-fold: from initial ca 10% (without Indispensability) to current <2%.

The method presented here could be subject to criticism on a number of fronts. Firstly, one might argue that the effect of a particular mutation in a living cell depends on many other factors, not necessarily caused by the mutation per se, e.g. may be determined by gene regulation or a combination of many mutations including in other genes, by an accumulation of history of direct and indirect effects, or by other external factors. It is known that the pathogenicity of some hereditary mutations exhibits itself only with age, such as in BRCA1,2 genes. Each mutation might contribute a small fraction to overall risk of developing a pathological condition. This can render predictions based on association of a mutation and its causative manifestation either infeasible or simply too ambiguous. In this approach it translates to many predictions falling into the "uncertain significance" class, in which variants could be conditionally pathogenic depending on other unknown factors. Secondly, some of the pathogenicity prediction scores, in particular those based on gene sequence conservation, are measures obtained from alignments of genetic sequences in the mammalian families, of which the human population is only a part. This implies an indirect relationship between the values of mammalian conservation scores and a likelihood to observe clinical/biological pathogenicity effects in humans. Nonetheless, when a gene variant is conserved (and common) in a

mammalian species family, it might be common in humans as well, and thus is more likely to be considered benign. The indirect negative connection between variant allele frequency and its pathogenicity, however, stems from a long history of negative evolutionary selection pressure on dysfunctional or pathogenic variants, and positive selection for those which are beneficial. Overall, the evolutionary link between all three characteristics of an allele variant: human frequency, mammalian family conservation and pathogenicity is undeniable: pathogenic variants are less conserved in families and less frequent in genetic sequences, while beneficial and benign mutations are more prevalent. Thirdly, although these results are already useful today, we suggest treating this attempt to classify all variants as preliminary, and to use the predictions with a degree of caution, similar to the usage of the allele frequencies and the impact-effect predictive scores. Nonetheless, the consensus VEP class has better accuracy in comparison to each of the individual input scores and might be better than using the population-derived AAFs for ultra-rare and very-rare variants, which currently suffer from a significant bias in sampling of human populations. Finally, the genome-wide pathogenicity class predictions may become a novel source of leads of causative pathogenic variants and genes suitable for further testing in addition to those discovered in GWAS.

5. Conclusions

In this work, predicted pathogenicity consensus scores and classifications were obtained for all potentially possible non-synonymous human exome SNVs, using a machine learning method with primary inputs of 39 functional impact scores. The input of the gene indispensability score accounted for gene variants differences in essentiality and mutation-tolerance. We showed that nearly all of the functional, conservation and impact scores used in this study correlate with each other and the evidence pathogenicity data on average, and anti-correlate with human population allele frequencies. The fact suggests that these measures are indeed highly related, and that using a combination approach is an appropriate method of predicting variant pathogenicity. Assessment of the performance of the methods using five-fold cross-validation on the ClinVar dataset confirmed the reliability and robustness of prediction. The good performance of the method can be attributed to the consensus combination of a wide array of multiple conservation and impact effect scores, which are based on each score's respective domain of evidence, e.g. all protein 3D structures for CADD. Another advantage of the method is that the models are trained based on multiple available pathogenicity classes, in comparison to some of the existing methods utilizing only two (pathogenic/benign) classes. Finally, the predicted consensus classification are aimed to help to distinguish between pathogenic, uncertain_significance and benign variants for all human exome nsSNVs in biological and clinical applications of Whole Exome Sequencing.

Supplementary Materials: The following are available online at www.mdpi.com/xxx/s1-3, Table S1 with the list of the impact effect scores used for input in the machine learning procedure; Table S2 with the individual parameters of the base- and meta-learners; Tables S3a-e with the outputs of the model fits containing cross-validations metrics for the base- and ensemble-learners. Availability: the predicted consensus VEP (cVEP) class scores for the 84,013,118 human exome nsSNVs are available here: https://cvep.imbi.uni-freiburg.de/cvep.tgz, in the form of 1.2Gb tar-gzip archive, in BED (Browser Extensible Data) format split by the chromosome regions; the archive can be searched using provided Java program by input of hg19, hg38 (1-based coordinates) or gene, chromosome; it is compatible with any of the variant annotation tools that can use the dbNSFP database Version 4.0.

Author Contributions: Conceptualization, methodology, data curation, software programming, calculations, analysis, results validation, visualization, V.J.; writing—original draft preparation, V.J., J.B.; writing—review and editing, M.B., Me.B., H.B.; supervision, funding acquisition, M.B., Me.B.,H.B. All authors have read and agreed to the published version of the manuscript, please turn to the CRediT taxonomy for the term explanation.

Funding: The research was funded by the Medical Informatics Funding Scheme (Medizininformatik-Initiative, MI-I)MIRACUM from the German Federal Ministry of Education and Research (Bundesministerium für Bildung und Forschung, BMBF).

Acknowledgments: The article processing charge was funded by the Baden-Wuerttemberg Ministry of Science, Research and Art and the University of Freiburg in the funding programme Open Access Publishing.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Pagon, R.A., *GeneTests: Integrating Genetic Services into Patient Care*.* The American Journal of Human Genetics, 2007. **81**(4): p. 658-659.
2. Soussi, T., et al., *High prevalence of cancer-associated TP53 variants in the gnomAD database: A word of caution concerning the use of variant filtering.* Hum Mutat, 2019. **40**(5): p. 516-524.
3. Bao, R., et al., *Review of current methods, applications, and data management for the bioinformatics analysis of whole exome sequencing.* Cancer Inform, 2014. **13**(Suppl 2): p. 67-82.
4. Hoskinson, D.C., A.M. Dubuc, and H. Mason-Suares, *The current state of clinical interpretation of sequence variants.* Curr Opin Genet Dev, 2017. **42**: p. 33-39.
5. Richards, S., 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.* Genet Med, 2015. **17**(5): p. 405-24.
6. Lek, M., et al., *Analysis of protein-coding genetic variation in 60,706 humans.* Nature, 2016. **536**(7616): p. 285-91.
7. Karczewski, K.J., et al., *Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes.* bioRxiv, 2019: p. 531210.
8. Song, W., et al., *Exploring the landscape of pathogenic genetic variation in the ExAC population database: insights of relevance to variant classification.* Genetics in Medicine, 2016. **18**(8): p. 850-854.
9. Huang, K.-l., et al., *Pathogenic Germline Variants in 10,389 Adult Cancers.* Cell, 2018. **173**(2): p. 355-370.e14.
10. Ioannidis, N.M., et al., *REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants.* Am J Hum Genet, 2016. **99**(4): p. 877-885.
11. Hassan, M.S., et al., *Evaluation of computational techniques for predicting non-synonymous single nucleotide variants pathogenicity.* Genomics, 2019. **111**(4): p. 869-882.
12. Niu, B., et al., *Protein-structure-guided discovery of functional mutations across 19 cancer types.* Nat Genet, 2016. **48**(8): p. 827-37.
13. Kircher, M., et al., *A general framework for estimating the relative pathogenicity of human genetic variants.* Nature genetics, 2014. **46**(3): p. 310-5.
14. Bailey, M.H., et al., *Comprehensive Characterization of Cancer Driver Genes and Mutations.* Cell, 2018. **173**(2): p. 371-385.e18.
15. Schiemann, A.H. and K.M. Stowell, *Comparison of pathogenicity prediction tools on missense variants in RYR1 and CACNA1S associated with malignant hyperthermia.* Br J Anaesth, 2016. **117**(1): p. 124-8.
16. van der Laan, M.J., E.C. Polley, and A.E. Hubbard, *Super learner.* Statistical applications in genetics and molecular biology, 2007. **6**: p. Article25.
17. *H2O.ai Team. H2O R Package Documentation.* 2018 [cited 2020 23.04.2020]; Available from: http:// h2o-release.s3.amazonaws.com/h2o/latest_stable_Rdoc.html.
18. Nelder, J.A. and R.W.M. Wedderburn, *Generalized Linear Models.* Journal of the Royal Statistical Society, 1972. **Series A**(General): p. 370-384.

19. Jane, E., J.R. Leathwick, and T. Hastie, *A Working Guide to Boosted Regression Trees.* Journal of Animal Ecology 2008. **77**(4): p. 802-813.
20. Geurts, P., D. Ernst., and L. Wehenkel, *Extremely randomized trees.* Machine Learning, 2006(63(1)): p. 3-42.
21. Gneiting, T. and A.E. Raftery, *Strictly Proper Scoring Rules, Prediction, and Estimation.* Journal of the American Statistical Association, 2007. **102**(477): p. 359-379.
22. Landrum, M.J., et al., *ClinVar: public archive of interpretations of clinically relevant variants.* Nucleic Acids Research, 2015. **44**(D1): p. D862-D868.
23. Liu, X., et al., *dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs.* Hum Mutat, 2016. **37**(3): p. 235-41.
24. Khurana, E., et al., *Interpretation of Genomic Variants Using a Unified Biological Network Approach.* PLOS Computational Biology, 2013. **9**(3): p. e1002886.
25. Wang, T., et al., *Identification and characterization of essential genes in the human genome.* Science, 2015. **350**(6264): p. 1096-101.