Informatics Human Mutation

DOI 10.1002/humu.22225

Supporting Information for this preprint is available from the *Human Mutation* editorial office upon request (humu@wiley.com)

Predicting the Functional, Molecular and Phenotypic Consequences of Amino Acid Substitutions using Hidden Markov Models

Hashem A. Shihab[1†], Julian Gough[2†], David N. Cooper[3], Peter D. Stenson[3], Gary L. A. Barker[4], Keith J. Edwards[4], Ian N. M. Day[1‡], Tom R. Gaunt[1‡*]

[1] Bristol Centre for Systems Biomedicine and MRC CAiTE Centre, School of Social and Community Medicine, University of Bristol, Bristol, BS8 2BN, UK

[2] Department of Computer Science, University of Bristol, The Merchant Venturers Building, Bristol, BS8 1UB, UK

[3] Institute of Medical Genetics, School of Medicine, Cardiff University, Cardiff, CF14 4XN, UK

[4] School of Biological Sciences, University of Bristol, Woodland Road, Bristol, BS8 1UG, UK

[†] Joint First Authorship

[‡]Joint Last Authorship

* Correspondence

Tom Gaunt, Bristol Centre for Systems Biomedicine and MRC CAiTE Centre, School of Social and Community Medicine, University of Bristol, Bristol, BS8 2BN, UK. E-mail: Tom.Gaunt@Bristol.ac.uk

Grant Sponsor

This work was supported by the UK Medical Research Council (MRC) and was carried out in the Bristol Centre for Systems Biomedicine (BCSBmed) doctoral training centre (director INMD) using the computational facilities of the Advanced Computing Research Centre, University of Bristol - http://www.bris.ac.uk/acrc. TRG & INMD acknowledge financial support from the MRC (G1000427). JG's contribution was supported by BBSRC grant BB/G022771. DNC and PDS gratefully acknowledge the financial support of BIOBASE GmbH.

Abstract

The rate at which nonsynonymous single nucleotide polymorphisms (nsSNPs) are being identified in the human genome is increasing dramatically owing to advances in whole-genome/whole-exome sequencing technologies. Automated methods capable of accurately and reliably distinguishing between pathogenic and functionally neutral nsSNPs are therefore assuming ever-increasing importance. Here, we describe the Functional Analysis Through Hidden Markov Models (FATHMM) software and server: a species-independent method with optional species-specific weightings for the prediction of the functional effects of protein missense variants. Using a model weighted for human mutations, we obtained performance accuracies that outperformed traditional prediction methods (i.e. SIFT, PolyPhen and PANTHER) on two separate benchmarks. Furthermore, in one benchmark, we achieve performance accuracies that outperform current state-of-the-art prediction methods (i.e. SNPs&GO and MutPred). We demonstrate that FATHMM can be efficiently applied to high-throughput/large-scale human and non-human genome sequencing projects with the added benefit of phenotypic outcome associations. To illustrate this, we evaluated nsSNPs in wheat (*Triticum* spp.) in order to identify some of the important genetic variants responsible for the phenotypic differences introduced by intense selection during domestication. A web-based implementation of FATHMM, including a high-throughput batch facility and a downloadable standalone package, is available at http://fathmm.biocompute.org.uk.

Key Words: SNP, hidden Markov models, bioinformatics, FATHMM

Introduction

Nonsynonymous single nucleotide polymorphisms (nsSNPs) lead to amino acid substitutions (AASs) and have the potential to affect the function of the protein product of a gene via the structure, biochemistry and/or splicing of the protein. Advances in high-throughput sequencing technologies have accelerated the rate at which nsSNPs are now being identified [The 1000 Genomes Project, 2010]. Accurate automated computational methods capable of predicting the effects of AASs and amenable to high-throughput analyses of large datasets are therefore of increasing importance for identifying and prioritising functional nsSNPs for further studies [Thusberg and Vihinen, 2009].

The majority of computational prediction methods utilize evolutionary sequence conservation and/or structural annotations within homologous (orthologous and/or paralogous) proteins from a database of known sequences and/or structures [Ng and Henikoff, 2006]. Traditionally, the BLAST range of pairwise alignment [Altschul et al., 1990] and sequence profile algorithms [Altschul et al., 1997] have been used to search large sequence databases for homologous proteins falling within a pre-defined similarity threshold. However, weaknesses of these algorithms include the position-invariant scoring matrices in BLAST and the *ad hoc* estimation of algorithm parameters, i.e. position-invariant gap penalties, in PSI-BLAST [Bateman and Haft, 2002]. On the other hand, hidden Markov models (HMMs) [Krogh et al., 1994; Eddy, 1996] are powerful probabilistic models that can be used to capture position-specific information within a multiple sequence alignment (MSA) of homologous sequences. Here, a MSA is represented as a series of match, insert and delete states linked together via state-transitions. A match state

models the position-specific amino acid probabilities (with Dirichlet mixtures [Sjölander et al., 1996]) at each column within the sequence alignment whereas insert/delete states allow for particular residues/states to be inserted and skipped, respectively, throughout the sequence alignment (position-specific insertions/deletions). HMM profiles are similar to PSI-BLAST profiles except they are applied within a more rigorous statistical framework and have been shown to perform considerably better when detecting distant relationships between homologous sequences [Madera and Gough, 2002].

Inspired by previous work [Ng and Henikoff, 2001; Thomas et al., 2003; Calabrese et al., 2009], we have capitalized upon recent advances in the HMMER3 software suite [Eddy, 2009] to potentiate the computational prediction of the functional effects of AASs using HMMs. First, we present an unweighted/species-independent method in which homologous sequences are automatically collected and aligned using an iterative search procedure. The resulting MSA is then used to build an *ab initio* HMM where sequence conservation is then interrogated through the internal match states of the model. In conjunction, sequence conservation within manually curated HMMs representing the alignment of conserved protein domain families: SUPERFAMILY [Gough et al., 2001] and Pfam [Sonnhammer et al., 1997], is interrogated. This additional domain-based analysis is capable of capturing important structural and evolutionary constraints (via priors) that are potentially missed when using an automatically collected alignment of homologous sequences. Next, we introduce a weighted/species-specific method which incorporates "pathogenicity weights". These weights are derived from the relative frequencies of disease-associated and functionally neutral AASs mapping onto conserved protein domains. Using a model weighted for human mutations, we obtained performance accuracies

that outperformed traditional prediction methods: SIFT, PolyPhen and PANTHER; on two separate benchmarks. Furthermore, in one benchmark, we achieve performance accuracies that outperform current state-of-the-art prediction methods: SNPs&GO and MutPred. We demonstrate that our method, Functional Analysis Through Hidden Markov Models (FATHMM), can be efficiently applied to all foreseeable high-throughput large-scale genomic datasets, and advances the field with the added benefit of providing phenotypic outcome associations. In addition to demonstrating the predictive capabilities of FATHMM on multiple benchmarks representing human mutations, we have applied it in practice to a large dataset of nsSNPs in wheat (*Triticum* spp.) in order to identify some of the key genetic variants responsible for the phenotypic differences introduced by intense selection during domestication and have made this analysis publicly available to the scientific community.

Materials & Methods

The Mutation Datasets

A collection of five human mutation datasets from online databases and the literature were downloaded and used in this study (Table 1). First, inherited disease-causing AASs annotated as DMs (damaging mutations) in the Human Gene Mutation Database [Stenson et al., 2009] (HGMD - November 2011; http://www.hgmd.org) and inherited putative functionally neutral AASs in the UniProt database [Apweiler et al., 2004] (UniProt - November 2011; http://www.uniprot.org/docs/humsavar) were downloaded and used to calculate the pathogenicity weights implemented in our weighted/species-specific method. Next, we obtained two human

mutation datasets to assess the performance of FATHMM against the performance of other computational prediction algorithms previously reported in the literature: the VariBench database (VariBench - November 2011; http://bioinf.uta.fi/VariBench) used in a comprehensive review [Thusberg et al., 2011] of nine other computational prediction methods [Ng and Henikoff, 2001; Ramensky et al., 2002; Thomas et al., 2003; Bao et al., 2005; Capriotti et al., 2006; Bromberg and Rost, 2007; Calabrese et al., 2009; Li et al., 2009; Adzhubei et al., 2010; Mort et al., 2010] and 267 AASs in four cancer-associated genes (*BRCA1*, *MSH2*, *MLH1* and *TP53*) used in a recent review [Hicks et al., 2011] of four alternative computational prediction algorithms [Ng and Henikoff, 2001; Tavtigian et al., 2006; Adzhubei et al., 2010; Reva et al., 2011]. Finally, we downloaded a human mutation dataset consisting of disease-associated and putative functionally neutral AASs from the SwissVar portal [Mottaz et al., 2010] (SwissVar - February 2011; http://swissvar.expasy.org) and performed an independent benchmark of FATHMM against eight other computational prediction algorithms [Ng and Henikoff, 2001; Ramensky et al., 2002; Thomas et al., 2003; Ferrer-Costa et al., 2004; Capriotti et al., 2006; Calabrese et al., 2009; Li et al., 2009; Adzhubei et al., 2010; Mort et al., 2010].

Scoring the Magnitude of Effect of Amino Acid Substitutions

The procedure for predicting the functional consequences on the protein function is as follows (see Supp. Figure S1 for a flow-diagram detailing the procedure): the *JackHMMER* component of HMMER3 (one iteration with the optional --*hand* parameter applied; see Supp. Figure S2) is used to search for homologous sequences within the UniRef90 [Suzek et al., 2007] database (November 2011). As part of this procedure, an *ab initio* HMM representing the MSA of

homologous sequences (with Dirichlet mixtures [Sjölander et al., 1996]) is constructed and used. In conjunction, protein domain annotations from the SUPERFAMILY [Gough et al., 2001] (version 1.75) and Pfam [Sonnhammer et al., 1997] (Pfam-A and Pfam-B; version 26.0) databases are made. The relevant SUPERFAMILY and Pfam HMMs are then extracted only if and when the domain assignment is deemed significant (e-value ≤ 0.01) and the AAS maps onto a match state within the model.

The information gain (as measured by the Kullback-Leibler [Kullback and Leibler, 1951] divergence from the SwissProt/TrEMBL [Apweiler et al., 2004] amino acid composition) is then calculated at the corresponding match states within the HMMs extracted above. Next, we interrogate the underlying amino acid probabilities modelled by the most informative HMM and assume that a reduction in the amino acid probabilities (when comparing the wild-type to the mutant residue) indicates a potentially negative impact upon protein function whereas a gain in the amino acid probabilities indicates a more favourable substitution. Furthermore, we assume that larger reductions in amino acid probabilities have more substantial effects than smaller reductions in amino acid probabilities. Here, the predicted magnitude of the effect upon protein function is calculated as follows:

$$unweighted = \ln \frac{P_m/(1.0 - P_m)}{P_w/(1.0 - P_w)} \quad (1)$$

where P_w and P_m represent the underlying probabilities for the wild-type and mutant amino acid residues, respectively.

Incorporating Species-Specific Pathogenicity Weights

As before, we interrogate the amino acid probabilities within the most informative SUPERFAMILY [Gough et al., 2001] or Pfam [Sonnhammer et al., 1997] (Pfam-A and Pfam-B) HMM (as measured by the Kullback-Leibler [Kullback and Leibler, 1951] divergence from the SwissProt/TrEMBL [Apweiler et al., 2004] amino acid composition). However, for an improved performance in human, the predicted magnitude of effect is weighted by the relative frequency of disease-associated (HGMD) and functionally neutral (UniProt) AASs mapping onto the relevant SUPERFAMILY/Pfam HMM:

$$weighted = \ln \frac{(1.0 - P_w) \cdot (W_n + 1.0)}{(1.0 - P_m) \cdot (W_d + 1.0)} \quad (2)$$

where P_w and P_m represent the underlying probabilities for the wild-type and mutant amino acid residues, respectively, and the pathogenicity weights, W_d and W_n, represent the relative frequencies of disease-associated and functionally neutral AASs mapping onto the relevant HMM, respectively. The pathogenicity weights also include a pseudo-count of 1.0 to avoid a zero divisible term.

Annotating the Molecular and Phenotypic Consequences of Amino Acid Substitutions

The overall biological function of a protein is commonly governed by the various combinations of protein domains within it [Peterson et al., 2010]. Therefore, we annotate the potential molecular and phenotypic consequences of pathogenic mutations via domain-centric ontologies

[de Lima Morais et al., 2011]. For example, the molecular consequences of AASs are statistically inferred by mapping SUPERFAMILY [Gough et al., 2001] HMMs onto the Gene Ontology [Ashburner et al., 2000]. Moreover, the phenotypic consequences of AASs are annotated by extending these mappings onto several phenotype ontologies including the Human Phenotype Ontology [Robinson et al., 2008], the Mammalian Phenotype Ontology [Smith and Eppig, 2009] and the Plant Phenotype Ontology [Pujar et al., 2006; Ilic et al., 2007].

Performance Evaluation

In accordance with previous computational prediction methods, the following six parameters (formulae 3-8) were used to assess the performance of our models:

$$Accuracy = \frac{tp + tn}{tp + tn + fp + fn} \quad (3)$$

$$Precision = \frac{tp}{tp + fp} \quad (4)$$

$$Sensitivity = \frac{tp}{tp + fn} \quad (5)$$

$$Specificity = \frac{tn}{fp + tn} \quad (6)$$

$$Negative\,Predictive\,Value\,(NVP) = \frac{tn}{tn + fn} \quad (7)$$

$$Matthew's\,Correlation\,Coefficient\,(MCC) = \frac{(tp \cdot tn) - (fn \cdot fp)}{\sqrt{(tp + fn) \cdot (tp + fp) \cdot (tn + fn) \cdot (tn + fp)}} \quad (8)$$

where *tp* and *fp* refer to the number of true positives and false positives reported and *tn* and *fn* denote the number of true negatives and false negatives reported.

Results

Calculating a Prediction Threshold

Theoretically, using our prediction formulae, scores approximately equal to zero indicate that there is no significant change in the underlying amino acid probabilities whereas scores less than zero indicate that an unfavourable substitution has been observed, i.e. the mutant residue is less likely to be observed than the wild-type residue, and scores greater than zero indicate that a favourable substitution has been observed, i.e. the mutant residue is more likely to be observed than the wild-type residue. However, in practice, FATHMM is sensitive to small fluctuations in the amino acid probabilities modelled within the HMMs. For example, the slightest reduction in amino acid probabilities would yield a pathogenic prediction in our unweighted/species-independent algorithm. Therefore, to eliminate the effects of these fluctuations, we plotted the distribution of the predicted magnitude of effect for both disease-associated and functionally neutral AASs within the SwissVar dataset (Figure 1). From this, we calculated prediction

thresholds for our unweighted and weighted methods at which the specificity and sensitivity were both maximised (-3.0 and -1.5, respectively). Using our unweighted method, we noted that the majority of disease-associated AASs (>60%) fell below our threshold whereas the majority of functionally neutral polymorphisms (80%) fell above this threshold. Furthermore, using our weighted method, the majority of disease-associated AASs (80%) fell below our threshold whereas a significant proportion of functionally neutral polymorphisms (>80%) fell above this threshold.

A Performance Comparison Against Published Reviews

The performance of FATHMM was compared against the performance of other computational prediction algorithms reported in two previously published reviews [Hicks et al., 2011; Thusberg et al., 2011]. First, the VariBench database was used to benchmark our method against nine alternative computational prediction algorithms [Ng and Henikoff, 2001; Ramensky et al., 2002; Thomas et al., 2003; Bao et al., 2005; Capriotti et al., 2006; Bromberg and Rost, 2007; Calabrese et al., 2009; Li et al., 2009; Adzhubei et al., 2010; Mort et al., 2010] (Table 2). Typically, the performance of trained/weighted computational prediction algorithms is superior to that of theoretical/unweighted algorithms. Therefore, to allow for a fair comparison to be made, we opted to compare our unweighted/species-independent method against other theoretical/unweighted computational algorithms and our weighted/species-specific method against other trained/weighted computational prediction algorithms. From Table 2, and in terms of performance accuracies, PANTHER [Thomas et al., 2003] appears to be the best performing theoretical/unweighted prediction method with an accuracy of 76%. It appears that both SIFT

[Ng and Henikoff, 2001] (another sequence-based method) and our unweighted method perform less favourably with accuracies of 65% and 69%, respectively, indicating that FATHMM is somewhat the better option of the two. The observed performances in our analysis indicate that our weighted method is the best performing method available with an overall performance accuracy of 86%, thereby outperforming the current state-of-the-art prediction methods MutPred [Li et al., 2009; Mort et al., 2010] (81%) and SNPs&GO [Capriotti et al., 2006] (82%).

Next, we used the Hicks dataset to benchmark FATHMM against four other computational prediction algorithms (using their native alignments) [Ng and Henikoff, 2001; Tavtigian et al., 2006; Adzhubei et al., 2010; Reva et al., 2011] (Table 3). Overall, Align-GVGD [Tavtigian et al., 2006] appears to be the best performing method. However, Align-GVGD employs gene-specific alignments and its performance is severely affected when automatically generated alignments are used [Hicks et al., 2011]. These results appear to indicate that our unweighted method is more specific than either Align-GVGD or SIFT; however, we also noted higher false positive rates when compared with the other prediction methods. In general, and perhaps more surprisingly, it appears that the performance of all trained/weighted computational prediction methods is inferior across the four genes when compared to their theoretical/unweighted counterparts. Again, although no one trained/weighted prediction method performs best across the four genes, it would appear that our weighted method is, on average, the most specific/least sensitive.

An Independent Benchmark Against Other Computational Prediction Methods

Although we recognise the importance of comparing prediction methods in relation to previously

established benchmarks, we also conducted our own benchmark (using the SwissVar mutation dataset – see Materials & Methods) comparing the performance of FATHMM against eight published computational prediction methods [Ng and Henikoff, 2001; Ramensky et al., 2002; Thomas et al., 2003; Ferrer-Costa et al., 2004; Capriotti et al., 2006; Calabrese et al., 2009; Li et al., 2009; Adzhubei et al., 2010; Mort et al., 2010] (Table 3 – see Supp. Table S1). In contrast to the VariBench benchmark, and in terms of performance accuracies, it appears that both SIFT [Ng and Henikoff, 2001] and our own unweighted method outperform PANTHER [Thomas et al., 2003] (68%) with performance accuracies of 74% and 71%, respectively, indicating that SIFT is somewhat the better option. The best performing method is MutPred [Li et al., 2009; Mort et al., 2010] with a performance accuracy of 90%. However, the observed performances show that our weighted method once again performs favourably when compared to other state-of-the-art prediction methods: SNPs&GO [Calabrese et al., 2009], despite the domain-based restriction inherited from our pathogenicity weights. Next, we compared the performance of our unweighted method via Receiver Operating Characteristic (ROC) curves against the top ranking theoretical/unweighted computational prediction methods: SIFT and PANTHER (Figure 2; A & B – see Supp. Figure S3 for a comprehensive ROC curve against all evaluated methods). Impressively, given a 10% false positive rate, it seems that the performance of our unweighted method is comparable to SIFT thereby highlighting the sensitivity of our method to small fluctuations within the underlying amino acid probabilities. Furthermore, we compared the performance of our weighted method via ROC curves against the top-ranking trained/weighted prediction algorithms: MutPred and SNPs&GO (Figure 2; C & D). These results confirm that our weighted method performs favourably when compared to SNPs&GO.

The pathogenicity weights incorporated in FATHMM were not directly used to train for, or recognise, pathogenic sequences and/or mutations. We do nevertheless recognise the potential for bias in the performances observed. Therefore, in order to remove this bias, we performed a 'leave-one-out' analysis on all benchmarking datasets. Here, we adjusted our pathogenicity weights, W_d and W_n, if and only when the AAS being evaluated was present in either the HGMD [Stenson et al., 2009] or UniProt [Apweiler et al., 2004] datasets. We observed no significant deviations in the performance measures reported above and hence concluded that the performances observed were not biased towards the pathogenicity weights employed (see Supp. Table S2).

To understand the potential complementarity/redundancy of FATHMM to other methods, we assessed the intersection of disease-associated AASs correctly identified (true positives) by our method and the top-ranking computational prediction algorithms (Figure 3). From this analysis, it was clear that no one method completely encapsulates all other methods i.e. each method succeeded in correctly and uniquely identifying some disease-associated AASs where other methods did not. These results reaffirm previous suggestions that combining predictions from multiple prediction methods has the potential to perform better than any individual method [Liu et al., 2011; Olatubosun et al., 2012].

Facilitating the High-Throughput Analysis of Large Genomic Datasets

Anticipating a massive increase in the number of available whole-genome and whole-exome datasets, the need for accurate computational prediction methods capable of processing these

datasets in a timely fashion is increasingly apparent. As a result, the majority of computational prediction algorithms now offer some form of pre-computed facility allowing for near-instant predictions to be returned (see Supp. Table S3). However, only SIFT [Ng and Henikoff, 2001] and PolyPhen-2 [Adzhubei et al., 2010] allow for batch submissions (with restrictions) to be made. To facilitate the high-throughput analysis of large-scale genomic datasets, our public web-server provides up-to-date (pre-computed) domain assignments for several large sequence collections, including SwissProt/TrEMBL [Apweiler et al., 2004]; thereby enabling (unrestricted) near-instant predictions to be made for AASs falling within conserved protein domains. Furthermore, our pre-computed database is available as an optional download enabling near-instant predictions to be made while running our software locally.

Annotating Phenotypic Outcome Associations

As previously alluded to, FATHMM not only predicts the potentially deleterious nature of AASs but is also capable of annotating the molecular and phenotypic consequences of these mutations via domain-centric ontologies. To illustrate this, we evaluated the predicted phenotypic consequences of disease-associated AASs within the SwissVar dataset (Supp. Table S4). As expected, the phenotypic consequences of well-characterised diseases are correctly identified. For example, the cardiovascular consequences of the C1971Y mutation in *FBN1* (Marfan syndrome; MIM# 154700) are correctly identified via domain-based ontological associations. However, potential issues of using domain-centric ontologies arise when a common domain harbours multiple mutations with distinct and uniquely expressed phenotypes. In these instances, domain-centric ontological associations may have become diluted and should therefore be used

with caution. For example, the predicted phenotypic consequences for the R239C mutation in *CHRNG* (Escobar syndrome; MIM# 265000) are consistent with the associated syndrome, which is characterised by a decrease in fetal movement and overall muscle weakness. However, phenotypes not associated with (or secondary to) Escobar syndrome, for example abnormalities in temperature regulation, were also predicted. Nevertheless, we foresee that these annotations will be most prominent in protein sequences of unknown function and/or ongoing non-human genome sequencing projects, as demonstrated below.

Case Study: Annotating the Functional and Phenotypic Consequences of nsSNPs in Wheat

As the world's population continues to grow, so does the demand for crops with particular characteristics such as drought resistance, high yield and resistance to pests and pathogens. The cultivation and repeat harvesting of wild "landrace" wheat varieties has led over time to the emergence of domesticated "elite" wheat varieties with desirable phenotypic characteristics. In an attempt to elucidate some of the important genetic variants responsible for these characteristics, we collected single nucleotide variants (SNVs) from four elite UK bread wheat varieties (Avalon, Cadenza, Rialto and Savannah) and have predicted the functional effects of these mutations when compared to four landrace wheat varieties from the Watkins collection held at the John Innes Centre, Norwich, UK (304, 306, 311 and 328). For this analysis, SNVs were mapped onto the draft wheat genome assembly and six-frame translated. For each reading frame, SUPERFAMILY [Gough et al., 2001] and Pfam [Sonnhammer et al., 1997] domain assignments on the full length amino acid sequence were made and the corresponding AASs were evaluated using our unweighted method. We found several biologically interesting SNV

differences between the landrace and elite wheat varieties (see Supp. Table S5). For example, wheat contig F0Z7V0F01D2DA5 had a SNP at position 172 in the casein kinase II beta subunit domain with phenotypic consequences predicted to affect the flower developmental stages and vegetative growth. The casein kinase II beta subunit domain has a putative function in flowering time regulation in the model plant *Arabidopsis* [Ogiso et al., 2010] and is likely to be biologically significant as European domestic wheat will have been selected to grow under shortened seasons and different day lengths to the landraces. Next, wheat contig GIZP4PP04H5FGF had a SNP at position 219 which lies within the Pfam starch synthase catalytic domain. Once again, this is likely to be biologically significant as the quantity and properties of starch are important to the baking properties of cultivated wheat and will thus have been under strong selection. Finally, wheat contig 09781 had a SNP at position 368 in the cysteine proteinase domain with predicted phenotypic consequences affecting plant structure development. In cereals, cysteine proteases are known to be important in the laying down of storage proteins [Fahmy et al., 2004]. As with starch, the properties of wheat storage proteins will have come under intense selection during domestication as they are the most important determinant of baking qualities and economic yield. These results, made publicly available to the wheat genomics community at http://www.cerealsdb.uk.net/functional_snps/index.htm, illustrate the potential additional utility of FATHMM in predicting the functional consequences of variants identified in ongoing non-human genome sequencing projects (even in species very distantly related to human).

Discussion

Here, we have introduced and discussed the Functional Analysis Through Hidden Markov Models (FATHMM) software and server: a species-independent method with optional species-specific weightings for the prediction of the functional effects of protein missense variants. Inspired by previous sequence-based computational prediction algorithms [Ng and Henikoff, 2001; Thomas et al., 2003], our unweighted/species-independent method interrogates sequence conservation through the underlying amino acid probabilities modelled by the internal match states of several HMMs representing the alignment of homologous sequences and conserved protein domains. Following a similar weighting scheme implemented in SNPs&GO [Calabrese et al., 2009], our weighted/species-specific method amalgamates sequence conservation within the HMMs with "pathogenicity weights" representing the relative frequencies of disease-associated and functionally neutral AASs mapping onto conserved protein domains. The pathogenicity weights incorporated here are not directly used to train for, or recognise, pathogenic sequences and/or mutations. Instead, these weights are capable of recognising protein domains (species-independent/evolutionary units) sensitive to or intolerant of missense mutations. Therefore, the pathogenicity weights implemented in FATHMM are also likely to represent an improvement for non-human organisms (especially those not too distantly related to human) [Ferrer-Costa et al., 2005].

The performance of FATHMM was compared to the performances of alternative computational prediction methods previously reported in two published reviews [Hicks et al., 2011; Thusberg et al., 2011]. Furthermore, we performed our own independent benchmark comparing the performance of FATHMM against the performance of other computational prediction methods. In two benchmarks (VariBench/SwissVar), the performance of our unweighted method is

comparable to another sequence-based method: SIFT [Ng and Henikoff, 2001], and to a sequence/structure based method: PolyPhen-1 [Ramensky et al., 2002]. This performance reaffirms the ability of FATHMM to recognise important structural and/or evolutionary constraints (via priors) modelled within manually curated HMMs representing the alignment of conserved protein domains: SUPERFAMILY [Gough et al., 2001] and Pfam [Sonnhammer et al., 1997]. A detailed analysis of four cancer-associated genes (Hicks; *BRCA1*, *MSH2*, *MLH1* and *TP53*) shows Align-GVGD [Tavtigian et al., 2006] to be the best performing prediction method. However, this can be attributed to the manually curated (gene-specific) sequence alignments employed in the prediction method. On average, the performance of our unweighted method in this benchmark is comparable to SIFT.

An important issue to consider when comparing the performance of trained/weighted computational prediction methods is the cross-validation dataset, i.e. these prediction methods should ideally be tested using "blind" datasets to minimise the bias in the performances observed. Unfortunately, this level of testing is not possible as it would require retraining/validating all prediction methods with common datasets. However, the majority of disease-associated AASs in the VariBench database were collected from Locus-Specific Databases (LSDB) and are not found in commonly used training datasets, e.g. SwissProt/TrEMBL [Apweiler et al., 2004]. Therefore, the curators claim this bias is minimised in this dataset [Thusberg et al., 2011]. Here, the performance of our weighted method appears to outperform the current state-of-the-art prediction methods: MutPred [Li et al., 2009; Mort et al., 2010] and SNPs&GO [Calabrese et al., 2009]. By contrast, the mutation dataset used in our independent benchmark was collected from the SwissVar [Mottaz et al., 2010] portal. As a

result, the estimated performances of other computational prediction methods which have been trained on SwissProt/TrEMBL mutations may be over-inflated. Here, MutPred is the best performing method; however, the performance of our weighted method is comparable to SNPs&GO. In order to alleviate the potential bias in our method, we performed a leave-one-out analysis and found no significant deviations in the observed performances; we therefore concluded that the performances observed in FATHMM were not an artefact of the weighting scheme employed. The performances of all trained/weighted computational prediction algorithms were, somewhat surprisingly, inferior when compared to their theoretical/unweighted counterparts across four cancer-associated genes. The performances observed within Align-GVGD (gene-specific alignments) suggest that there may be some benefit in incorporating disease-specific weightings into our algorithm, e.g. cancer-specific weightings similar to those employed by [Capriotti and Altman, 2011].

A potential disadvantage of our weighted method is the inherited restriction (via the weighting scheme employed) to AASs falling within conserved protein domains. However, protein domain annotations from the SUPERFAMILY and Pfam databases encompass around 80% of the SwissProt/TrEMBL database [Punta et al., 2012]. In our analysis, we were able to analyse a large proportion (>70%) of the VariBench and SwissVar benchmarking datasets. On the other hand, unlike other sequence-based prediction methods (including our own unweighted method), which are too slow for practical use in large-scale sequencing projects, our weighted method uses computationally inexpensive domain assignments. Therefore, FATHMM can be efficiently applied to all foreseeable high-throughput large-scale genomic datasets with minimal reduction in coverage. In addition, our method advances the field with its unique ability to annotate the

molecular and phenotypic consequences of AASs using several domain-centric ontologies [de Lima Morais et al., 2011] including the Human Phenotype Ontology [Robinson et al., 2008] and the Mammalian Phenotype Ontology [Smith and Eppig, 2009]. Thus, by coupling the functional predictions generated by FATHMM with domain-based ontological associations, as opposed to protein level annotations, we have developed a tool which is capable of providing useful insights into the underlying mechanisms disrupted by AASs without any prior/background information on the protein itself. A web-based implementation of FATHMM, which facilitates the high-throughput analysis of large-scale genomic datasets, and includes a downloadable open-source software package, is available at http://fathmm.biocompute.org.uk.

Acknowledgments

This work was supported by the UK Medical Research Council (MRC) and was carried out in the Bristol Centre for Systems Biomedicine (BCSBmed) doctoral training centre (director INMD) using the computational facilities of the Advanced Computing Research Centre, University of Bristol - http://www.bris.ac.uk/acrc. TRG & INMD acknowledge financial support from the MRC (G1000427). JG's contribution was supported by BBSRC grant BB/G022771. DNC and PDS gratefully acknowledge the financial support of BIOBASE GmbH. The authors thank Predrag Radivojac and Biao Li from Indiana University (MutPred) and Adam Hospital from the University of Barcelona (PMut) for their help in analysing the SwissVar dataset. The authors also thank Dr. Philip Guthrie for his help in reviewing the manuscript.

References

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. 2010. A method and server for predicting damaging missense mutations. Nat Methods 7:248–249

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol 215:403–410

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402

Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale DA, O'Donovan C, Redaschi N, Yeh L-SL. 2004. UniProt: the Universal Protein knowledgebase. Nucleic Acids Res 32:D115–119

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25:25–29

Bao L, Zhou M, Cui Y. 2005. nsSNPAnalyzer: identifying disease-associated nonsynonymous single nucleotide polymorphisms. Nucleic Acids Res 33:W480–482

Bateman A, Haft DH. 2002. HMM-based databases in InterPro. Brief. Bioinformatics 3:236–245

Bromberg Y, Rost B. 2007. SNAP: predict effect of non-synonymous polymorphisms on function. Nucleic Acids Res 35:3823–3835

Calabrese R, Capriotti E, Fariselli P, Martelli PL, Casadio R. 2009. Functional annotations improve the predictive score of human disease-related mutations in proteins. Hum Mutat 30:1237–1244

Capriotti E, Altman RB. 2011. A new disease-specific machine learning approach for the prediction of cancer-causing missense variants. Genomics 98:310–317

Capriotti E, Calabrese R, Casadio R. 2006. Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. Bioinformatics 22:2729–2734

Eddy SR. 1996. Hidden Markov models. Curr Opin Struct Biol 6:361–365

Eddy SR. 2009. A new generation of homology search tools based on probabilistic inference. Genome Inform 23:205–211

Fahmy AS, Ali AA, Mohamed SA. 2004. Characterization of a cysteine protease from wheat Triticum aestivum (cv. Giza 164). Bioresour Technol 91:297–304

Ferrer-Costa C, Orozco M, Cruz X de la. 2004. Sequence-based prediction of pathological mutations. Proteins 57:811–819

Ferrer-Costa C, Orozco M, Cruz X de la. 2005. Use of bioinformatics tools for the annotation of disease-associated mutations in animal models. Proteins 61:878–887

Gough J, Karplus K, Hughey R, Chothia C. 2001. Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. J Mol Biol 313:903–919

Hicks S, Wheeler DA, Plon SE, Kimmel M. 2011. Prediction of missense mutation functionality depends on both the algorithm and sequence alignment employed. Hum Mutat 32:661–668

Ilic K, Kellogg EA, Jaiswal P, Zapata F, Stevens PF, Vincent LP, Avraham S, Reiser L, Pujar A, Sachs MM, Whitman NT, McCouch SR, Schaeffer ML, Ware DH, Stein LD, Rhee SY. 2007. The plant structure ontology, a unified vocabulary of anatomy and morphology of a flowering plant. Plant Physiol 143:587–599

Krogh A, Brown M, Mian IS, Sjölander K, Haussler D. 1994. Hidden Markov models in computational biology. Applications to protein modeling. J Mol Biol 235:1501–1531

Kullback S, Leibler RA. 1951. On Information and Sufficiency. The Annals of Mathematical Statistics 22:79–86

Li B, Krishnan VG, Mort ME, Xin F, Kamati KK, Cooper DN, Mooney SD, Radivojac P. 2009. Automated inference of molecular mechanisms of disease from amino acid substitutions. Bioinformatics 25:2744–2750

Lima Morais DA de, Fang H, Rackham OJL, Wilson D, Pethica R, Chothia C, Gough J. 2011. SUPERFAMILY 1.75 including a domain-centric gene ontology method. Nucleic Acids Res 39:D427–434

Liu X, Jian X, Boerwinkle E. 2011. dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. Hum Mutat 32:894–899

Madera M, Gough J. 2002. A comparison of profile hidden Markov model procedures for remote homology detection. Nucleic Acids Res 30:4321–4328

Mort M, Evani US, Krishnan VG, Kamati KK, Baenziger PH, Bagchi A, Peters BJ, Sathyesh R, Li B, Sun Y, Xue B, Shah NH, Kann MG, Cooper DN, Radivojac P, Mooney SD. 2010. In silico functional profiling of human disease-associated and polymorphic amino acid substitutions. Hum Mutat 31:335–346

Mottaz A, David FPA, Veuthey A-L, Yip YL. 2010. Easy retrieval of single amino-acid

polymorphisms and phenotype information using SwissVar. Bioinformatics 26:851–852

Ng PC, Henikoff S. 2001. Predicting deleterious amino acid substitutions. Genome Res 11:863–874

Ng PC, Henikoff S. 2006. Predicting the effects of amino acid substitutions on protein function. Annu Rev Genomics Hum Genet 7:61–80

Ogiso E, Takahashi Y, Sasaki T, Yano M, Izawa T. 2010. The role of casein kinase II in flowering time regulation has diversified during evolution. Plant Physiol 152:808–820

Olatubosun A, Väliaho J, Härkönen J, Thusberg J, Vihinen M. 2012. PON-P: Integrated predictor for pathogenicity of missense variants. Hum Mutat 33:1166-1174

Peterson TA, Adadey A, Santana-Cruz I, Sun Y, Winder A, Kann MG. 2010. DMDM: domain mapping of disease mutations. Bioinformatics 26:2458–2459

Pujar A, Jaiswal P, Kellogg EA, Ilic K, Vincent L, Avraham S, Stevens P, Zapata F, Reiser L, Rhee SY, Sachs MM, Schaeffer M, Stein L, Ware D, McCouch S. 2006. Whole-plant growth stage ontology for angiosperms and its application in plant biology. Plant Physiol 142:414–428

Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J, Heger A, Holm L, Sonnhammer ELL, Eddy SR, Bateman A, Finn RD. 2012. The Pfam protein families database. Nucleic Acids Res 40:D290–301

Ramensky V, Bork P, Sunyaev S. 2002. Human non-synonymous SNPs: server and survey. Nucleic Acids Res 30:3894–3900

Reva B, Antipin Y, Sander C. 2011. Predicting the functional impact of protein mutations: application to cancer genomics. Nucleic Acids Res 39:e118

Robinson PN, Köhler S, Bauer S, Seelow D, Horn D, Mundlos S. 2008. The Human Phenotype

Ontology: a tool for annotating and analyzing human hereditary disease. Am J Hum Genet 83:610–615

Sjölander K, Karplus K, Brown M, Hughey R, Krogh A, Mian IS, Haussler D. 1996. Dirichlet mixtures: a method for improved detection of weak but significant protein sequence homology. Comput Appl Biosci 12:327–345

Smith CL, Eppig JT. 2009. The mammalian phenotype ontology: enabling robust annotation and comparative analysis. Wiley Interdiscip Rev Syst Biol Med 1:390–399

Sonnhammer EL, Eddy SR, Durbin R. 1997. Pfam: a comprehensive database of protein domain families based on seed alignments. Proteins 28:405–420

Stenson PD, Mort M, Ball EV, Howells K, Phillips AD, Thomas NS, Cooper DN. 2009. The Human Gene Mutation Database: 2008 update. Genome Med 1:13

Suzek BE, Huang H, McGarvey P, Mazumder R, Wu CH. 2007. UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinformatics 23:1282–1288

Tavtigian SV, Deffenbaugh AM, Yin L, Judkins T, Scholl T, Samollow PB, Silva D de, Zharkikh A, Thomas A. 2006. Comprehensive statistical study of 452 BRCA1 missense substitutions with classification of eight recurrent substitutions as neutral. J Med Genet 43:295–305

The 1000 Genomes Project. 2010. A map of human genome variation from population-scale sequencing. Nature 467:1061–1073

Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, Daverman R, Diemer K, Muruganujan A, Narechania A. 2003. PANTHER: a library of protein families and subfamilies indexed by function. Genome Res 13:2129–2141

Thusberg J, Olatubosun A, Vihinen M. 2011. Performance of mutation pathogenicity prediction

methods on missense variants. Hum Mutat 32:358–368

Thusberg J, Vihinen M. 2009. Pathogenic or not? And if so, then how? Studying the effects of missense mutations using bioinformatics methods. Hum Mutat 30:703–714

Figure Legends

Figure 1. The distribution of the predicted magnitude of effect for disease-associated (shaded region) and functionally neutral (unshaded region) AASs in the SwissVar dataset using our unweighted and weighted methods (A & B, respectively). From this, we calculated prediction thresholds at which both specificity and sensitivity were maximised (-3.0 & -1.5, respectively).

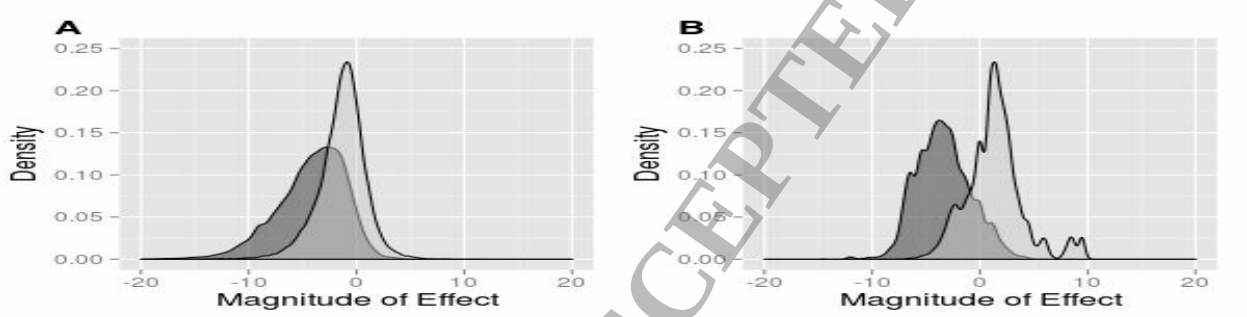

Figure 2. Receiver Operating Characteristic (ROC) curves for the top-ranking computational prediction algorithms evaluated using the SwissVar dataset. Here, we compare our unweighted method against SIFT and PANTHER (A – full curve; B - 10% false positive rate) whereas our weighted method is compared to SNPs&GO and MutPred (C – full curve; D - 10% false positive rate). Full ROC curves for all computational prediction algorithms evaluated are made available in Supp. Figure S3.

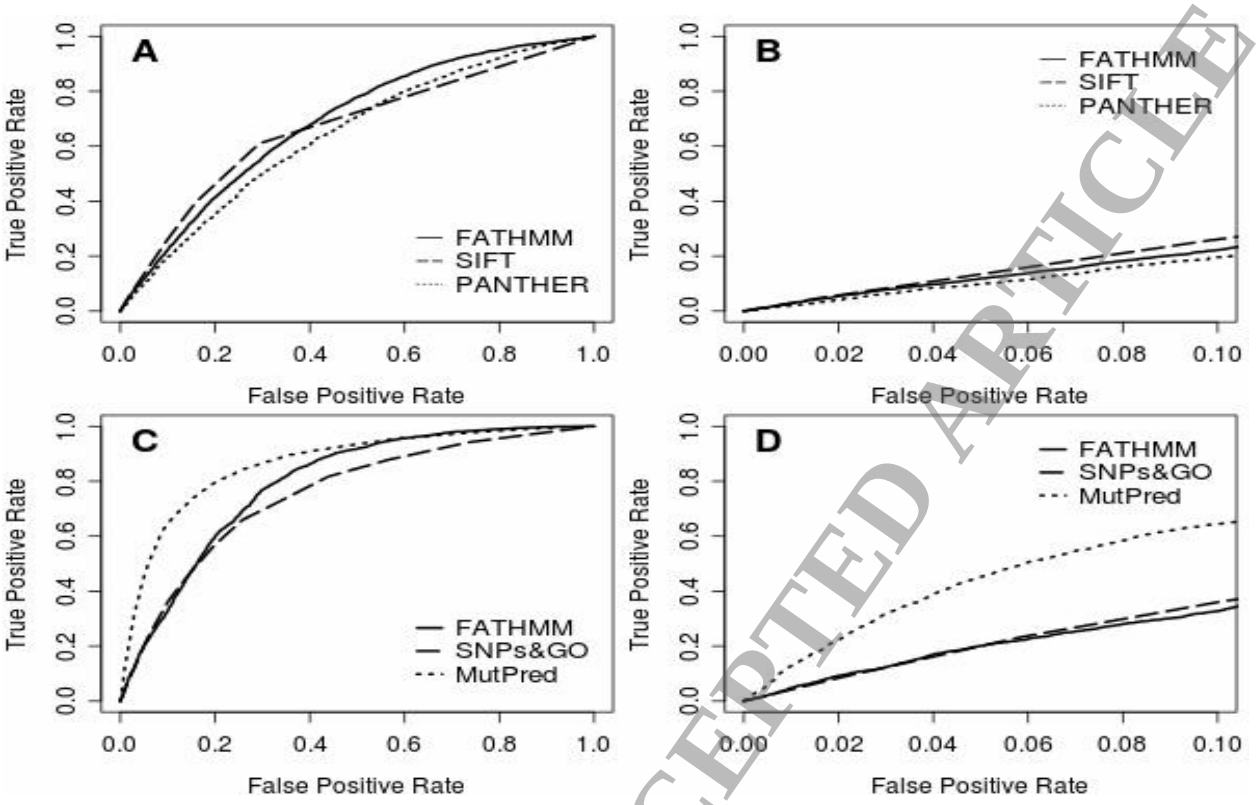

Figure 3. The intersection of disease-associated amino acid substitutions correctly identified by the top-ranking computational prediction algorithms evaluated using the SwissVar dataset. Here, we compare our unweighted method against SIFT and PANTHER (A) whereas our weighted method is compared to SNPs&GO and MutPred (B).

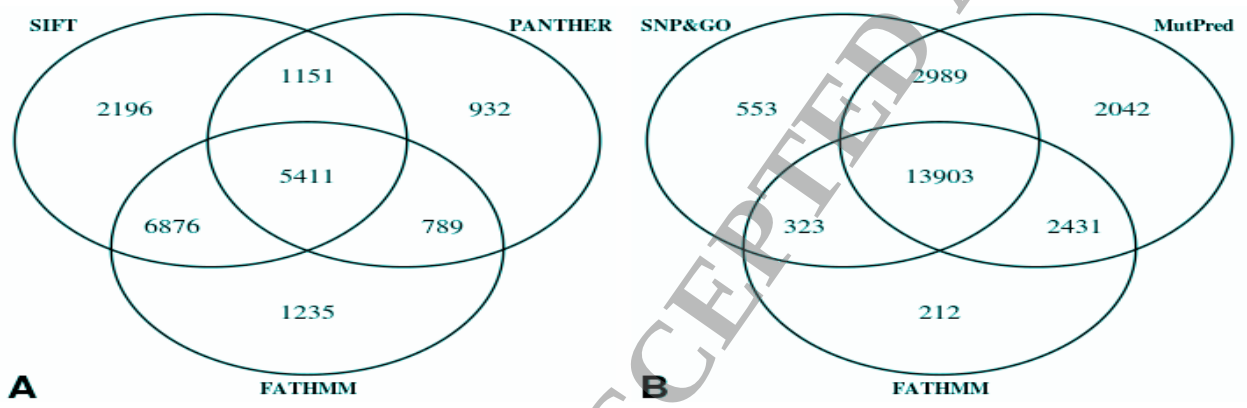

Table 1. Summary of Mutation Datasets

Dataset	Proteins	Amino Acid Substitutions	Description
HGMD	2,298	49,532	Inherited disease-causing mutations from HGMD used to calculate our pathogenicity weights
UniProt	11,548	36,928	Inherited putative functionally neutral mutations from UniProt used to calculate our pathogenicity weights
VariBench	9,684	40,470	Benchmarking dataset used in a review of nine alternative computational prediction algorithms [Thusberg et al., 2011]
Hicks et. al.	4	267	Benchmarking dataset consisting of both disease-causing and functionally neutral mutations in four well-characterised genes (*BRCA1*, *MSH2*, *MLH1*, *TP53*) used in a recent review of four alternative prediction algorithms [Hicks et al., 2011]
SwissVar	11,986	59,976	Benchmarking dataset used as an independent benchmark of eight alternative prediction algorithms

Table 2. Performance of Computational Prediction Methods using the VariBench Benchmarking Dataset

	tp	fp	tn	fn	Accuracy†	Precision†	Specificity†	Sensitivity†	NVP†	MCC†
Theoretical/Unweighted Computational Prediction Methods										
SIFT	10464	4856	12188	7433	0.65	0.64	0.62	0.68	0.66	0.30
PolyPhen 1_a	10093	9185	17669	3199	0.69	**0.77**	**0.85**	0.52	0.64	0.39
PolyPhen 1_b	14285	4993	13671	7197	0.70	0.68	0.66	0.74	0.72	0.40
PANTHER	9689	2859	8676	2797	**0.76**	0.76	0.76	**0.77**	**0.77**	**0.53**
FATHMM (unweighted)	11561	4839	16257	7707	0.69	0.72	0.77	0.60	0.66	0.38
Trained/Weighted Computational Prediction Methods										
PolyPhen 2_a	13807	5102	13863	6010	0.71	0.71	0.70	0.73	0.72	0.43
PolyPhen 2_b	16206	2703	10199	9674	0.69	0.64	0.51	0.86	0.78	0.39
PhD-SNP	11900	6896	16788	4377	0.71	0.75	0.79	0.63	0.68	0.43
SNPs&GO	13736	5487	17028	1382	0.82	**0.90**	**0.92**	0.71	0.76	0.65
nsSNPAnalyzer	4360	2778	1319	943	0.60	0.59	0.58	0.61	0.60	0.19
SNAP	16000	2146	8190	6387	0.72	0.67	0.56	**0.88**	0.83	0.47
MutPred	13829	2507	15891	4557	0.81	0.79	0.78	0.85	0.84	0.63
FATHMM (weighted)	14231	1633	10146	2336	**0.86**	0.86	0.86	0.86	**0.86**	**0.72**

tp, *fp*, *tn*, *fn* refer to the number of true positives, false positives, true negatives and false negatives, respectively.
† Accuracy, Precision, Specificity, Sensitivity, NVP and MCC are calculated from normalised numbers
a "Probably Pathogenic" predictions classed as disease-causing
b "Probably Pathogenic" predictions classed as functionally neutral

The performances of alternative computational prediction algorithms have been reproduced with permission from [Thusberg et al., 2011] - copyright (2012) Wiley.

Table 3. Specificity and Sensitivity of Computational Prediction Methods in Four Well-Characterised Genes (*BRCA1*, *MSH2*, *MLH1* and *TP53*)

Algorithm	*BRCA1*		*MSH2*		*MLH1*		*TP53*	
	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity
Theoretical/Unweighted Computational Prediction Methods								
SIFT	0.31	**0.94**	0.46	**0.90**	0.52	0.72	0.75	**0.84**
Align-GVGD	**0.94**	0.71	0.55	**0.90**	0.52	**0.97**	**1.00**	0.82
FATHMM (unweighted)	0.56	0.65	**0.73**	0.84	**0.71**	0.77	**1.00**	0.71
Trained/Weighted Computational Prediction Methods								
PolyPhen-2	0.38	0.77	0.36	0.90	**0.67**	0.90	**1.00**	0.84
X-Var	0.56	**0.82**	0.27	**1.00**	0.33	**1.00**	0.50	0.96
FATHMM (weighted)	**0.70**	0.47	**0.50**	0.79	0.24	0.97	NA[†]	**1.00**

[†] The specificity for our weighted method in this instance is uninformative as there was only one neutral mutation falling within conserved protein domains.

The performances of alternative computational prediction algorithms have been reproduced with permission from [Hicks et al., 2011] - copyright (2012) Wiley.

Table 4. Performance of Computational Prediction Methods using the SwissVar Benchmarking Dataset

	tp	fp	tn	fn	Accuracy†	Precision†	Specificity†	Sensitivity†	NVP†	MCC†
Unweighted Computational Prediction Methods										
SIFT	15634	6318	28236	7716	**0.74**	**0.79**	**0.82**	0.67	0.71	**0.49**
PolyPhen 1	12803	8759	18603	4497	0.71	0.70	0.68	**0.74**	**0.72**	0.42
PANTHER	8283	5842	17447	5162	0.68	0.71	0.75	0.62	0.66	0.37
FATHMM (unweighted)	14311	6717	29454	9429	0.71	0.76	0.81	0.60	0.67	0.43
Weighted/Trained Computational Prediction Methods										
PolyPhen 2 (HumDiv)	19782	13592	20874	3204	0.73	0.69	0.61	0.86	0.81	0.48
PolyPhen 2 (HumVar)	19928	13239	21227	3058	0.74	0.69	0.62	0.87	0.82	0.50
PhD-SNP Sequence	15695	9380	26838	8062	0.70	0.72	0.74	0.66	0.69	0.40
PhD-SNP Profile	17548	7233	27731	5161	0.78	0.79	0.79	0.77	0.78	0.57
PMut	13498	12156	23636	10159	0.62	0.63	0.66	0.57	0.61	0.23
SNPs&GO	17768	3768	29101	5655	0.82	0.87	0.89	0.76	0.79	0.65
MutPred	21365	3500	32719	2392	**0.90**	**0.90**	**0.90**	**0.90**	**0.90**	**0.80**
FATHMM (weighted)	15916	3017	19713	4496	0.82	0.85	0.87	0.78	0.80	0.65

tp, *fp*, *tn*, *fn* refer to the number of true positives, false positives, true negatives and false negatives, respectively
† Accuracy, Precision, Specificity, Sensitivity, NVP and MCC are calculated from normalised numbers