*DOI:*
10.1016/j.ygeno.2013.06.005

*Document Version*
Publisher's PDF, also known as Version of record

Genomics 102 (2013) 223–228

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

# Predicting the functional consequences of non-synonymous DNA sequence variants — evaluation of bioinformatics tools and development of a consensus strategy

Kimon Frousios [a], Costas S. Iliopoulos [a], Thomas Schlitt [b,c], Michael A. Simpson [c,*]

[a] Department of Informatics, King's College London, Strand Campus, The Strand, London WC2R 2LS, United Kingdom
[b] Institute for Mathematical and Molecular Biomedicine, King's College London, Hodgkin Building, Guy's Campus, London SE1 1UL, United Kingdom
[c] Department of Medical and Molecular Genetics, King's College London, School of Medicine, 8th Floor Tower Wing, Guy's Hospital, London SE1 9RT, United Kingdom

ARTICLE INFO

*Article history:*
Received 10 January 2013
Accepted 21 June 2013
Available online 3 July 2013

*Keywords:*
Coding
DNA
Variant
Function
Prediction
SNP

ABSTRACT

The study of DNA sequence variation has been transformed by recent advances in DNA sequencing technologies. Determination of the functional consequences of sequence variant alleles offers potential insight as to how genotype may influence phenotype. Even within protein coding regions of the genome, establishing the consequences of variation on gene and protein function is challenging and requires substantial laboratory investigation. However, a series of bioinformatics tools have been developed to predict whether non-synonymous variants are neutral or disease-causing. In this study we evaluate the performance of nine such methods (SIFT, PolyPhen2, SNPs&GO, PhD-SNP, PANTHER, Mutation Assessor, MutPred, Condel and CAROL) and developed CoVEC (Consensus Variant Effect Classification), a tool that integrates the prediction results from four of these methods. We demonstrate that the CoVEC approach outperforms most individual methods and highlights the benefit of combining results from multiple tools.



## 1. Introduction

DNA sequencing has entered a new era, in which high throughput technologies enable large amounts of sequence data to be generated quickly and at low cost. Much of the data generated is aimed at the discovery of disease causing genetic variation. A key challenge in this process is the interpretation of the functional consequences of variant alleles. Given the extensive number of variants identified by whole genome or exome re-sequencing studies, it is infeasible to interrogate the functional consequences of all variant alleles at all gene loci experimentally.

A number of bioinformatics solutions for the annotation, scoring and classification of variants have been developed to address this challenge [1–27] and several comprehensive overviews of the available tools and methods have been carried out [28–31]. Such tools are providing a supportive role in the experimental validation of disease-related alleles, by prioritizing candidate variants with predicted functional consequences as causes of specific inherited diseases and cancers. These bioinformatics approaches draw from a broad range of existing knowledge about the structure, function and conservation of genes, transcripts and proteins in which the variants are located.

In this article, our aim is to identify the best method with which to prioritize non-synonymous substitutions as candidate causes of monogenic diseases. We assess the performance of eight existing tools that draw on different resources to classify non-synonymous single nucleotide substitutions as likely or unlikely to have a serious impact on a protein's function. We also propose our own consensus strategy. The tools evaluated here are SIFT [1], PolyPhen2 [2], MutPred [3], SNPs&GO [4], PANTHER [5], PhD-SNP [6] and Mutation Assessor [7], as well as the consensus classifiers Condel [17] and CAROL [9].

SIFT, PANTHER and Mutation Assessor base their predictions on sequence conservation only, using multiple sequence alignments they each build independently, whereas PolyPhen2, SNPs&GO, MutPred and PhD-SNP combine homology information with various types of structural and functional annotation of the proteins, such as amino acid properties, the location of functional sites, and the secondary structure and membrane topology of the protein. MutPred additionally evaluates the effect of the variant to protein stability based on just the amino acid sequence, while PolyPhen2 can optionally use information from an available homologous protein 3D structure to assess the effect to protein stability. An alternative version of SNPs&GO also integrates information from 3D structures [12]. Aside from the final classification

* Corresponding author.
*E-mail addresses:* kimon.frousios@kcl.ac.uk (K. Frousios), costas.iliopoulos@kcl.ac.uk (C.S. Iliopoulos), thomas.schlitt@kcl.ac.uk (T. Schlitt), michael.simpson@kcl.ac.uk (M.A. Simpson).

0888-7543 
http://dx.doi.org/10.1016/j.ygeno.2013.06.005

of the variant, PolyPhen2, MutPred and Mutation Assessor offer additional information on what the actual biological effect of the variant may be. As far as the consensus approaches are concerned, CAROL combines the prediction scores of SIFT and PolyPhen2, while Condel additionally uses the score from Mutation Assessor.

SIFT, PANTHER, PolyPhen2, Mutation Assessor and Condel employ explicit rules or mathematical models in order to reach their verdict and are, thus, not biased by the selection of a training dataset of mutations. On the contrary, PhD-SNP and SNPs&GO use support vector machine (SVM) classifiers trained on variants extracted from Uniprot [32], whereas MutPred uses a Random Forest (RF) classifier trained on disease variants from HGMD [33] and neutral variants from Uniprot. CAROL's classification threshold has been calibrated using disease variants from HGMD and neutral variants from the 1000 Genomes Project [34].

All of these tools require minimal prior information from the user and, with the exception of PhD-SNP and SNPs&GO, are able to process batch queries directly through their available interfaces. All seven stand-alone methods accept amino acid substitutions as queries, whilst PolyPhen2 and SIFT also accept nucleotide substitution queries. Both consensus methods accept pre-obtained scores as input, although Condel's website can also mediate query submission to the three tools it combines.

Together, these tools encompass a broad range of non-synonymous DNA sequence variant classification criteria and methods. They have each been independently evaluated previously [17,31], but never together on a common dataset. In this study we propose a consensus classifier based on these tools, and demonstrate improvements in accuracy.

## 2. Results

### 2.1. Evaluation of the selected tools

We submitted our *positive* and *negative* data sets (see Materials and methods), representing a total of 15570 variants to the seven selected individual classification tools and evaluated their classification performance. We measured the prediction rate, the sensitivity and specificity, the overall correct rate, the Matthews correlation coefficient and the area under the ROC curve. As these performance metrics only apply to binary classifications, different scenarios were considered (see Methods). Table 1 presents the most favorable of the scenarios, whereas the full results are available in Additional file 1.

The prediction rate for all methods apart from PANTHER is >0.9. SNPs&GO, PhD-SNP and MutPred report predictions for nearly all of the variants submitted. When the intermediate class is considered as damaging, MutPred is the most sensitive of all of the tested classifiers (0.94) and also has the highest overall correct rate (0.92). However, SNPs&GO demonstrated the highest specificity (0.95).

PolyPhen2 was executed several times, using different parameters and producing notably different results. PolyPhen2 provides two classifiers trained on different datasets [2]: The HumVar set consists of the disease-causing mutations annotated in Uniprot as positive cases, and variants without annotated implication in disease as negative cases. The HumDiv is a smaller and stricter set, consisting only of variants implicated in monogenic (Mendelian) diseases in the Uniprot annotation as positive cases, and of observed differences between human proteins and their closely related mammalian equivalents as negative cases. Both displayed the same prediction rate, correct ratio and correlation. However, the HumVar classifier proved to be more sensitive, whereas the HumDiv one was more selective. When the corresponding amino acid sequence was explicitly provided to PolyPhen2, the prediction rate increased.

Evaluating the area under the ROC curves (Table 1) demonstrates that MutPred has the highest classification power, followed by SNPs&GO. PhD-SNP does not output any score values, therefore it was not possible to evaluate its performance using this method. The actual ROC curves are available in Additional file 2.

### 2.2. Evaluation of consensus strategies

Our independent evaluation of these seven prediction tools demonstrates how each of the different approaches has different attributes. We therefore evaluated methods of combining the outputs from these tools in order to improve the predictive performance. Previously, Gonzalez-Perez and Lopez-Bigas [17] developed Condel, which combines the output from PolyPhen2, SIFT and Mutation Assessor. Our evaluation of this approach demonstrates that in comparison to the three methods it combines, it performed better than SIFT and comparably to Mutation Assessor. It also showed a better correct rate and correlation than PolyPhen2, but the latter retained considerably higher sensitivity. Another consensus tool, CAROL [9], which combines SIFT and PolyPhen2, showed similar performance to Condel.

For the development of our own consensus approach, we evaluated combinations of different subgroups of the six individual classifiers (MutPred was excluded because of the direct overlap of our test data with the data used in the development of this tool; see Discussion) using a weighted majority vote score (*WMV*) and a support vector machine (SVM) approach (Table 2). Using the *WMV*, the highest accuracy in prediction was obtained from the combination of PolyPhen2 (HumDiv), SNPs&GO and Mutation Assessor, which produced both higher correlation and a higher correct rate than any of the individual methods involved. Its specificity was high and comparable to that of

**Table 1**
Evaluation results for the individual tools. Prediction rate, sensitivity, specificity, correct rate, Matthews correlation coefficient and area under the ROC curve, for SIFT, PolyPhen2, MutPred, SNPs&GO, PANTHER, PhD-SNP and Mutation Assessor, [a] treating low confidence predictions as damaging, and [b] treating them as neutral. PhD-SNP does not output its classification score, thus it was not possible to plot a ROC curve for this tool. PolyPhen2 was executed with various parameters: Two classifiers are offered, trained on different datasets (HumDiv and HumVar [2]). Additionally, we supplied the amino acid substitution (aa) instead of the nucleotide change, and also explicitly supplied the amino acid sequence (seq).

| Tool | Pred. Rate | Sensitivity | | Specificity | | Correct Rate | | MCC | | AUC |
|---|---|---|---|---|---|---|---|---|---|---|
| | | [a] | [b] | [a] | [b] | [a] | [b] | [a] | [b] | |
| SIFT | 0.93 | 0.73 | 0.64 | 0.86 | 0.88 | 0.79 | 0.75 | 0.59 | 0.55 | 0.87 |
| PolyPhen2 | | | | | | | | | | |
| (HumDiv) | 0.92 | 0.84 | 0.71 | 0.77 | 0.87 | 0.80 | 0.74 | 0.61 | 0.61 | 0.88 |
| (HumVar) | 0.92 | 0.76 | 0.61 | 0.86 | 0.94 | 0.81 | 0.73 | 0.62 | 0.59 | 0.90 |
| (Humdiv, aa) | 0.91 | 0.84 | 0.71 | 0.76 | 0.87 | 0.80 | 0.74 | 0.60 | 0.62 | 0.88 |
| (HumDiv, aa + seq) | 0.96 | 0.84 | 0.71 | 0.76 | 0.87 | 0.80 | 0.74 | 0.60 | 0.62 | 0.88 |
| MutPred | 1.00 | 0.94 | 0.56 | 0.90 | 0.95 | 0.92 | 0.73 | 0.84 | 0.56 | 0.97 |
| SNPs&GO | 1.00 | 0.71 | 0.71 | 0.95 | 0.95 | 0.83 | 0.83 | 0.68 | 0.68 | 0.93 |
| PANTHER | 0.68 | 0.69 | 0.69 | 0.84 | 0.84 | 0.75 | 0.75 | 0.52 | 0.52 | 0.84 |
| PhD-SNP | 1.00 | 0.62 | 0.62 | 0.78 | 0.78 | 0.70 | 0.70 | 0.41 | 0.41 | – |
| Mut. Assessor | 0.90 | 0.76 | 0.34 | 0.85 | 0.98 | 0.81 | 0.57 | 0.62 | 0.04 | 0.89 |

**Table 2**
Evaluation results for the consensus methods. Prediction rate, sensitivity, specificity, correct rate and Matthews correlation coefficient for different consensus strategies (abbreviations: PPH2 = PolyPhen2, PNTH = PANTHER, S&G = SNPs&GO, M/A = Mutation Assessor, PhD = PhD-SNP).

| Tool | Pred. Rate | Sens. | Spec. | Correct Rate | MCC |
|---|---|---|---|---|---|
| Condel | 1.00 | 0.77 | 0.88 | 0.83 | 0.66 |
| CAROL | 0.99 | 0.79 | 0.85 | 0.82 | 0.64 |
| *WMV* | | | | | |
| (SIFT,PPH2)[a] | 0.89 | 0.79 | 0.88 | 0.84 | 0.68 |
| (SIFT,PPH2,M/A)[b] | 0.96 | 0.79 | 0.88 | 0.84 | 0.67 |
| (PPH2,S&G,PNTH,PhD + M/A) | 0.97 | 0.75 | 0.93 | 0.84 | 0.69 |
| (PPH2,S&G,PhD,M/A) | 0.97 | 0.74 | 0.94 | 0.84 | 0.70 |
| (PPH2,S&G,M/A) | 0.96 | 0.80 | 0.92 | 0.86 | 0.73 |
| (SIFT,PPH2,S&G,PNTH,PhD,M/A) | 0.98 | 0.74 | 0.93 | 0.84 | 0.69 |
| (SIFT,PPH2,S&G,M/A) | 0.97 | 0.76 | 0.92 | 0.84 | 0.69 |
| CoVEC | | | | | |
| (Lin. kernel SVM) | 1.00 | 0.83 | 0.90 | 0.87 | 0.74 |
| (RBF kernel SVM) | 1.00 | 0.84 | 0.89 | 0.87 | 0.74 |

[a] Similar to CAROL.
[b] Similar to Condel.

the most specific method involved (SNPs&GO), and its sensitivity was higher than two of the constituent methods, SNPs&GO and Mutation Assessor, but not as high as PolyPhen2. Interestingly, the application of the *WMV* to the combination of SIFT, PolyPhen2 and Mutation Assessor, which is the combination employed by Condel, was not the highest performing combination. Its performance was, however, almost identical to Condel's. Similar observations are made for the combination of SIFT and PolyPhen2, which are the tools employed by CAROL.

The SVM classification approach using scores generated by SIFT, PolyPhen2, SNPs&GO and Mutation Assessor provided elevated accuracy and correlation in comparison to our most accurate *WMV* combination and also in comparison to Condel and CAROL. The use of the linear and RBF kernels produced almost identical predictions. The specificity of the SVM classifiers was marginally lower than the best performing *WMV* classifier but sensitivity was improved. In comparison to the individual constituent tools the SVM approach matched the highest observed sensitivity of the individual tools (PolyPhen2) and provided equivalent or improved correct rate and correlation compared to the respective highest values obtained by individual tools. It also provides the second highest observed specificity out of all the tools discussed in this article.

### 2.3. Additional validation

As experimentally validated variants are relatively few, there is a high risk of datasets overlapping the training sets of methods and causing biases, especially with regards to disease-causing variants. Therefore, we collected an additional separate set of 4985 disease-associated variants (see Materials and methods) and submitted them for individual and consensus classification. The results are shown in Table 3.

**Table 3**
Classification performance for a random set of 4985 variants annotated as disease-associated in PhenCode.

| Tool | Sensitivity | False negatives % | Not predicted % |
|---|---|---|---|
| SIFT | 0.74 | 0.22 | 0.05 |
| PolyPhen2 | 0.88 | 0.12 | 0.00 |
| MutPred | 0.95 | 0.05 | 0.00 |
| SNPs&GO | 0.85 | 0.15 | 0.00 |
| PANTHER | 0.60 | 0.23 | 0.17 |
| PhD-SNP | 0.70 | 0.30 | 0.00 |
| Mut. Assessor | 0.64 | 0.19 | 0.17 |
| Condel | 0.81 | 0.19 | 0.00 |
| CAROL | 0.85 | 0.15 | 0.00 |
| *WMV* | 0.81 | 0.15 | 0.04 |
| CoVEC (linear SVM) | 0.91 | 0.09 | 0.00 |

The highest sensitivity is demonstrated by MutPred, followed by our own CoVEC. By comparison to the sensitivities measured on the HGMD-derived dataset (Tables 1 and 2), most of the individual tools displayed a similar respective sensitivity. However, SNPs&GO and PhD-SNP increased noticeably in sensitivity, as did all the consensus tools.

## 3. Discussion

The results summarized in Table 1 show that, in comparison to all the other tools, PANTHER produced a significantly lower prediction rate (0.68 compared to >0.9). This may result from some substitutions not falling in positions covered by the multiple sequence alignments in its library [5]. Indeed, the authors of SNPs&GO, through which we obtained our PANTHER predictions, reported a similarly low prediction rate (0.76) when benchmarking PANTHER [4].

Also notable is the observation that, PolyPhen2's prediction rate improved when the amino acid sequence was explicitly provided. This likely highlights shortcomings in the automated retrieval of annotation from diverse resources. Specifying the sequence takes this task away from the tool and reduces the chance of mismatch between the variant and the protein. SNPs&GO and Mutation Assessor also accommodate the explicit specification of the amino acid sequence to be used, whereas the other tools retrieve the sequence automatically, based on the supplied identifier code or genomic coordinates.

Also of interest, is the comparison between Condel, CAROL and our *WMV* combination of the same tools employed by Condel and CAROL respectively (Table 2). Condel utilizes a sophisticated weighting system, based on the probability that a prediction by a method is a false positive or false negative, pre-calculated on a specified reference data set (HumVar by default) [17]. This weighting was designed to favor confident classifications by penalizing scores that are near the threshold and up-weighting scores that are near either end of each tool's scoring range. CAROL uses a different but also sophisticated weighting system in combining the scores, with the same goal of upweighting scores that are far from the threshold. In contrast, the weighting in our *WMV* method relies on each tool's self-evaluation of the classification. Despite being comparatively very simplistic, the resulting performance of the *WMV* is identical to that of Condel and very similar to CAROL.

It must be noted, that an important consideration in benchmarks such as the one conducted in this study, is the choice of the dataset.

The set must be adequately large, in order to have statistical power, and must comprise of distinct and unambiguous cases. The proportional composition of positive versus negative cases also influences the reliability of the statistical metrics used to evaluate the performance [35]. Accordingly, our evaluation was undertaken with equal numbers of disease-causing mutations from HGMD (*positive* set) and common variants from the 1000 Genomes Project Pilot project (*negative* set). The composition of these datasets gives us confidence that the *positive* set consists of non-synonymous variants with strong likelihood of functional effect. The choice of common variants for the *negative* set is based on the hypothesis that the presence of an allele at high frequency in each of genetically discrete populations may reflect a scenario in which the variant is less likely to have a substantial functional effect.

The nature and availability of the data, however, imposed a different compromise on the benchmarking set. The pressure to collect a large enough set has likely resulted in an overlap between our test set and the training sets of some of the evaluated methods. Indeed, the enhanced performance of MutPred (Table 1) could reflect the use of HGMD as a training set in the development of this tool [3] and thus a complete overlap with our *positive* set. Use of the PhenCode-derived set of positive variants, however, which was selected to exclude overlap with HGMD, showed that MutPred maintained its high sensitivity. Interestingly, SNPs&GO and PhD-SNP which used Uniprot data in their training showed a noticeable increase in sensitivity with the PhenCode dataset, likely due to the fact that a large section of PhenCode consists of the Uniprot data. Overlap can be found in the negative set as well, as CAROL exploited a similar approach to this study in drawing neutral variants out of the 1000 Genomes Project data.

Previously, evaluations of functional variant prediction tools have been presented [17,31], but they differ in their selection of methods and testing data. Though some differences in the measured accuracy values are expected, one would also expect that some trends should be apparent across studies. Indeed there is complete agreement between our results and those of Thusberg et al. [31] for SNPs&GO and PhD-SNP. The correct rate and MCC for PANTHER were also in agreement, though our sensitivity and specificity differed significantly. Our metric values for SIFT, PolyPhen2 and MutPred were higher than those reported by Thusberg et al., but were similar to those reported by Gonzalez-Perez and Lopez-Bigas [17]. These observed differences likely result from the source and composition of the test data. As discussed in the above, overlap of test data with datasets used in the individual development of these tools is clearly a concern, and the different proportion of positive and negative sets may also play a role in the differences observed between evaluation studies. We attempted to control for these biases by ensuring the positive and negative datasets were balanced and compliant with all of the tools' requirements, resulting in comparable numbers of predicted variants in both datasets across all tools. Random sampling factors may also have had an effect on the performance, though our jackknifing validation and cross-validation results (data not shown) suggest that this is unlikely to be a major source of variability.

In conclusion, we evaluated the performance of seven independently published methods that aim to predict the functional consequences of alleles that result in amino acid substitutions, as well as consensus approaches based on these methods. We conclude that, out of the individual tools that we evaluated, MutPred offers the best all around performance. PolyPhen2 provides the second best sensitivity, after MutPred, for studies in which not missing potential functional effects outweighs the cost of false positives. Both tools offer a range of information on what the actual biochemical, structural or functional effect of the substitution may be. SNPs&GO, on the other hand, is the most applicable in situations where false positives need to be minimized. We demonstrate that combining multiple prediction tools provides a more even balance between sensitivity and specificity than most of the individual methods, and that our SVM-based consensus classifier CoVEC is robust and well-suited to the task of combining scores and can match or outperform existing consensus solutions.

A website implementing both CoVEC and the *WMV* classifier is available at http://www.dcs.kcl.ac.uk/pg/frousiok/variants/index.html. Alternatively, Perl scripts and modules aimed at assisting with the preparation of data, the local execution of batch queries and the integration in local pipelines are also available from SourceForge: http://sourceforge.net/projects/covec/files.

## 4. Materials and methods

### 4.1. Tool selection and execution

Seven individual tools and two consensus tools for predicting the functional consequences of non-synonymous DNA sequence variation were selected for comparison. All selected tools enable predictions of both previously observed and novel variants and enable evaluation of large numbers of variants through batch queries or scripted submission to their web-APIs.

The task of obtaining functional effect predictions from multiple tools can be simplified with the use of meta-tools such as PON-P [24] and the Ensembl SNP Effect Predictor [22], both of which serve as gateways to a multitude of bioinformatics resources relevant to the functional study of variants, including several of the selected tools for this study (SIFT, PolyPhen2, SNPs&GO, PhD-SNP). However, we opted to generate predictions for each tool individually, using their dedicated interfaces. SIFT[1] and PolyPhen2[2] were run as batch queries on the respective web-servers. PhD-SNP, SNPs&GO and Mutation Assessor were queried on-line by scripted submission of individual variants. MutPred was kindly executed locally by its authors. We obtained PANTHER predictions indirectly via SNPs&GO. Condel and CAROL were used through respective local installations.

### 4.2. Benchmarking data sets

In order to evaluate the prediction tools, we selected two sources of human non-synonymous variant data; one that is enriched for variants with confirmed functional consequences and a second variant dataset likely to contain a reduced level of functional variation.

The set of DNA variants with functional consequences comprise variants previously implicated in the pathogenesis of inherited human disease and were extracted from HGMD Pro v.2011.1 [33]. The set of putative neutral variation was selected from variants identified by the 1000 Genomes Project Pilot project [34] (released July 2010). The pilot data is based on low coverage whole genome sequencing of 179 individuals, distributed in three groups with distinct geographic origin (African, Caucasian, East-Asian). From the >15 million distinct SNPs contained in the pilot data we selected non-synonymous variants with a minor allele frequency greater or equal to 5% in each of the three populations. This selection of variants should be enriched for variants that do not have a functional consequence.

All variants were originally derived from the NCBI36/hg18 human genome assembly and annotated with respect to the genes in which they reside using Annovar [25]. Prior to being submitted for prediction, they were converted to the NCBI37/hg19 assembly using the UCSC Genome Browsers's liftOver utility[3] [36] and cross-referenced with the NCBI transcript RefSeq [37] and the Uniprot [32] databases in order to obtain the amino acid sequence and protein identifiers. After all processing, our putatively neutral data set from the 1000 Genomes Project comprised 7791 non-synonymous variants across

[1] http://sift.jcvi.org/www/SIFT_chr_coords_submit.html
[2] http://genetics.bwh.harvard.edu/pph2/bgi.shtml
[3] http://genome.ucsc.edu/cgi-bin/hgLiftOver

4555 genes — we define this set of variants as our *negative set*. In order to size-match our neutral and functional cohorts, we randomly selected a subset of the functionally enriched variants from the HGMD catalog of disease-causing mutations with a maximum of 10 disease-causing variants per gene, resulting in 7779 variants across 1448 genes — we define this set of variants as our *positive set*.

As HGMD has been used to train some of the tools examined here, an independent additional validation dataset was compiled. Single amino acid substitution variants annotated in PhenCode [38] as disease-associated were collected and compared to the HGMD Pro catalog. After removing any potential overlap, a sample subset of the remaining variants from PhenCode was randomly extracted, consisting of 4985 amino acid substitutions in 1164 proteins.

### 4.3. Evaluation

The selected classification tools were evaluated with metrics applicable for binary classification problems [35]. We define the following metrics: sensitivity — the proportion of the *positive set* classified as having a functional consequence, specificity — the proportion of the *negative set* not predicted to have a functional consequence, correct rate — the proportion of correctly classified cases from both sets together. We also calculated the Matthews correlation coefficient (MCC) and area under the ROC curve (AUC).

These binary classification metrics are limited in certain situations within this study. Whilst SNPs&GO, PANTHER, PhD-SNP and Condel offer a simple binary prediction, SIFT and PolyPhen2 generate three ranked classes, Mutation Assessor generates five and MutPred none. In order for the metrics to be applicable, the ranked classes were merged down to simulate a binary classification. Two scenarios were considered: (i) The intermediate class was considered to be neutral and (ii) it was considered to be damaging. In the case of Mutation Assessor, which employs five classes, the two lowest probability classes were binned together as neutral, the two highest probability classes were merged as damaging, and the middle class was considered under the two scenarios mentioned above. MutPred does not explicitly classify the variants. Therefore we created three ranked classes as follows: We considered all the variants which scored below the threshold advised in the tool's documentation as neutral. Variants scoring above the threshold, for which MutPred additionally offered hypotheses about the nature of the mutation's effect, comprised our damaging class. Variants scoring above the threshold, but lacking any hypothesis about the mutation's effect were treated as the intermediate class, to be considered under the two scenarios.

Direct comparison between these tools was further confounded, as it is not uncommon for the tools to be unable to classify certain variants. To address the issue of the number of submitted variants not equaling the number of predictions, we recorded the prediction rate (proportion of submitted cases for which a prediction was returned) for each tool and calculated two versions of the accuracy metrics, one based on the number of classified cases and one based on the number of submitted cases (see Additional file 1).

To test the robustness of the performance metrics, we performed 10-fold jackknifing, removing each time a different 10% of the variants, evenly from positive and negative cases and re-calculating the correct rate each time. The values obtained from the 10 iterations where consistent for each method (data not shown), with the highest observed standard deviation <0.005.

### 4.4. Consensus classification development

In order to evaluate the potential benefit of incorporating outputs from multiple tools to improve predictions we implemented two consensus approaches: a weighted majority vote score (*WMV*) and a support vector machine (SVM).

The *WMV* approach assigned a numerical value ($V_i$) to each of the three defined classes (damaging, intermediate and neutral) from each of the selected tools. We assigned the value +2 for the damaging class, +1 for the intermediate class, and −2 for the neutral class. A score of 0 was assigned when a method did not generate a prediction. The weighted vote score was calculated by adding up the individual values generated for each of the tools incorporated in the consensus model:

$$WMV = \sum_i V_i \,. \tag{1}$$

We define 0 as the threshold value for the WVS. Negative values lead to classification of the variant as neutral and positive values as damaging. If the votes add up to exactly 0 classification is not possible.

The SVM-based consensus classifier incorporates the raw output scores generated by SIFT, PolyPhen2, SNPs&GO and Mutation Assessor. We used SVMlight [39] to build the SVM model and perform the classification using either a linear kernel or a radial based function (RBF) kernel with a $g$ parameter of 0.0625 and the default $C$ parameter. A grid search for optimal $C$ and $g$ parameters for the RBF kernel demonstrated a broad range of similarly well performing value combinations. To test the robustness of the predictor, we performed 10-fold cross-validation, by splitting our data into pairs of subsets and using one subset of each pair to train the model and the other to evaluate the accuracy. The training was performed in quadruplicate, changing the number of variants used to train the model (2000, 5000, 10000, 13000). The results were very consistent across all 40 iterations (data not shown) with the standard deviation <0.01, and accuracy improving less than 0.005 between the model trained on 2000 variants and the model trained on 13000 variants.

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.ygeno.2013.06.005.

## Acknowledgments

KF is funded by the Greek State Scholarships Foundation. The authors also acknowledge support from the Department of Health via the National Institute for Health Research (NIHR) comprehensive Biomedical Research Centre award to Guy's & St Thomas' NHS Foundation Trust in partnership with King's College London and King's College Hospital NHS Foundation Trust. The authors also thank Dr Biao Li for kindly running MutPred on our behalf.

## References

[1] P. Ng, S. Henikoff, SIFT: predicting amino acid changes that affect protein function, Nucleic Acids Res. 31 (2003) 3812.
[2] I. Adzhubei, S. Schmidt, L. Peshkin, et al., A method and server for predicting damaging missense mutations, Nat. Methods 7 (2010) 248.
[3] B. Li, V. Krishnan, M. Mort, et al., Automated inference of molecular mechanisms of disease from amino acid substitutions, Bioinformatics 25 (2009) 2744.
[4] R. Calabrese, E. Capriotti, P. Fariselli, et al., Functional annotations improve the predictive score of human disease-related mutations in proteins, Hum. Mutat. 30 (2009) 1237.
[5] P. Thomas, A. Kejariwal, N. Guo, et al., Applications for protein sequence-function evolution data: mRNA/protein expression analysis and coding SNP scoring tools, Nucleic Acids Res. 34 (2006) W645.
[6] E. Capriotti, R. Calabrese, R. Casadio, Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information, Bioinformatics 22 (2006) 2729.
[7] B. Reva, Y. Antipin, C. Sander, Predicting the functional impact of protein mutations: application to cancer genomics, Nucleic Acids Res. 39 (2011) e118.
[8] L. Bao, M. Zhou, Y. Cui, nsSNPAnalyzer: identifying disease-associated nonsynonymous single nucleotide polymorphisms, Nucleic Acids Res. 33 (2005) W480.
[9] M. Lopes, C. Joyce, G. Ritchie, et al., A combined functional annotation score for non-synonymous variants, Hum. Hered. 73 (2012) 47.
[10] M. Barenboim, M. Masso, I. Vaisman, D. Jamison, Statistical geometry based prediction of nonsynonymous SNP functional effects using random forest and neuro-fuzzy classifiers, Proteins Struct. Funct. Genet. 71 (2008) 1930.
[11] Y. Bromberg, B. Rost, SNAP: predict effect of non-synonymous polymorphisms and function, Nucleic Acids Res. 35 (2007) 3823.
[12] E. Capriotti, R. Altman, Improving the prediction of disease-related variants using protein three-dimensional structure, BMC Bioinforma. 12 (2011) S3.

[13] C. Chelala, A. Khan, N. Lemoine, SNPnexus: a web database for functional annotation of newly discovered and public domain single nucleotide polymorphisms, Bioinformatics 25 (2008) 655.
[14] L. Conde, J. Vaquerizas, H. Dopazo, et al., PupaSuite: finding functional single nucleotide polymorphisms for large-scale genotyping purposes, Nucleic Acids Res. 34 (2006) W621.
[15] J. Dantzer, C. Moad, R. Heiland, S. Mooney, MutDB services: interactive structural analysis of mutation data, Nucleic Acids Res. 33 (2005) W311.
[16] C. Ferrer-Costa, J. Gelpi, L. Zamakola, et al., PMUT: a web-based tool for the annotation of pathological mutations of proteins, Bioinformatics 21 (2005) 3176.
[17] A. Gonzalez-Perez, N. Lopez-Bigas, Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel, Am. J. Hum. Genet. 88 (2011) 440.
[18] B. Hemminger, B. Saelim, F. Sullivan, TAMAL: an integrated approach to choosing SNPs for genetic studies of human complex traits, Bioinformatics 22 (2006) 626.
[19] H. Kang, K. Choi, B. Kim, et al., FESD: a functional element SNPs database in human, Nucleic Acids Res. 33 (2005) D518.
[20] R. Karchin, M. Diekhans, L. Kelly, et al., LS-SNP: large-scale annotation of coding non-synonymous SNPs based on multiple information sources, Bioinformatics 21 (2005) 2814.
[21] P. Lee, H. Shatkay, F-SNP: computationally predicted functional SNPs for disease association studies, Nucleic Acids Res. 36 (2008) D820.
[22] W. McLaren, B. Pritchard, D. Rios, et al., Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor, Bioinformatics 26 (2010) 2069.
[23] J. Reumers, S. Maurer-Stroh, J. Schymkowitz, F. Rousseau, SNPeffect v2.0: a new step in investigating the molecular phenotypic effects of human non-synonymous SNPs, Bioinformatics 22 (2006) 2183.
[24] J. Thusberg, M. Vihinen, Pathogenic or not? And if so, then how? Studying the effects of missense mutations using bioinformatics methods, Hum. Mutat. 30 (2009) 703.
[25] K. Wang, M. Li, H. Hakonarson, ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data, Nucleic Acids Res. 38 (2010) e164.
[26] H. Yuan, J. Chiou, W. Tseng, et al., FASTSNP: an always up-to-date and extendable service for SNP function analysis and prioritization, Nucleic Acids Res. 34 (2006) W635.
[27] P. Yue, E. Melamud, J. Moult, SNPs3D: candidate gene and SNP selection for association studies, BMC Bioinforma. 7 (2006) 166.
[28] R. Karchin, Next generation tools for the annotation of human SNPs, Brief. Bioinform. 10 (2008) 35.
[29] S. Coassin, A. Brandstätter, F. Kronenberg, Lost in the space of bioinformatics tools: a constantly updated survival guide for genetic epidemiology. The GenEpi Toolbox, Atherosclerosis 209 (2010) 321.
[30] M. Cline, R. Karchin, Using bioinformatics to predict the functional impact of SNVs, Bioinformatics 27 (2011) 441.
[31] J. Thusberg, A. Olatubosun, M. Vihinen, Performance of mutation pathogenicity prediction methods on missense variants, Hum. Mutat. 32 (2011) 1.
[32] The UniProt Consortium, Ongoing and future developments at the Universal Protein Resource, Nucleic Acids Res. 39 (2011) D214.
[33] P. Stenson, M. Mort, E. Ball, et al., The human gene mutation database: 2008 update, Genome Med. 1 (2009) 13.
[34] The 1000 Genomes Project Consortium, A map of human genome variation from population-scale sequencing, Nature 467 (2010) 1061.
[35] P. Baldi, S. Brunak, Y. Chauvin, et al., Assessing the accuracy of prediction algorithms for classification: an overview, Bioinformatics 16 (2000) 412.
[36] P. Fujita, B. Rhead, A. Zweig, et al., The UCSC Genome Browser database: update 2011, Nucleic Acids Res. 39 (2011) D876.
[37] K. Pruitt, T. Tatusova, D. Maglott, NCBI reference sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins, Nucleic Acids Res. 33 (2005) D501.
[38] B. Giardine, C. Riemer, T. Hefferon, et al., PhenCode: connecting ENCODE data with mutations and phenotype, Hum. Mutat. 28 (2007) 554.
[39] T. Joachims, Making large-scale SVM learning practical, in: B. Schoelkopf, C. Burges, A. Smola (Eds.), Advances in Kernel Methods — Support Vector Learning, MIT-Press, 1999.