*Published online 3 October 2016*

*Nucleic Acids Research, 2017, Vol. 45, No. 3* ***e13***
*doi: 10.1093/nar/gkw886*

# IMHOTEP—a composite score integrating popular tools for predicting the functional consequences of non-synonymous sequence variants

**Carolin Knecht[1], Matthew Mort[2], Olaf Junge[1], David N. Cooper[2], Michael Krawczak[1] and Amke Caliebe[1,*]**

[1]Institute of Medical Informatics and Statistics, Kiel University, 24105 Kiel, Germany and [2]Institute of Medical Genetics, School of Medicine, Cardiff University, Cardiff, CF14 4XN, UK

Received June 23, 2015; Revised September 19, 2016; Accepted September 26, 2016

## ABSTRACT

**The *in silico* prediction of the functional consequences of mutations is an important goal of human pathogenetics. However, bioinformatic tools that classify mutations according to their functionality employ different algorithms so that predictions may vary markedly between tools. We therefore integrated nine popular prediction tools (PolyPhen-2, SNPs&GO, MutPred, SIFT, MutationTaster2, Mutation Assessor and FATHMM as well as conservation-based Grantham Score and PhyloP) into a single predictor. The optimal combination of these tools was selected by means of a wide range of statistical modeling techniques, drawing upon 10 029 disease-causing single nucleotide variants (SNVs) from Human Gene Mutation Database and 10 002 putatively 'benign' non-synonymous SNVs from UCSC. Predictive performance was found to be markedly improved by model-based integration, whilst maximum predictive capability was obtained with either random forest, decision tree or logistic regression analysis. A combination of PolyPhen-2, SNPs&GO, MutPred, MutationTaster2 and FATHMM was found to perform as well as all tools combined. Comparison of our approach with other integrative approaches such as Condel, CoVEC, CAROL, CADD, MetaSVM and MetaLR using an independent validation dataset, revealed the superiority of our newly proposed integrative approach. An online implementation of this approach, IMHOTEP ('Integrating Molecular Heuristics and Other Tools for Effect Prediction'), is provided at http://www.uni-kiel.de/medinfo/cgi-bin/predictor/.**

## INTRODUCTION

Unravelling the genetic basis of inherited diseases has been a major focus of human genetics research for decades. Over the last 10 years, the development of fast, accurate and inexpensive DNA sequencing technologies has brought within reach the comprehensive characterization of all genetic variants carried by a given individual, many of them rare or even private. As a consequence, one of the major challenges of human genetics research has been to distinguish those relatively few sequence variants that are functionally significant from the many thousands that are not. To tackle this problem, a number of 'pathogenicity prediction tools' have been developed that profess *in silico* assessment of the structural and functional impact of a given gene mutation upon its corresponding gene product (1). Particular progress has been made in this regard for non-synonymous (ns) variants because their consequences for the structural and functional characteristics of a given gene product are the easiest to infer.

Currently available prediction tools for ns variants employ different algorithms and exploit different types of information, including amino acid sequence, limited DNA sequence context (e.g. CpG) and protein structure as well as functional annotation (2). In addition, the performance of the tools depends critically upon the training data used for their development (3,4). Consequently, predictions made by different tools can differ greatly when applied to one and the same variant, which is why in practice the combined use of different tools has been recommended (2,5,6).

A number of software packages are available for the annotation of DNA genetic variation, including dbNSFP (7,8) and ANNOVAR (9). These annotation tools also provide output of selected prediction tools, including PolyPhen-2 (10) and SIFT (11) etc. However, to our knowledge, only dbNSFP also integrates the output of different prediction tools (MetaSVM and MetaLR scores) so as to yield a single 'consensus' prediction that should be useful in practice, for example, when evaluating the clinical significance of a newly

*To whom correspondence should be addressed. Tel: +49 431 500 30711; Fax: +49 431 500 30704; Email: caliebe@medinfo.uni-kiel.de

detected genetic variant. Moreover, in a research context, the inconsistent classification of variants may render their prioritization for further analysis difficult. Whilst 'consensus' prediction is undoubtedly useful, it is not straightforward computationally for two main reasons. First, different tools usually employ different definitions of whether a variant is 'consequential' or 'inconsequential'. Second, the scores generated by different tools usually scale differently and are therefore not directly comparable.

The above notwithstanding, all currently available prediction tools are designed to reflect the influence of the genetic variant of interest on the respective gene product. An efficient strategy is therefore required to standardize and combine the quantitative output of the different tools into a single numerical value that could form the basis of a consensus prediction. Integration approaches such as Condel (12), CoVEC (13), CAROL (14), MetaSVM and MetaLR (15) provide some kind of solution to this problem and have consistently been found to perform better than single tools do on their own (13). However, most of the pipelines only integrate tools at the level of running batch queries (12,16) so that interactive tools like SNPs&GO (17) and MutPred (18) had to be excluded, despite their evidently good performance (5).

In the present study, we combined the output of nine popular prediction tools, namely PolyPhen-2 (10), SNPs&GO (17), MutPred (18), SIFT (11), MutationTaster2 (19), Mutation Assessor (20) and FATHMM (21) as *sensu stricto* predictors, and the conservation-based PhyloP (22) and Grantham Score (23), into a single composite score. In contrast to other integration approaches, we took different statistical methods systematically into consideration to generate such a score, including (i) simple summation (binary and continuous), (ii) majority vote, (iii) logistic regression, (iv) decision tree and (v) random forest. Our final choice of methods drew upon the analysis of 10 029 putatively consequential ns single nucleotide variants (nsSNVs) from the Human Gene Mutation Database (HGMD) (24), and 10002 putatively 'inconsequential' nsSNVs obtained using the UCSC Genome Browser (http://genome.ucsc.edu/cgi-bin/hgTables). We also demonstrate that the inclusion of imputed predictions does not notably impede tool integration, which implies that variants that cannot be handled by one or the other tool can still be assigned a valid composite score. Finally, we validated our newly proposed integration approach in an independent set of variants and compared the results to the output of established analysis pipelines. An online implementation of our integration approach is provided under the acronym IMHOTEP ('Integrating Molecular Heuristics and Other Tools for Effect Prediction') at http://www.uni-kiel.de/medinfo/cgi-bin/predictor/. Imhotep was an Egyptian polymath who was chief minister to Djoser, the second king of the third dynasty who reigned 2630-2611 BCE. Imhotep, whose name means 'he who cometh in peace', is thought to have been the architect of the step pyramid of Saqqara. He was described as being 'the first figure of a physician to stand out clearly from the mists of antiquity' by the eminent Canadian physician Sir William Osler (http://www.gutenberg.org/files/1566/1566-h/1566-h.htm) (25). In 525 BCE, Imhotep was deified as a god of medicine. He is associated with the Greek god of medicine, Asclepios (https://www.britannica.com/biography/Imhotep) (26).

## MATERIALS AND METHODS

### Mutation data

Two sets of nsSNVs with putatively known functional consequences were used to construct and evaluate a composite prediction score. The first set of data was retrieved by means of the UCSC Genome Browser (http://genome.ucsc.edu/cgi-bin/hgTables, table: snp135Common) and filtered according to the following criteria: (i) minor allele frequency (MAF) >10%, (ii) missense variant, (iii) validated by at least two methods (e.g. presence in both HapMap and 1000 Genomes project), (iv) no more than two alleles reported so far and (v) absent from HGMD. For some of the variants, different transcripts of the affected gene were found to be logged in UCSC. In these instances, the variant in question was mapped to the longest transcript available. This procedure yielded 10 801 UCSC nsSNVs from 5666 genes suitable for further analysis. Owing to their high MAF, these variants were deemed likely to lack any significant deleterious effect and were therefore classified as 'inconsequential' for the purpose of the present study. We deliberately chose not to label these variants 'neutral' because 'neutral' has an established meaning in population genetics (27). The corresponding amino acid substitutions were derived from dbSNP (http://www.ncbi.nlm.nih.gov/SNP/) and the underlying amino acid sequences were downloaded from UCSC. All information was combined using an in-house Perl script. The second set of data was provided by HGMD Pro 2015.3 (24) and comprised 12 355 nsSNVs from 2165 genes that were added to HGMD after 1 January 2013. The SNVs were annotated as 'disease-causing' in HGMD and were therefore deemed to be 'consequential' for the purpose of the present study. We considered only post-2013 HGMD data to ensure that the SNVs under study were not used for the development of any prediction tool investigated here, thereby avoiding an overlap between the testing and training sets.

Integration models were validated by 10-fold cross-validation using cross-validation datasets of ∼1000 consequential and inconsequential variants each. To ensure that the cross-validation datasets contained roughly the same number of genes and that all variants of a gene were in one dataset, we pursued the following strategy separately for each type of variant (consequential and inconsequential): for each cross-validation dataset, genes were repeatedly drawn at random without replacement and all variants in a given gene allocated to the respective cross-validation dataset. This procedure was repeated until the total number of variants of the variant type in question exceeded 1000 in the dataset. If the excess number of variants was smaller than the deficit without the last gene included (i.e. if the number of variants including the last gene minus 1000 was smaller than 1000 minus the number of variants without the last gene), all variants of the last gene were indeed added to the dataset. Otherwise, the last gene and its variants were returned to the pool. This procedure resulted in 10 equally large cross-validation sets with comparable number of genes. The resulting 10 datasets, henceforth referred

to as 'the cross-validation data', contained a total of 10 002 putatively inconsequential nsSNVs from UCSC and 10 029 putatively consequential nsSNVs from HGMD (see Supplementary Table S1 for the exact number of variants and genes per individual dataset).

As with the cross-validation data, an additional dataset was generated for use as an external validation base and for comparison with other integration approaches. This dataset, termed 'the external validation data', contained 800 consequential and 799 inconsequential variants each. Note that cross-validation and external validation data were generated such that they did not overlap in terms of the variants included. If required, the combination of both datasets will be referred to as 'the complete data'.

### Prediction tools

We considered nine prediction tools in our study, namely PhyloP (22), Grantham Score (23), PolyPhen-2 (HumVar model) (10), SNPs&GO (17), MutPred (18), SIFT (11), MutationTaster2 (19), Mutation Assessor (20) and FATHMM (inherited disease model) (21). These tools were selected because of their popularity and their previously demonstrated fine performance (5).

Note that PhyloP and the Grantham Score are not prediction tools *sensu stricto* because they do not classify variants as either consequential or inconsequential. Instead, PhyloP evaluates the difference in evolutionary conservation between nucleotides from different species at a given genomic position whilst the Grantham Score quantifies the biochemical difference between two amino acids. This notwithstanding, for a given nsSNV, the scores of both tools are likely to be related to the impact of the nsSNV upon the functionality of the respective gene product, which is why both tools have been used extensively to prioritize genetic variants for functional follow-up (28). Moreover, both tools are embedded into other integrative bioinformatics platforms, including ANNOVAR (9). To facilitate consistent binary categorization in our study, we used classification schemes previously described for PhyloP (7) and the Grantham Score (29).

Other than PhyloP and the Grantham Score, the remaining tools considered in our study were genuine effect predictors that use different statistical methods to classify variants as either consequential or inconsequential. Thus, PolyPhen-2 and MutationTaster2 employ naïve Bayes classifiers, which are usually both simple and robust. SNPs&GO and MutPred follow classical machine learning approaches, namely support vector machine and random forest, respectively, with SNPs&GO also utilizing Gene Ontology information. A more sophisticated hidden Markov model is implemented in FATHMM to represent multiple alignments of homologous protein sequences. SIFT is also based upon sequence conservation information and is the only prediction tool *sensu stricto* that needs no training on variant data of known effect. Finally, Mutation Assessor involves a statistical technique termed 'combinatorial entropy optimization' to find an 'optimal' hierarchical clustering of proteins into subfamilies. In addition to a binary classification into consequential or inconsequential, all tools considered here also provide some kind of continuous score (Table 1).

The output of PhyloP was generated by means of the UCSC Genome Browser (using the phyloP46wayAll table of hg19). The Grantham Score was calculated with R on the basis of Grantham's original data (http://www.genome.jp/dbget-bin/www_bget?aax2:GRAR740104). PolyPhen-2 (http://genetics.bwh.harvard.edu/pph2/), SIFT (http://sift.bii.a-star.edu.sg), MutationTaster2 (http://www.mutationtaster.org/), Mutation Assessor (http://mutationassessor.org/r3/) and FATHMM (http://fathmm.biocompute.org.uk/inherited.html) were accessed by way of batch queries. Owing to the large number of queries necessary, the output of MutPred (http://mutpred.mutdb.org/) had to be generated locally by staff from the School of Informatics and Computing, Indiana University. The output of SNPs&GO (http://snps-and-go.biocomp.unibo.it/snps-and-go/) was obtained using a Perl script. All result files were merged and processed with R (R v.3.2.2, https://www.R-project.org.) (30).

To ensure comparability between the nine tools, we normalized each score to the interval [0,1] such that the maximum evidence in favor of a functional consequence equalled unity. For PhyloP, PolyPhen-2, SNPs&GO, MutPred and MutationTaster2, the corresponding output already had this format. The output of SIFT was transformed by simple inversion, $x_{SIFT\_new} = 1\text{-}x_{SIFT}$. The Grantham Score, was normalized by division through its maximum value, i.e. $x_{Grantham\_new} = x_{Grantham}/215$. Since the ensuing threshold $t_0$ distinguishing consequential from inconsequential variants still differed between tools, we applied a second transformation to each score so as to allow use of a universal threshold $t_{0\_new} = 0.5$. If necessary, score x was thus replaced by $x_{new} = 0.5 \times x/t_0$ if $x \leq t_0$ or by $x_{new} = 0.5 + 0.5 \times (x\text{-}t_0)/(1\text{-}t_0)$ otherwise. Since the scores produced by Mutation Assessor and FATHMM can also be negative, the output of these tools was normalized as follows: $x_{MutAss\_new} = 0.5 \times (x_{MutAss} + 5.545)/7.483$ if $x_{MutAss} \leq 1.938$ and $x_{MutAss\_new} = 0.5 + 0.5 \times (x_{MutAss}\text{ -}1.938)/3.999$ otherwise; $x_{FATHMM\_new} = 0.5 \times (10.64\text{-}x_{FATHMM})/12.14$ if $x_{FATHMM} > -1.5$ and $x_{FATHMM\_new} = 0.5\text{-}0.5 \times (x_{FATHMM} +1.5)/14.63$ otherwise. MutationTaster2 offers additional functionality to assess the pathogenicity of variants. Variants are predicted as 'polymorphism_automatic' if either all three corresponding genotypes are observed in at least one HapMap population, or a variant was found in the homozygous state in more than four participants from the 1000 Genomes Project (31). Moreover, if a variant is termed as 'probable-pathogenic' or 'pathogenic' in ClinVar (32) it is labeled as 'disease_causing_automatic'. Despite the automatic prediction, the score of the Bayes classification model is given as well. Since the automatic classifications involving 'looking up' a variant in existing databases is an intrinsic feature of MutationTaster2, we decided to include the automatic predictions in our evaluation. For the generation of the normalized score, we used both the binary prediction (irrespective of being automatic or not) and the Bayes score: $x_{MutTas2\_new} = 0.5 - 0.5 \times x_{MutTas2}$ if prediction is inconsequential and $x_{MutTas2\_new} = 0.5 + 0.5 \times x_{MutTas2}$ otherwise.

**Table 1.** Description of the prediction tools studied

| Prediction tool | Score range | Classification scheme | Additional output | Algorithmic basis |
|---|---|---|---|---|
| PhyloP | [0;1]* | >0.95 (conserved)<br>≤0.95 (non-conserved) | | phylogenetic analysis |
| Grantham Score | [0;215] | ≤50 (conservative)#<br>51-100 (moderately conservative)<br>101-150 (moderately radical)<br>≥151 (radical) | | biochemical distances between amino acids |
| PolyPhen-2 | [0;1] | probably damaging<br>possibly damaging<br>benign<br>unknown | false positive rate<br>true positive rate | naïve Bayes classifier |
| SNPs&GO | [0;1] | ≤0.5 (neutral)<br>>0.5 (disease) | reliability index | support vector machine |
| MutPred | [0;1] | ≤0.5 (not deleterious)<br>>0.5 (deleterious)<br>>0.75 (more confidently deleterious) | hypotheses about molecular cause of amino acid substitution | random forest |
| SIFT | [0;1] | ≥0.05 (tolerated)<br><0.05 (damaging) | median information<br>number of sequences aligned at position | protein sequence alignment |
| MutationTaster2 | [0;1] | disease causing<br>disease causing automatic<br>polymorphism<br>polymorphism_automatic | affected features | naïve Bayes classifier |
| Mutation Assessor | [−5.545,5.937] | >3.5 (high functional impact)<br>1.938-3.5 (medium)<br>0.8-1.938 (low)<br>≤0.8 (neutral) | | combinatorial entropy optimization |
| FATHMM | [−16.13;10.64] | ≤−1.5 (damaging)<br>>−1.5 (tolerated) | | hidden Markov model |

*Transformation by Liu *et al*. (7).
#classification by Li *et al*. (29).

## Missing predictions

For some variants, one or more prediction tools failed to yield any output. Therefore, all analyses were performed twice. First, variants lacking an output (henceforth called 'failure variants') were excluded from further analysis, leaving 14 233 variants (8163 from HGMD, 6070 from UCSC) with complete prediction by all tools. For the second round of analysis, missing output values were imputed for failure variants as described below.

Imputation was performed with R using one of three different methods, namely (i) random imputation from a uniform distribution on [0,1], (ii) use of R package `AMELIA` (33) and (iii) use of R package `mice` (34). Both `AMELIA` and `mice` are multiple imputation methods that take into account potential interdependencies between variables, i.e. data missing for one variable are imputed from data on the other variables, if present. While `AMELIA` combines bootstrapping with the expectation-maximization algorithm, `mice` employs a Gibbs sampler. The performance of each imputation method was measured separately for each tool using the Matthews correlation coefficient (MCC, see below). The MCC was calculated on two disjoint sets of variants for each tool, namely those for which output was available ('reference'; $MCC_{ref}$) and those for which output had to be imputed ($MCC_{imp}$). Imputation can be deemed most satisfactory when the two MCCs are as similar as possible for as many tools as possible. Therefore, we aimed to minimize the Euclidean distance between $MCC_{ref}$ and $MCC_{imp}$, weighting each tool by the number of missing output values, i.e

$$\sqrt{\sum_{k=1}^{9} \frac{(MCC_{ref,k} - MCC_{imp,k})^2}{\sum_{i=1}^{9} n_i} \cdot n_k}.$$

Here, $n_i$ denotes the number of failure variants for the ith tool. PhyloP, the Grantham Score and MutPred yielded too few failure variants to sensibly call for imputation. Therefore, these three tools were omitted from the calculation of the Euclidean distance, which rendered the optimization criterion a sum over just six tools. The imputation method with the lowest Euclidean distance between $MCC_{ref}$ and $MCC_{imp}$ was eventually selected for further analysis.

## Statistical analysis

All statistical analyses were performed with R v.3.2.2 (30). Agreement analysis of categorical tool output involved computation of the proportion of agreement and Cohen's kappa, together with 95% confidence intervals, using R-package `vcd` v.1.4-1 (35). A third output category was

added to the original classification in order to allow for failure variants. Agreement analysis of continuous tool output was performed drawing upon Spearman rank correlation coefficients.

The performance of a particular tool on a given set of nsSNVs was quantified by means of the MCC (36), calculated from the number of true positives (tp), true negatives (tn), false positives (fp) and false negatives (fn) as:

$$\mathrm{MCC} = \frac{tp \times tn - fn \times fp}{\sqrt{(tp + fn)(tp + fp)(tn + fn)(tn + fp)}}.$$

The MCC is often used in the context of prediction tool validation and is generally held to provide a more balanced assessment of prediction performance than other measures such as sensitivity and specificity (37). The MCC was calculated here taking the origin of an nsSNV from either HGMD or UCSC as the gold standard.

The primary aim of our study was to develop and validate a composite score that combines the output of different prediction tools. For binary output, this combination was fashioned by either logistic regression modeling or majority vote; for continuous output, we carried out decision tree and random forest analysis. In addition, we evaluated the simple sum of scores. All composite prediction models were validated by 10-fold cross-validation in two independent runs, with failure variants either excluded or imputed.

In the logistic regression analysis, the binary classifications of a given nsSNV by the different prediction tools were treated as explanatory variables. The statistical significance of including a given tool in a regression model *via* backward selection was assessed by means of a Wald test (38) at the 5% significance level. An optimal cut-off for the model output was obtained by maximizing the sum of sensitivity and specificity with R package `pROC` v.1.8 (39). Majority vote was also based upon the binary output of different prediction tools and, starting with the inclusion of all tools, backward selection of voters was carried out by maximization of the MCC.

In the decision tree and random forest analyses, the continuous prediction scores were treated as explanatory variables. Decision tree analysis was carried out with `rpart` v.4.1-10 (40) whilst package `randomForest` v.4.6-2 was used for random forest analysis (41). For the latter, the random number of explanatory variables (e.g. the prediction scores) included at a given split was calculated using the tuneRF command.

Finally, the sum of the different (binary or continuous) tool output values was calculated without any additional modifications, and an optimal cut-off for classification was determined maximizing the sum of sensitivity and specificity with R package `pROC`.

The integration methods that performed best, both with and without imputed tool output, were random forest, decision tree and logistic regression (see below). These methods were finally trained on the complete cross-validation data. To evaluate their performance specifically for imputed data, the methods were also trained on the cross-validation data with only half of the failure variants included, and validated on the output imputed for the other half.

Random forest, decision tree and logistic regression as well as simple summation were also evaluated by 10-fold cross-validation under the inclusion of only those five tools that had the strongest impact upon prediction performance. The latter was judged from the complete data either by the Gini indices (random forest), the variable selection (logistic regression) or the contribution to the final decision tree (decision tree). In view of the apparently similar performance achieved by a combination of just the five top tools, we repeated random forest, decision tree and logistic regression and, for comparison, simple summation for these tools in the cross-validation complete data.

For external validation and comparison to other approaches, our top integration methods (random forest, decision tree and logistic regression) and six commonly used integration approaches, namely Condel (12), CoVEC (13), CAROL (14), CADD (with a threshold of 20 for the scaled score) (16), MetaSVM and MetaLR (15), were also applied to the external validation data.

Receiver operating characteristic (ROC) curves were plotted on the one hand for the nine individual tools (complete data) and on the other hand for the two integration methods random forest and logistic regression of this study, and for the six integration approaches under evaluation (Condel, CoVEC, CAROL, CADD, MetaSVM and MetaLR; all applied to the external validation data excluding failure variants). ROC curves were derived with the R package `pROC` and the corresponding area under curve (AUC) values calculated. The decision tree is a binary classification method and therefore only classifies variants as consequential or inconsequential, but does not yield a continuous score. Therefore, no ROC curves could be derived for this method.

All relevant model parameters of the logistic regression analysis and the final decision tree can be found in Figure 1 and Supplementary Table S12. Furthermore, a web server has been set-up which automatically scores variants according to our best integration models (http://www.uni-kiel.de/medinfo/cgi-bin/predictor/). The input for this web server is the prediction scores of the individual prediction tools (e.g. SIFT) which can either be entered directly or uploaded from a file. The server offers the opportunity to choose between either all nine individual tools or only the five top tools as input.

## RESULTS

### Performance and agreement of individual prediction tools

The nine prediction tools (*sensu stricto* or conservation score-based) considered in our study varied widely in terms of their classification of the 21 630 variants of the complete data used for training and external validation (Table 2). Thus, the proportion of nsSNVs predicted to be inconsequential ranged from 38% (PhyloP) to 71% (Grantham Score). Moreover, the proportion of nsSNVs for which no prediction was possible (i.e. failure variants) ranged from 0% (Grantham Score) to 25% (SNPs&GO). The performance of the different tools was also found to vary considerably (MCC: 0.21 to 0.87, Table 3) which is also highlighted by the corresponding ROC curves (Figure 2). Outstanding performance was achieved by MutationTaster2,

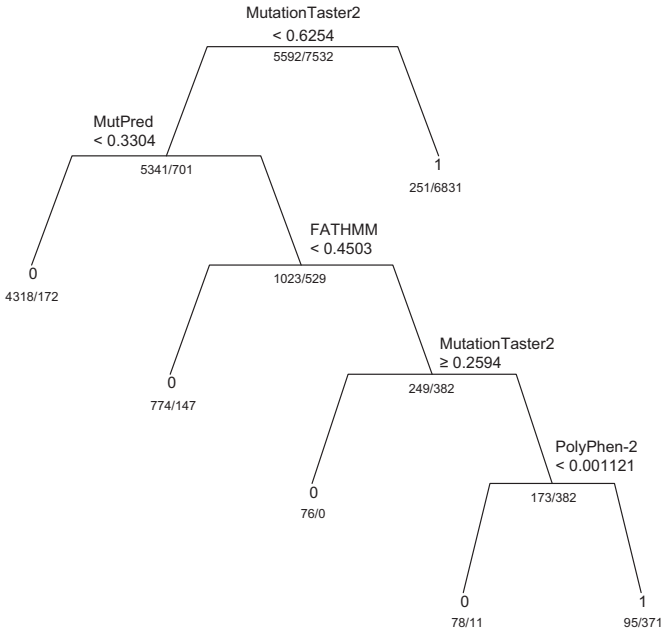


**Figure 1.** Decision tree developed on complete cross-validation data. Numbers refer to inconsequential (left) and consequential variants (right).

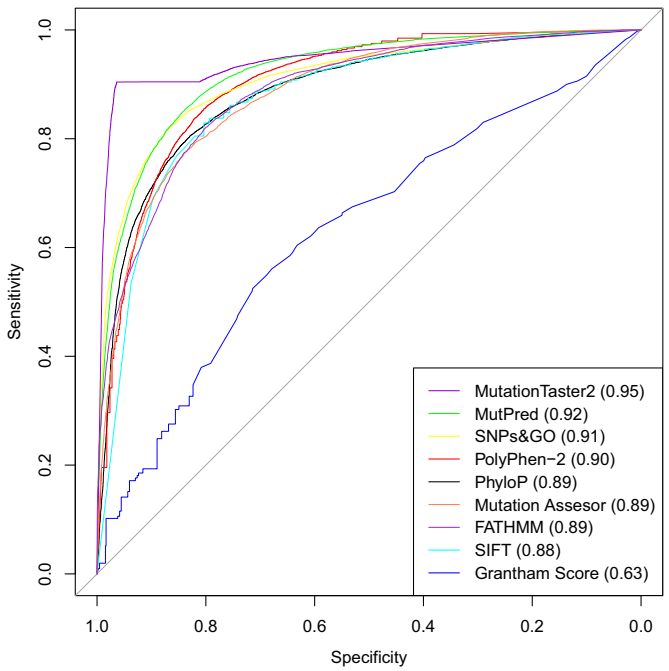


**Figure 2.** ROC curves of individual prediction tools (based upon complete data). Numbers in brackets are the AUCs of the respective tools.

which had the highest specificity (0.96) and—together with PhyloP—the highest sensitivity (0.90). Its MCC (0.87) by far exceeded that of the other eight tools. Because of its low specificity (0.66), PhyloP yielded an MCC of only 0.58. Owing to the large proportion of variants predicted to be inconsequential, the Grantham Score had high specificity (0.81) but rather low sensitivity (0.38). Consequently, its MCC was only 0.21 (the smallest MCC of all). The MCC of the other six tools ranged between 0.58 and 0.68.

The pairwise level of agreement between tools was generally low, ranging from 0.49 (SNPs&GO and Grantham Score) to 0.81 (MutPred and MutationTaster2) (Supplementary Table S2). When measured by the Spearman rank correlation coefficient (Supplementary Table S3), the Grantham Score and FATHMM exhibited the poorest agreement (0.17). Categorical and continuous tool output showed generally similar degrees of agreement.

### Missing predictions and imputation

For some variants, one or more tools failed to yield a prediction. The number of such 'failure variants' differed greatly between tools (Table 2), from none for the Grantham Score via one for PhyloP to 43 for MutPred and over 600 for MutationTaster2. PolyPhen-2 and SIFT both failed for >1000 variants whilst Mutation Assessor and FATHMM failed with >3000 predictions and SNPs&GO even failed for >5000 variants.

We adopted two different strategies to deal with missing predictions in our integrative analyses. In the first tier, a variant was excluded altogether if no prediction was possible with at least one tool. In the second tier, missing predictions were imputed using one of three different methods, namely random sampling from a uniform distribution on [0,1] and the application of either the `AMELIA` or the `mice` R package. Judged by the ensuing tool-wise sensitivity, specificity and MCC values (Supplementary Table S4), `AMELIA` and `mice` were found to perform similarly. However, since the weighted Euclidean distance between the reference and the imputed MCCs was as low as 0.08 for `AMELIA`, but 0.10 for `mice` and 0.62 for chance imputation, we opted to use `AMELIA` in our study.

### Performance of different tool integration methods

Various methods to integrate the normalized output of the different prediction tools were evaluated in our study, namely (i) simple summation (binary and continuous), (ii) majority vote, (iii) logistic regression, (iv) decision tree and (v) random forest. As judged by their MCCs, random forest performed best, closely followed by decision tree and logistic regression analysis. Simple summation showed the poorest predictive performance by far (Table 4 and Supplementary Tables S5–10).

No notable differences were observed between an exclusion of failure variants and the use of imputed data. Therefore, we shall henceforth present detailed results only for those analyses from which failure variants were excluded; the results of analyses involving imputed predictions can be found in the Supplementary Data. Binary and continuous summation performed worst (MCC equal to 0.80 and 0.79, respectively; Table 4; Supplementary Tables S5a and 6a). With majority voting, PhyloP, SNPs&GO, MutPred, MutationTaster2 and FATHMM were retained in the model in all validation rounds, sometimes together with PolyPhen-2 and Mutation Assessor, whereas the Grantham Score and SIFT were never retained. The average MCC of majority voting was 0.83 (Table 4 and Supplementary Table S7a). With logistic regression analysis, the same tools always remained

**Table 2.** Binary variant classification provided by prediction tools on the complete data (10 829 consequential, 10 801 inconsequential variants)

| Variant type | Prediction tool | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | PhyloP | Grantham Score | PolyPhen-2 | SNPs&GO | MutPred | SIFT | Mutation Taster2 | Mutation Assessor | FATHMM |
| Consequential | 13 446 (62%) | 6180 (29%) | 9888 (46%) | 6355 (29%) | 9701 (45%) | 9699 (45%) | 10008 (46%) | 8218 (38%) | 7257 (34%) |
| Inconsequential | 8183 (38%) | 15450 (71%) | 9859 (46%) | 9891 (46%) | 11887 (55%) | 10583 (49%) | 10998 (51%) | 9636 (45%) | 11008 (51%) |
| Failure | 1 (0%) | 0 (0%) | 1883 (9%) | 5384 (25%) | 42 (0%) | 1348 (6%) | 624 (3%) | 3776 (17%) | 3365 (15%) |

**Table 3.** Individual performance of prediction tools on the complete data (10 829 consequential, 10 801 inconsequential variants)

| Performance measure | Prediction tool | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | PhyloP | Grantham Score | PolyPhen-2 | SNPs&GO | MutPred | SIFT | Mutation Taster2 | Mutation Assessor | FATHMM |
| Sensitivity* | 0.90 | 0.38 | 0.82 | 0.67 | 0.79 | 0.78 | 0.90 | 0.74 | 0.68 |
| Specificity* | 0.66 | 0.81 | 0.83 | 0.95 | 0.89 | 0.84 | 0.96 | 0.87 | 0.89 |
| MCC* | 0.58 | 0.21 | 0.65 | 0.63 | 0.68 | 0.61 | 0.87 | 0.61 | 0.58 |
| Accuracy* | 0.78 | 0.59 | 0.82 | 0.80 | 0.84 | 0.81 | 0.93 | 0.80 | 0.79 |
| AUC* | 0.89 | 0.63 | 0.90 | 0.91 | 0.92 | 0.88 | 0.95 | 0.89 | 0.89 |
| Number of variants with prediction | 21 629 | 21 630 | 19 747 | 16 246 | 21 588 | 20 282 | 21 006 | 17 854 | 18 265 |

*Performance measures were calculated for each tool only from variants with prediction. MCC: Matthews correlation coefficient; AUC: area under ROC curve.

**Table 4.** Cross-validation of integration methods (all tools)

| Integration method | Matthews correlation coefficient | |
|---|---|---|
| | failure variants excluded (n = 13 124) | failure variants imputed (n = 20 031) |
| Random forest | 0.90 | 0.89 |
| Decision tree | 0.88 | 0.86 |
| Logistic regression | 0.87 | 0.87 |
| Majority vote | 0.83 | 0.83 |
| Summation binary | 0.80 | 0.79 |
| Summation continuous | 0.79 | 0.80 |

Matthews correlation coefficients are averages taken over 10 cross-validation datasets. For detailed results, see Supplementary Tables S5–10.

in the model upon backward selection, namely PolyPhen-2, SNPs&GO, MutPred, MutationTaster2 and FATHMM. The average MCC for logistic regression analysis equalled 0.87 (Table 4 and Supplementary Table S8a). Decision tree analysis performed well, with an average MCC of 0.88. Only MutationTaster2 was consistently retained, occasionally complemented by PolyPhen-2, MutPred and FATHMM (Table 4 and Supplementary Table S9a). The other tools were never included. Finally, random forest was found to perform best of all integration methods, with an average MCC of 0.90 (Table 4 and Supplementary Table S10a).

### Performance of different integration methods after 'smart selection' of tools

Owing to their superior performance in different settings, random forest, decision tree and logistic regression analysis were selected for final composite score definition. For comparison, and because of its simplicity, binary summation was also considered. Inspection of the Gini indices (random forest), tree topology (decision tree) and variable selection (logistic regression) suggests that PolyPhen-2, SNPs&GO, MutPred, MutationTaster2 and FATHMM represent the best choice of tools when it comes to composite nsSNV classification. Random forest, decision tree and logistic regression analysis performed almost as well with these five tools as with all tools combined (Table 5). Notably, the simple summation performed even better when confined to the selected tools (MCC = 0.84 versus MCC = 0.80 for all tools included). For the detailed results of the 10-fold cross-validation, see Supplementary Table S11.

### Final model definition and performance on imputed data

The three best integration methods, namely random forest, decision tree and logistic regression analysis, were finally trained on the whole cross-validation data (i.e. 7532 consequential variants from HGMD, 5592 inconsequential variants from UCSC, failure variants excluded). The parameter estimates of the logistic regression analysis and the Gini indices of the random forest model are summarized in Supplementary Tables S12 and 13, for all tools as well as for the five tools of the smart selection. The final decision tree is included in Figure 1. An online implementation of the integration approach developed in this study (IMHOTEP) is provided at http://www.uni-kiel.de/

**Table 5.** Cross-validation of integration methods (PolyPhen-2, SNPs&GO, MutPred, MutationTaster2 and FATHMM only; failure variants excluded)

| Performance measure | Integration method | | | |
|---|---|---|---|---|
| | Summation binary | Logistic regression | Decision tree | Random forest |
| Sensitivity | 0.94 | 0.93 | 0.94 | 0.95 |
| Specificity | 0.90 | 0.95 | 0.93 | 0.94 |
| MCC | 0.84 | 0.87 | 0.88 | 0.90 |
| Accuracy | 0.92 | 0.93 | 0.94 | 0.95 |

Performance measures are averages taken over 10 cross-validation datasets. MCC: Matthews correlation coefficient. For detailed results, see Supplementary Table S11.

medinfo/cgi-bin/predictor/. To assess how the three best integration approaches performed on imputed data alone, training was also carried out on data comprising all variants with complete predictions and half of the failure variants (with imputed predictions). Validation using the other half of the imputed predictions revealed that the methods performed almost equally well (MCC = 0.83 for logistic regression analysis, MCC = 0.85 for both decision tree and random forest; see Table 6).

### Evaluation of performance and comparison to other integration approaches

The performance of our three best integration models was also evaluated on the external validation data comprising 631 putatively consequential variants from HGMD and 478 putatively inconsequential variants from UCSC for which all individual nine tools returned a prediction. These external validation data did not overlap with the cross-validation data on which the final models were trained. For comparison, the commonly used integration approaches Condel, CoVEC, CAROL, CADD, MetaSVM and MetaLR as well as the best individual prediction tool, MutationTaster2, were also investigated. Being included only for comparison and because of its simplicity, binary summation not surprisingly performed worst of all integration methods considered in our study (Table 7). Random forest, decision tree and logistic regression showed excellent performance, with sensitivity, specificity and accuracy values of ≈0.95 and an MCC between 0.89 and 0.93. Again, random forest performed best (sensitivity = 0.97, specificity = 0.96, accuracy = 0.97, MCC = 0.93). When considering only the five top tools of the smart selection, no difference in performance was observed.

Of the established integration approaches, Condel outperformed CoVEC, CAROL and CADD (sensitivity = 0.85, specificity = 0.95, accuracy = 0.89, MCC = 0.79). Interestingly, individual prediction tool MutationTaster2 performed better than the other six established integration approaches (sensitivity = 0.92, specificity = 0.96, accuracy = 0.94, MCC = 0.88). Inspection of the ROC curves mainly confirmed these results. The best ROC curves and largest AUCs were obtained for the two integration approaches developed in this study, random forest and logistic regression, and for MetaLR (Figure 3). No ROC curve could be derived for binary summation and decision tree because these integration methods do not provide a continuous score.

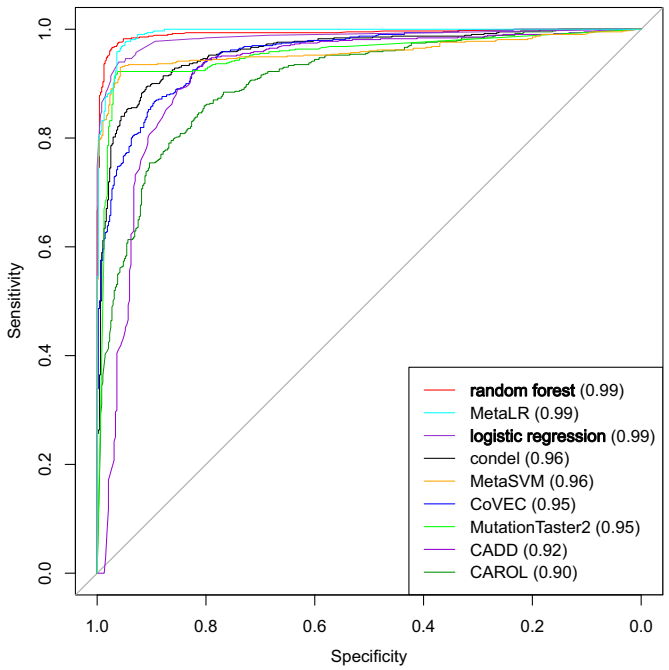


**Figure 3.** ROC curves of integration approaches random forest, logistic regression, Condel, CoVEC, CAROL, CADD, MetaSVM, MetaLR and individual prediction tool MutationTaster2 (based upon external validation data, excluding failure variants). Integration approaches of this study (random forest and logistic regression (all nine tools)) are in bold. Numbers in brackets are the AUCs of the respective approaches.

## DISCUSSION

The present study aimed at integrating the output of existing functionality prediction tools into a single numerical score. In contrast to previous integration endeavors (12,13,16,42), we systematically scrutinized all popular statistical techniques fit to such a task, from simple summation via decision tree analysis to random forest, most of which allowed the integration of both binary and continuous tool output. Since even the normalized continuous output of the tools was still not normally distributed, but had maxima at 0 and 1, integration by discriminant analysis was not deemed sensible. Also distinct from other approaches, we not only included software that was accessible by batch queries but also developed an in-house Perl script to integrate interactive tool SNPs&GO. The output of interactive tool MutPred had to be calculated locally by staff of the School of Informatics and Computing, Indiana University, owing to the large number of queries necessary. Such efforts nevertheless

**Table 6.** Performance of best integration methods on imputed output (complete data)

| Performance measure | Integration method | | | |
|---|---|---|---|---|
| | Summation binary | Logistic regression | Decision tree | Random forest |
| Sensitivity | 0.87 | 0.89 | 0.87 | 0.90 |
| Specificity | 0.89 | 0.94 | 0.97 | 0.95 |
| MCC | 0.75 | 0.83 | 0.85 | 0.85 |
| Accuracy | 0.88 | 0.92 | 0.93 | 0.93 |

Integration methods were trained on a combination of 14 233 variants with complete predictions and 3700 failure variants with imputed predictions from the complete data. Performance measures refer to the remaining 3697 failure variants with imputed predictions. MCC: Matthews correlation coefficient.

**Table 7.** External validation of integration methods (failure variants excluded)

| Integration method | Tool selection | Performance measure | | | | |
|---|---|---|---|---|---|---|
| | | Sensitivity | Specificity | MCC | Accuracy | AUC |
| Summation binary | All | 0.91 | 0.91 | 0.82 | 0.91 | n.a. |
| | Smart | 0.93 | 0.94 | 0.86 | 0.93 | n.a. |
| Logistic regression | All | 0.94 | 0.95 | 0.89 | 0.94 | 0.99 |
| | Smart | 0.94 | 0.95 | 0.89 | 0.94 | 0.99 |
| Decision tree | All | 0.96 | 0.95 | 0.91 | 0.96 | n.a. |
| | Smart | 0.96 | 0.95 | 0.91 | 0.96 | n.a. |
| Random forest | All | 0.97 | 0.96 | 0.93 | 0.97 | 0.99 |
| | Smart | 0.97 | 0.96 | 0.93 | 0.97 | 0.99 |
| Condel | | 0.85 | 0.95 | 0.79 | 0.89 | 0.96 |
| CoVEC | | 0.83 | 0.91 | 0.74 | 0.86 | 0.95 |
| CAROL | | 0.80 | 0.86 | 0.65 | 0.83 | 0.90 |
| CADD | | 0.93 | 0.81 | 0.75 | 0.88 | 0.92 |
| MetaSVM | | 0.73 | 1.00 | 0.73 | 0.84 | 0.96 |
| MetaLR | | 0.75 | 1.00 | 0.75 | 0.86 | 0.99 |
| MutationTaster2 | | 0.92 | 0.96 | 0.88 | 0.94 | 0.95 |

Integration methods considered in the present study (summation binary, logistic regression, decision tree) were trained on the complete cross-validation data. MCC: Matthews correlation coefficient; AUC: area under ROC curve; n.a.: not applicable; 'smart' tool selection: PolyPhen-2, SNPs&GO, MutPred, MutationTaster2, FATHMM.

appeared well warranted because MutPred and SNPs&GO, in particular, have been found to be of high predictive capability, both by ourselves and others (5).

We confirmed that the integration of individual prediction tools is usually a worthwhile undertaking because the best integration methods identified here outperformed all individual tools. With hindsight, such a result is not surprising because we also noted that the output of individual tools was only weakly correlated between tools. Moreover, the nine tools considered here employed a wide range of statistical methods and different types of biological information so that a combination of such complementary characteristics should almost inevitably enhance the predictive performance of the individual tools.

Decision tree, logistic regression analysis and especially random forest were found to perform best of all integration methods tested. Whilst all three allow classification of variants as consequential or inconsequential, logistic regression analysis additionally yields a posterior probability for each class, which might be of extra value when it comes to prioritizing variants for experimental follow-up.

Simple summation and majority vote were found to perform rather poorly, which was unfortunate because both methods are easy to interpret, computationally fast and require no training of a statistical model. However, simple summation and majority vote have the obvious drawback of not weighting individual tools according to their predictive performance. Since both sensitivity and specificity were found in our study to vary markedly between tools, such a lack of discrimination may indeed be one reason for the poor performance observed. Both methods therefore seem appropriate only in cases where other integration approaches are either inappropriate or impracticable, for example, when new tools are to be included that have not yet been related to other tools in the context of composite modeling.

Compared to existing integration approaches, Condel, CoVEC, CAROL and CADD, our prediction models achieved not only slightly higher specificity (random forest: 0.96, decision tree: 0.95, logistic regression: 0.95 versus Condel: 0.95, CoVEC: 0.91, CAROL: 0.86, CADD:0.81) but also considerably higher sensitivity (random forest: 0.97, decision tree: 0.96, logistic regression:0.94 versus Condel: 0.85, CoVEC: 0.83, CAROL:0.80 and CADD: 0.93) based upon the external validation data. This superiority was also reflected by higher MCC values and a generally higher accuracy of our approaches (e.g. random forest: MCC = 0.93, accuracy = 0.97) compared to the other four above mentioned existing integration approaches (where the best performance was observed for Condel: MCC = 0.79 and accuracy = 0.89). Integration approaches MetaSVM and MetaLR both achieved a specificity of almost unity in combination with a rather low sensitivity of 0.73 and 0.75, respectively. Especially for MetaLR, the sensitivity could be considerably improved without losing too much specificity. Currently, MetaLR applies a standard

threshold of 0.5 to the output of a logistic regression model which, in view of its AUC value of 0.99, does not appear to be an optimal choice. Instead, use of a different threshold should render the approach much more useful in practice.

Five of the seven *sensu stricto* prediction tools, namely PolyPhen-2, SNPs&GO, MutPred, MutationTaster2 and FATHMM, were found to perform particularly well in our study. Use of these tools alone led to almost the same predictive performance as the integration of all tools so that it may be sensible for reasons of efficiency to confine any practical implementation of the composite prediction models accordingly. Our study also revealed that failure variants, i.e. variants that lack a prediction for at least one tool, do not need to be excluded from integrated prediction even although neither random forest nor decision tree nor logistic regression analysis, by definition, can handle missing data. Missing predictions can be imputed, and their inclusion led to an acceptable overall performance with all integration methods studied.

MutationTaster2, which was trained on HGMD data (variants added up to February 2012), showed a particularly convincing performance and was clearly superior to all other individual tools. Since HGMD data were also used in the present study, although from a later time of inclusion, adaptation of MutationTaster2 to some intrinsic characteristics of these variants may have contributed to its exceptional performance. However, this cannot be the sole explanation since other tools, including MutPred and FATHMM, were trained on HGMD data as well. Moreover, there is a considerable overlap between HGMD and other publicly accessible databases like ClinVar, UniProt or dbSNP. Besides predictions, MutationTaster2 provides 'look up' variants in existing databases e.g. the 1000 Genomes Project (for inconsequential prediction) and in ClinVar (for consequential prediction) to generate automatic predictions. This is especially important for inconsequential variants. In our data, we found only 12 consequential variants which were predicted to be 'automatic disease causing'. All of these would have been predicted as consequential by the corresponding scores as well. However, out of our 10 801 inconsequential variants, 9552 were predicted to be 'automatic polymorphism'. If only the scores were regarded, 1583 of these variants would have been falsely predicted as consequential. Thus, without the look-up feature, the sensitivity of MutationTaster2 would not have changed but the specificity would have decreased from 0.96 to 0.81 highlighting the utility of the look-up feature. The inclusion of MutationTaster2 is also a probable reason for the stronger composite predictive capability achieved in our study as compared to established integration approaches that do not include this rather new tool. It is also fair to say that the use of MutationTaster2 alone may be a simple second-best alternative to tool integration in cases where the latter appears impracticable.

Grimm *et al*. (4) recently gave a detailed account of two potential sources of error in the development of variant effect prediction tools, namely type1 and type 2 circularity. Both flaws affect individual prediction tools and may accumulate in integration approaches. Type 1 circularity refers to an actual overlap of training and test data that could result in an overly optimistic judgement of the predictive performance of a given tool. To avoid this type of circularity, we included only HGMD data that were added to the database after 1 January 2013 because none of the tools considered here used any of these data for development. Moreover, we not only cross-validated our approaches internally but also compared their performance to that of other integration approaches in independent external validation data. Type 2 circularity refers to the intrinsic tendency of some prediction tools to predict the effect of a given variant mainly by the effect of variants on the same gene. Such bias can arise if variants in one and the same gene are more likely to be logged under the same functional label in mutation databases used for tool development. To counteract the consequence of type 2 circularity, we followed a variant selection strategy such that all variants of a gene were allocated to the same dataset thereby ensuring an almost even distribution of the number of variants per gene across the 10 cross-validation and the external validation datasets.

The Grantham Score performed somewhat worse than other tools in our study, which is not surprising bearing in mind that the Grantham Score was developed more than 40 years ago (a trailblazer at that time!) and was not originally intended for the classification of pathogenic variants but rather to explore protein evolution (23). Nevertheless, since the Grantham Score quantifies the biochemical difference between two amino acids in terms of their composition, polarity and molecular volume, the score can also reasonably be assumed to reflect the functional impact of missense mutations. However, all other tools considered in our study use additional information for prediction, such as secondary structure or evolutionary conservation, with the latter being particularly predictive (43). It seems likely that the lack of such information is the main reason for the poor predictive performance of the Grantham Score. Moreover, we used a threshold for the binary classification of Grantham Scores that was recommended decades ago (29). In the light of our own results, a revision of this threshold seems warranted and could well lead to better prediction in future studies.

Some of the tools considered in our study, particularly SNPs&GO, failed to yield predictions for a considerable proportion of variants. This shortcoming is largely explicable by the requirement of SNPs&GO to provide the tool with an accession number from UNIPROT, which may not be available in all cases. Moreover, the transcript that SNPs&GO associates with a given UNIPROT-ID may not be unequivocally defined which could have contributed to the large proportion of failure variants in our study because we consistently considered the longest transcript for UCSC variants. By contrast, for the consequential variants, HGMD provided transcripts that were chosen by criteria more likely to match those used by the developers of SNPs&GO. Not surprisingly, the number of failure variants was considerably smaller in the HGMD-derived than in the UCSC-derived data. However, since composite prediction worked equally well with imputed and non-imputed SNPs&GO output, we surmise that the large proportion of failure variants noted for this tool did not invalidate the general conclusions drawn from our study.

## CONCLUSION

We systematically studied different statistical techniques suitable to combine the output of nine popular prediction tools to identify putatively consequential missense variants. For our best integration approaches, the corresponding composite score clearly outperformed each single prediction tool in terms of both sensitivity and specificity. Although the best results were obtained with integrative methods random forest, decision tree and logistic regression, single tool prediction with MutationTaster2 was also found to work exceptionally well. The three top integration approaches allow prioritization of variants even in large numbers, such as can be expected to arise from whole exome sequencing. We also showed that, in cases where predictions by one or more tools were missing, imputation is an appropriate means to obtain a composite score for these variants as well. Finally, as is the case for all types of *in silico* prediction, even variants classified as consequential by the most sophisticated integration approaches are not necessarily of strong impact on a certain phenotype and additional experimental investigations will usually be required for their individual validation (2,44).

## SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

## ACKNOWLEDGEMENTS

We thank Predrag Radivojac, Biao Li and Vikas Rao Pejaver, Indiana University, Bloomington, USA, for assistance in calculating the MutPred scores. The authors gratefully acknowledge the continuous support of Almut Nebel and Friederike Flachsbart from Kiel University, Germany.

## FUNDING

Deutsche Forschungsgemeinschaft [NE1191/1-1]; Bundesministerium für Bildung und Forschung [sysINFLAME; 01ZX1306A]; Qiagen Inc through a License Agreement with Cardiff University (to D.N.C., M.M.). Funding for open access charge: sysINFLAME [01ZX1306A].

*Conflict of interest statement.* D.N.C. and M.M. acknowledge financial support from Qiagen Inc through a License Agreement with Cardiff University.

## REFERENCES

1. Coassin,S., Brandstatter,A. and Kronenberg,F. (2010) Lost in the space of bioinformatic tools: a constantly updated survival guide for genetic epidemiology. The GenEpi Toolbox. *Atherosclerosis*, **209**, 321–335.
2. Knecht,C. and Krawczak,M. (2014) Molecular genetic epidemiology of human diseases: from patterns to predictions. *Hum. Genet.*, **133**, 425–430.
3. Care,M.A., Needham,C.J., Bulpitt,A.J. and Westhead,D.R. (2007) Deleterious SNP prediction: be mindful of your training data! *Bioinformatics*, **23**, 664–672.
4. Grimm,D.G., Azencott,C.A., Aicheler,F., Gieraths,U., MacArthur,D.G., Samocha,K.E., Cooper,D.N., Stenson,P.D., Daly,M.J., Smoller,J.W. *et al.* (2015) The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity. *Hum. Mutat.*, **36**, 513–523.
5. Thusberg,J., Olatubosun,A. and Vihinen,M. (2011) Performance of mutation pathogenicity prediction methods on missense variants. *Hum. Mutat.*, **32**, 358–368.
6. Wu,J.X. and Jiang,R. (2013) Prediction of deleterious nonsynonymous single-nucleotide polymorphism for human diseases. *Sci. World J.*, **2013**, 675851.
7. Liu,X., Jian,X. and Boerwinkle,E. (2011) dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. *Hum. Mutat.*, **32**, 894–899.
8. Liu,X., Wu,C., Li,C. and Boerwinkle,E. (2016) dbNSFP v3.0: a one-stop database of functional predictions and annotations for human nonsynonymous and splice-site SNVs. *Hum. Mutat.*, **37**, 235–241.
9. Wang,K., Li,M.Y. and Hakonarson,H. (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.*, **38**, e164.
10. Adzhubei,I.A., Schmidt,S., Peshkin,L., Ramensky,V.E., Gerasimova,A., Bork,P., Kondrashov,A.S. and Sunyaev,S.R. (2010) A method and server for predicting damaging missense mutations. *Nat. Methods*, **7**, 248–249.
11. Ng,P.C. and Henikoff,S. (2001) Predicting deleterious amino acid substitutions. *Genome Res.*, **11**, 863–874.
12. Gonzalez-Perez,A. and Lopez-Bigas,N. (2011) Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. *Am. J. Hum. Genet.*, **88**, 440–449.
13. Frousios,K., Iliopoulos,C.S., Schlitt,T. and Simpson,M.A. (2013) Predicting the functional consequences of non-synonymous DNA sequence variants–evaluation of bioinformatics tools and development of a consensus strategy. *Genomics*, **102**, 223–228.
14. Lopes,M.C., Joyce,C., Ritchie,G.R., John,S.L., Cunningham,F., Asimit,J. and Zeggini,E. (2012) A combined functional annotation score for non-synonymous variants. *Hum. Hered.*, **73**, 47–51.
15. Dong,C., Wei,P., Jian,X., Gibbs,R., Boerwinkle,E., Wang,K. and Liu,X. (2015) Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Hum. Mol. Genet.*, **24**, 2125–2137.
16. Kircher,M., Witten,D.M., Jain,P., O'Roak,B.J., Cooper,G.M. and Shendure,J. (2014) A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.*, **46**, 310–315.
17. Calabrese,R., Capriotti,E., Fariselli,P., Martelli,P.L. and Casadio,R. (2009) Functional annotations improve the predictive score of human disease-related mutations in proteins. *Hum. Mutat.*, **30**, 1237–1244.
18. Li,B., Krishnan,V.G., Mort,M.E., Xin,F., Kamati,K.K., Cooper,D.N., Mooney,S.D. and Radivojac,P. (2009) Automated inference of molecular mechanisms of disease from amino acid substitutions. *Bioinformatics*, **25**, 2744–2750.
19. Schwarz,J.M., Cooper,D.N., Schuelke,M. and Seelow,D. (2014) MutationTaster2: mutation prediction for the deep-sequencing age. *Nat. Methods*, **11**, 361–36.
20. Reva,B., Antipin,Y. and Sander,C. (2011) Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res.*, **39**, e118.
21. Shihab,H.A., Gough,J., Cooper,D.N., Stenson,P.D., Barker,G.L., Edwards,K.J., Day,I.N. and Gaunt,T.R. (2013) Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Hum. Mutat.*, **34**, 57–65.
22. Pollard,K.S., Hubisz,M.J., Rosenbloom,K.R. and Siepel,A. (2010) Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.*, **20**, 110–121.
23. Grantham,R. (1974) Amino-acid difference formula to help explain protein evolution. *Science*, **185**, 862–864.
24. Stenson,P.D., Mort,M., Ball,E.V., Shaw,K., Phillips,A. and Cooper,D.N. (2014) The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Hum. Genet.*, **133**, 1–9.
25. Osler,W. (1921) *The Evolution of Modern Medicine*. Yale University Press, New Haven.
26. Imhotep. *Encyclopaedia Britannica Online*. 16 September 2016, date last accessed.
27. Kimura,M. (1983) *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge.

28. Gilissen,C., Hoischen,A., Brunner,H.G. and Veltman,J.A. (2012) Disease gene identification strategies for exome sequencing. *Eur. J. Hum. Genet.*, **20**, 490–497.
29. Li,W.H., Wu,C.I. and Luo,C.C. (1984) Nonrandomness of point mutation as reflected in nucleotide substitutions in pseudogenes and its evolutionary implications. *J. Mol. Evol.*, **21**, 58–71.
30. R Core Team (2015) *A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna.
31. The 1000 Genomes Project Consortium. (2015) A global reference for human genetic variation. *Nature*, **526**, 68–74.
32. Landrum,M.J., Lee,J.M., Benson,M., Brown,G., Chao,C., Chitipiralla,S., Gu,B., Hart,J., Hoffman,D., Hoover,J. *et al.* (2016) ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res.*, **44**, D862–D868.
33. Honaker,J., King,G. and Blackwell,M. (2011) Amelia II: a program for missing data. *J. Stat. Softw.*, **45**, 1–47.
34. van Buuren,S. and Groothuis-Oudshoorn,K. (2011) mice: Multivariate Imputation by Chained Equations in R. *J. Stat. Softw.*, **45**, 1–67.
35. Meyer,D., Zeileis,A. and Hornik,K. (2015) vcd: Visualizing Categorical Data. *R Package Version* , **1**, 4–1.
36. Matthews,B.W. (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta*, **405**, 442–451.
37. Baldi,P., Brunak,S., Chauvin,Y., Andersen,C.A.F. and Nielsen,H. (2000) Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, **16**, 412–424.
38. Wald,A. (1943) On a statistical generalization of metric spaces. *Proc. Natl. Acad. Sci. U.S.A.*, **29**, 196–197.
39. Robin,X., Turck,N., Hainard,A., Tiberti,N., Lisacek,F., Sanchez,J.C. and Muller,M. (2011) pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, **12**, 77.
40. Therneau,T., Atkinson,B. and Ripley,B. (2015) rpart: Recursive Partinioning and Regression Trees. R Package Version 4.1-10.
41. Liaw,A. and Wiener,M. (2002) Classification and regression by randomForest. *R. News*, **2**, 18–22.
42. Olatubosun,A., Valiaho,J., Harkonen,J., Thusberg,J. and Vihinen,M. (2012) PON-P: integrated predictor for pathogenicity of missense variants. *Hum. Mutat.*, **33**, 1166–1174.
43. Ng,P.C. and Henikoff,S. (2006) Predicting the effects of amino acid substitutions on protein function. *Annu. Rev. Genomics Hum. Genet.*, **7**, 61–80.
44. Cline,M.S. and Karchin,R. (2011) Using bioinformatics to predict the functional impact of SNVs. *Bioinformatics*, **27**, 441–448.