*Bioinformatics*, 32(12), 2016, 1797–1804
doi: 10.1093/bioinformatics/btw094
Advance Access Publication Date: 18 February 2016
Original Paper



Sequence analysis

# Multilevel biological characterization of exomic variants at the protein level significantly improves the identification of their deleterious effects

**Daniele Raimondi[1,2,3,4], Andrea M. Gazzo[1,2], Marianne Rooman[1,5], Tom Lenaerts[1,2,6] and Wim F. Vranken[1,3,4,*]**

[1]Interuniversity Institute of Bioinformatics in Brussels, ULB-VUB, Brussels 1050, Belgium, [2]Machine Learning Group, Université Libre De Bruxelles, Brussels 1050, Belgium, [3]Structural Biology Brussels, Vrije Universiteit Brussel, Brussels 1050, Belgium, [4]Structural Biology Research Centre, VIB, Brussels 1050, Belgium, [5]3BIO-BioInfo Group, Université Libre De Bruxelles, Brussels 1050, Belgium and [6]Artificial Intelligence Lab, Vrije Universiteit Brussel, Brussels 1050, Belgium

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on September 24, 2015; revised on February 11, 2016; accepted on February 15, 2016

## Abstract

**Motivation**: There are now many predictors capable of identifying the likely phenotypic effects of single nucleotide variants (SNVs) or short in-frame Insertions or Deletions (INDELs) on the increasing amount of genome sequence data. Most of these predictors focus on SNVs and use a combination of features related to sequence conservation, biophysical, and/or structural properties to link the observed variant to either neutral or disease phenotype. Despite notable successes, the mapping between genetic variants and their phenotypic effects is riddled with levels of complexity that are not yet fully understood and that are often not taken into account in the predictions, despite their promise of significantly improving the prediction of deleterious mutants.

**Results**: We present DEOGEN, a novel variant effect predictor that can handle both missense SNVs and in-frame INDELs. By integrating information from different biological scales and mimicking the complex mixture of effects that lead from the variant to the phenotype, we obtain significant improvements in the variant-effect prediction results. Next to the typical variant-oriented features based on the evolutionary conservation of the mutated positions, we added a collection of protein-oriented features that are based on functional aspects of the gene affected. We cross-validated DEOGEN on 36 825 polymorphisms, 20 821 deleterious SNVs, and 1038 INDELs from SwissProt. The multilevel contextualization of each (variant, protein) pair in DEOGEN provides a 10% improvement of MCC with respect to current state-of-the-art tools.

**Availability and implementation**: The software and the data presented here is publicly available at http://ibsquare.be/deogen.

**Contact**: wvranken@vub.ac.be

**Supplementary information**: Supplementary data are available at *Bioinformatics* online.

## 1 Introduction

The human exome covers ~1% of the entire genome and contains roughly 180 000 protein-coding exons (Ng *et al.*, 2009). There are currently 40 million human variants identified by genome sequencing over multiple individuals from different genetic backgrounds (Sherry *et al.*, 2001), with every individual carrying around four million variants in their whole genome. Whole exome sequencing (WES) identifies approximately 20 000 of these per individual (Abecasis *et al.*, 2012; Ng *et al.*, 2009) in a cost-effective manner (Bamshad *et al.*, 2011; Cooper *et al.*, 1995), and has led to the successful identification of many new disease–gene associations in the last years (Bamshad *et al.*, 2011; Boycott *et al.*, 2013).

The huge amount of WES data that became available over the last decade has allowed researchers to explore the genetic basis of human disease, and tools were developed that aid in the identification of disease-causing genetic variants. For example, variant prioritization tools (Moreau and Tranchevent, 2012) process exome sequencing data and rank the observed variants, so that the ones most likely to be causative for a specific disease are ranked highest. The major benefit of this ranking is that it focuses subsequent *in vitro* validation experiments to the subset of genetic candidates that are more likely to be related to the disease under investigation (Moreau and Tranchevent, 2012). The state of the art variant prioritization methods integrate heterogeneous sources of genetic and biological information (Sifrim *et al.*, 2013; Robinson *et al.*, 2014), where one of the most critical features are the scores of pathogenicity for single nucleotide variants (SNVs) as predicted by tools such as PolyPhen2 (Adzhubei *et al.*, 2010), Sift (Ng and Henikoff, 2001), MutationTaster (Schwarz *et al.*, 2010), Mutation Assessor (Reva *et al.*, 2011), PhyloP (Cooper *et al.*, 2005), GERP++ (Davydov *et al.*, 2010), and CADD (Kircher *et al.*, 2014). These tools use machine learning (ML) approaches to address the variant-effect prediction problem from many different angles related to the known proxies for pathogenicity, such as the evolutionary conservation of the mutated position (Adzhubei *et al.*, 2010; Ng and Henikoff, 2001; Zeng *et al.*, 2014), the stability change upon mutation (Dehouck *et al.*, 2009; De Baets *et al.*, 2011), possible structural alterations (Adzhubei *et al.*, 2010; De Baets *et al.*, 2011), the change in physico/chemical characteristics due to the mutation (Stone and Sidow, 2005) and functional annotations using GO terms (Calabrese *et al.*, 2009).

The phenotypical interpretation of protein-level alterations on the level of individuals is the ultimate goal of this field (Tavtigian *et al.*, 2008; van den Berg *et al.*, 2015), but this causal relationship is still far from being completely understood and it is confounded by many aspects related to the intrinsic complexity of cell life (Reumers *et al.*, 2009; Sahni *et al.*, 2015). A crucial restriction of variant-effect prediction is that an alteration of the protein's molecular phenotype, even if it is a *sine qua non* condition for the disease phenotype in the carrier individual, may not constitute in itself a sufficient cause for the disease: this also depends on the particular role that the affected protein plays in the well-being of the organism (Reumers *et al.*, 2009; Sahni *et al.*, 2015; Yates and Sternberg, 2013). Even the most commonly used features, which relate evolutionary constraints with likely functional damage, offer only a partial correlation with the pathogenicity of the variant (Tavtigian *et al.*, 2008). Consequently, additional information that bridges the variant-phenotype gap is crucial to improve variant-effect predictions.

Here, we present DEOGEN, a novel method for variant-effect prediction that integrates heterogeneous sources of information in order to analyze each (*protein*, *variant*) pair by combining different levels of contextualisation of the protein function with a Random Forest (RF) (Breiman, 2001) predictor. After a cross-validation with very strict settings, DEOGEN performs 10% better than the most recently developed variant-effect predictors that target missense SNVs. Additionally, we show that our approach is general enough to provide an improvement in the prediction of the phenotypic effects of short in-frame genetic variants (INDELs), where, after retraining the RF using almost the same features as for SNVs, DEOGEN performs 17% better than variant-effect predictors without multilevel contextualization. These improvements show that adding levels of contextualization to variant effect prediction helps to approach the complexity of cell life, an effect observed for both SNVs and INDELs. The increased contextualization will also, in time, allow a better understanding of why particular variants might have a particular deleterious effect.

## 2 Methods

### 2.1 Datasets

The main dataset used for benchmarking and training our predictor is the September 2011 version of Humsavar (Humsavar11), which contains 20 821 deleterious SNVs and 36 825 polymorphisms over 11 969 proteins. In order to provide further comparison with other state of the art methods, we ran also a separate cross-validation on Humsavar13, the January 2013 version of Humsavar, which contains 22 617 disease SNVs and 37 331 polymorphisms. For the benchmarking and training of the INDEL predictions, we used the dataset from PROVEAN (Choi *et al.*, 2012), which has been made publicly available by the authors and allows a direct comparison of the results. This dataset, which we call INDVAR, contains 1038 in-frame variants involving 1–6 aminoacids: 729 deletions, 171 insertions and 138 replacements (in-frame substitution of multiple amino acids), mapped on 520 proteins. 841 of them (84%) are annotated as deleterious and 197 (16%) are polymorphisms.

### 2.2 Variant-oriented features

We used three evolutionary derived scores calculated from multiple sequence alignments (MSAs) to assess how likely the protein is to tolerate the mutation. Conceptually these scores can be divided into *column-wise* and *sequence-wise* categories. In the former category, the value is derived by considering only the position of the sequence (and thus a column in the MSA) in which the mutation occurs, while in the latter the entire sequence is taken into account while calculating the score.

#### 2.2.1 Column-wise features

For each protein for which we want to predict the outcome of the mutations (the query protein) we obtained the MSAs using JackHMMER (Eddy 2011) with one iteration and $E$ value of 0.1 against NCBInr database. The collected homologs were further filtered as a function of a Sequence Identity greater than 85% and a coverage higher than 80% in order to keep only the sequences that are functionally similar to the query sequence. From this MSA, we calculated the Conservation Index (CI), as proposed in (Calabrese *et al.*, 2009). In the following equation, $A$ is the set of the possible 20 aminoacids and $i$ is the position of the MSA in which the mutation occurs. $f_a(i)$ is thus the frequency of occurrence of aminoacid $a$ at position $i$ while $f_a$ is the frequency of $a$ as observed in the entire alignment:

$$CI(i) = \sqrt{\sum_{a \in A} (f_a(i) - f_a)} \quad (1)$$

The other column-wise score is LOR, the log-odd ratio of observing the wildtype amino acid $w$ with respect to the mutated amino acid $m$ at the target position: $f_w(i)$ is the frequency of occurrence of wildtype ($w$) aminoacid at position $i$, while $f_m(i)$ is the frequency of the mutated ($m$) aminoacid in the same column of the MSA.

$$LOR = \log\left(\frac{f_w(i)}{1-f_w(i)}\right) - \log\left(\frac{f_m(i)}{1-f_m(i)}\right) \quad (2)$$

#### 2.2.2 Sequence-wise features

The PROV score is taken from the PROVEAN predictor (Choi *et al.*, 2012) and evaluates the deleteriousness of a mutation by comparing the wild-type and mutated sequence with closely related *functional* homologous sequences in the MSA. Mutations that reduce the similarity between the query protein and the *functional* homologs collected with PBLAST are considered more likely to be deleterious. This change in similarity is calculated by averaging many 'delta alignment scores' (Choi *et al.*, 2012) over close homologs. Each delta score is computed as $\Delta(Q, v, S) = A(Q', S) - A(Q, S)$, where $A$ is a semiglobal pairwise alignment method, $Q$ and $Q'$ are the query protein before and after the implementation of the variant $v$ on it and $S$ is one of the close functional homolog (Choi *et al.*, 2012). Because this score is computed on the entire length of the sequence it is natively able to deal with variants that range over multiple positions of the protein such as insertions, deletions and replacements (INDELs).

### 2.3 Protein-oriented features

#### 2.3.1 Node degree in PPI networks

We retrieved the protein–protein interaction (PPI) networks provided in ConsensusPathDB (Kamburov *et al.*, 2011), since it integrates in a complementary and non-redundant way *Homo sapiens* binary and complex protein–protein interactions from many existing public PPI databases. We collected 115 673 interactions between 11 370 proteins. The DGR feature reflects the degree of a node, which is defined by the number edges incident on it. It is the simplest measure of centrality for a given node, and can discriminate between nodes that are loosely connected to the rest of the networks and *hubs* with thousand of edges.

#### 2.3.2 Pathway annotations

We retrieved 3660 biological pathways annotated in ConsensusPathDB (Kamburov *et al.*, 2011), and adopted the log-odd score approach used in (Calabrese *et al.*, 2009) to encode the pathway information into an ML-understandable feature. Here, we briefly recap the log-odd score calculation procedure for sake of clarity.

The pathways' log-odd scores uses the cross-validation training sets to learn which pathways are more sensitive or more tolerant to deleterious mutations. For each pathway $\mathcal{P}$, we initialize to 1 two counters $C_n^{\mathcal{P}}$ and $C_d^{\mathcal{P}}$, respectively indicating the number of neutral ($n$) and deleterious ($d$) variants annotated on $\mathcal{P}$. The training phase follows the following procedure: for each protein $p$ in the training set, we collect from the database the set $\mathcal{S}^p$ of the pathways in which $p$ is involved, translating from protein to gene names. Then, for each pathway $x \in \mathcal{S}^p$ and for each variant $v$ mapped on $p$, we add 1 to $C_n^x$ if $v$ is neutral or 1 to $C_d^x$ if $v$ is deleterious. In this way we assign to each pathway $x$ the number of deleterious ($C_d^x$) and neutral ($C_n^x$) variants annotated on it in the training set. We then compute the probability of having an harmful or neutral variant on the pathway $x$ by calculating $P_x(\text{del}) = C_d^x/(C_n^x + C_d^x)$ and $P_x(\text{neut}) = C_n^x/(C_n^x + C_d^x)$. In order to provide an final measure of the tendency of pathway $x$ to be harmful we compute the log-odd score $\log(P_x(\text{del})/P_x(\text{neut}))$.

Once the log-odd scores have been inferred from the training set, in order to predict a variant $m$ for a protein $p$ in the test set, we collect the set of pathways $\mathcal{S}^p$ in which $p$ is involved and we sum all the log-odd scores learned from the training set as described in the previous paragraph. The final one-dimensional PATH feature is computed as:

$$PATH = \sum_{x \in \mathcal{S}^p} \log(P_x(\text{del})/P_x(\text{neut})) \quad (3)$$

#### 2.3.3 Recessiveness index

The recessiveness score (REC) is obtained from a linear predictor, developed in (MacArthur *et al.*, 2012) that discriminates between loss of function (LoF) tolerant genes and genes that can cause recessive disorders when homozygously lost. From dbNFSPv2.8 (Liu *et al.*, 2011, 2013), we collected REC predictions for 14 070 proteins.

#### 2.3.4 Essential genes in mouse

From dbNFSPv2.8 (Liu *et al.*, 2011, 2013) we collected also annotations for the degree of *essentiality* of 6283 proteins, as derived from knock-out experiments in mice on human orthologs genes (Georgi *et al.*, 2013). We included this knowledge in the ESS score because it can help characterizing the importance of the proteins on which the mutations occur, and so whether mutations will likely have a strong effect on the organism.

### 2.4 Machine learning

In summary, we used three variant-oriented evolutionary-based features (CI, LOR and PROV) to contextualize the mutation at the molecular level: CI and LOR are *column-wise* while PROV takes the entire sequence into account, making it a *sequence-wise* score. To complete the contextualization, we used also four protein-oriented features that encompass information about the biological relevance of the protein/gene in which the variants occur: the number of protein–protein interactors (DGR), the sensitivity to deleterious mutations of the biological pathway involved (PATH), the REC and the notion of to what extent the gene should be considered essential (ESS).

To obtain the final predictions from these features, we used the RF (Breiman, 2001) classifier implemented in the `scikit-learn` Python library (Pedregosa *et al.*, 2011). Different 200-trees models were developed for SNVs and INDELs: for SNVs we limited the minimum number of samples required to split an internal node to 100 and the minimum number of samples in newly created leaves to 20 further reducing the possibility of overfitting. For INDELs, we lowered these parameters due to the smaller size of the dataset and allowed node splitting and leaves creation with at least 10 samples.

The column-wise variant-oriented scores had to be treated differently for INDELs. The CI score was adapted for each INDEL $I$ involving $n$ positions by average the column-wise scores over $n$: $\text{CI}_{INDEL} = n^{-1}\sum_{i \in I} \text{CI}(i)$. Adapting the LOR score to variants spanning multiple positions is not always feasible due to the possible lack of correspondence between wild type and mutant amino acids, and we thus suppress this feature when dealing with INDELs.

We obtained the prediction performances shown in the Results Section through a 10-fold cross validation procedure. We designed

the folds by clustering the sequences in Humsavar11 and in INDVAR as a function of their homology using BlastClust. Proteins in different folds share less than 30% sequence similarity at 90% of coverage, in order to avoid any overestimation of the results due to overfitting. We evaluated the performances with the commonly adopted scores sensitivity (SEN), specificity (SPE), precision (PRE), balanced accuracy (BAC), and Matthews Correlation Coefficient (MCC). In particular, BAC and MCC are not affected by unbalance in the dataset between the positive (Deleterious) and negative (Polymorphism) classes.

## 3 Results and discussion

### 3.1 Contextualization of SNVs

We address the inherently complex variant-effect prediction problem through the integration of different sources of information. By describing each (*protein*, *variant*) pair from different perspectives corresponding to different *levels of contextualisation*, we assembled the most relevant and accessible pieces of information that are currently available, with the aim to elucidate the fuzzy and complex mapping between molecular-level alterations and the individual-level phenotypic outcome.

We visualize this mapping as a collection of concentric circles (see Fig. 1), where closest to the variant are the *variant-oriented* features that provide evolutionary-based descriptors of the molecular alteration introduced by the mutation in the protein. From the second layer on, the *protein-oriented* features provide higher level functional and biological contextualization of the gene/protein

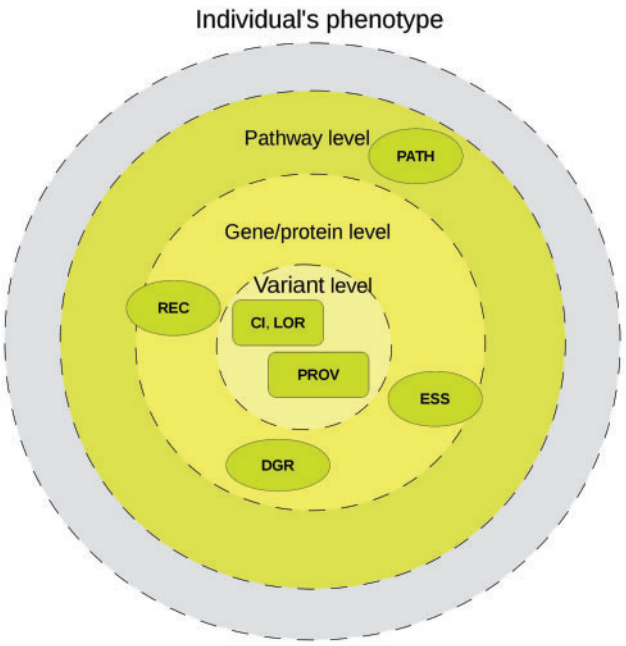


**Fig. 1.** Levels of contextualization within DEOGEN. Illustration of the heterogeneous sources of information combined within DEOGEN in order to contextualize each pair (*protein, variant*) from different levels of contextualization. PROV, CI, and LOR are variant-oriented features. In particular, CI and LOR are column-wise evolutionary conservation scores used to investigate the evolutionary constraints acting on the mutating position in the MSA while PROV is calculated using the entire sequence. ESS, DGR and REC are protein-oriented features, aiming at characterizing the functional role of the affected protein. The most abstract level of contextualization used is the pathway sensitivity to deleterious mutations, inferred using the log-odd score technique (PATH)

wherein the mutation occurs, so capturing parts of the complex role played by the affected protein in the cell and in the organism. The highest level of contextualization used here is the biological pathway level.

We use three *variant-oriented* features with different characteristics (see Methods and Fig. 1): the log-odd ratio (LOR) score and conservation index (CI) (Calabrese *et al.*, 2009), which are *column-wise* measures of the conservation of a mutated column within a multiple-sequence alignment (MSA), and the PROVEAN (Choi *et al.*, 2012) predictions (PROV), which provide a *sequence-wide* measure of the change in evolutionary distance between the mutated target protein and close functional homologs that correlates with the deleteriousness of variants (Choi *et al.*, 2012).

The *protein-oriented* features use pathway and protein–protein interaction (PPI) networks information (DGR) as well as genetic and clinical information, for instance an evaluation of how tolerant the carrier individual is to homozygous loss-of-function (LoF) mutations on the affected gene (REC). These protein-oriented features play an essential role in DEOGEN, and we explain and analyze them in more detail in the following sections, using Humsavar11 as reference dataset (see Methods).

### 3.2 Node degree in PPI networks

Protein function, from enzymatic reactions to participation in complex signaling cascades, requires interaction with other molecules, from proteins to nucleic acids and chemicals (Das *et al.*, 2014; Yates and Sternberg, 2013). An important piece of information at this level is the number of known interaction partners of the investigated protein; more interactions may indicate hub proteins that play a role in many processes. We retrieved the number of known interactors for Human proteins from ConsensusPathDB (Kamburov *et al.*, 2011) for use as a feature called degree (DGR). Supplementary Fig. 1 shows the DGR distributions separated into deleterious and polymorphism classes: although the distributions are partially superimposed, they are significantly different: both a two-sided Student's *t*-test over the expected values of independent samples and a two-sided Wilcoxon rank-sum test give a $P$ value smaller than $10^{-100}$. These results reveal that deleterious SNVs tend to occur in proteins with an higher number of binding partners than neutral SNVs. A possible problem underlying this data is that hub proteins tend to be more studied that peripheral ones, resulting in more known variants for these proteins. There is, however, no correlation between number of known variants for each gene versus its number of interactors (DGR) ($r = 0.089$), and therefore no bias (see Supplementary Fig. 5).

### 3.3 Recessiveness score

The REC (MacArthur *et al.*, 2012) predicts the ease with which a particular gene can cause deleterious phenotypes when homozygously lost by discriminating between LoF tolerant genes and genes involved in recessive diseases (MacArthur *et al.*, 2012). Supplementary Fig. 2 shows the distribution of the deleterious and polymorphism classes of SNVs as a function of the REC. Both the two-sided *t*-test and the Wilcoxon rank-sum test indicate significantly different distributions with $P$ values smaller than $10^{-100}$. Polymorphisms are denser at low scores (between 0.1 and 0.2), indicating an enrichment of neutral SNVs on genes predicted to be LoF tolerant, while deleterious SNVs are quasiuniformly distributed and range towards scores that suggest involvement in recessive diseases. There are 12 586 SNVs (22% of the total) which could not be mapped on REC scores because the carrier protein has not been predicted in (MacArthur *et al.*, 2012). The SNVs that were mapped to

the REC score amount to 68% of the polymorphisms and 89% of the disease related SNVs. Finally, there is also here no bias due to the preferential study of human health-related proteins, with a low correlation ($r = 0.23$) between each gene's REC score and the number of known SNVs mapped to it (see Supplementary Fig. 6).

### 3.4 Pathway annotations

We retrieved 3660 pathways involving 15 350 genes from ConsensusPathDB (Kamburov *et al.*, 2011), and mapped 23 231 SNVs of the total 57 646 (40%) available in the dataset Humsavar11 (see Methods) on at least one pathway. The mapped variants are divided into 11 020 deleterious (47%) and 12 195 neutral SNVs (53%), while of the remaining SNVs without pathway annotation 9801 are deleterious (28%) and 24 630 neutral (72%). This means that 53% of deleterious SNVs can be mapped on at least one pathway, while the same holds only for 33% of polymorphisms. Supplementary Fig. 4 shows the 25 pathways that are most likely (top panel) and least likely (bottom panel) to contain deleterious variants based on our analysis on Humsavar11. The likelihood of containing deleterious variants is expressed as a log-odd score (see Methods), where a zero score means that there is an equal balance between deleterious and non-deleterious variants, and a positive (negative) score indicates that the pathway contains mostly disease-related (neutral) variants.

The pathways most likely to contain deleterious variants are, in more detail, related to guanine and guanosine salvage (pathway name: PWY-6620, score: 3.93), Refsum Disease (SMP00451, 3.87), and Phytanic Acid Peroxisomal Oxidation (SMP00450, 3.87). All the 51 SNVs mapped on PWY-6620 are deleterious, while both SMP00451 and SMP00450 pathways have 2 neutral and 143 deleterious SNVs mapped to them. The genes in the top-ranking pathways tend to reoccur: for instance pathways HOMOCYSDEGR_PWY until SMP00515 contain the human gene CBS, which has 94 deleterious and 1 neutral variant mapped to it, and is instrumental in producing the high log-odd score of those pathways.

The bottom part of Supplementary Fig. 4 shows the pathways that are more likely to host neutral variants. These are the termination of O-glycan biosynthesis (REACT_115835, –5.02), the O-linked glycosylation of mucins (REACT_115606, –5.07) and the Olfactory Signaling Pathway (REACT_15488, –6.22). In particular, all of the 500 SNVs in Humsavar11 mapped on the Olfactory Signaling Pathway are polymorphisms, which is also the case for REACT_115835 and REACT_115606 (respectively 150 and 158 polymorphisms).

Supplementary Fig. 3 shows that in the majority of the pathways 15–30% of the genes are covered by at least one SNV. Moreover, it indicates that there is no correlation between the percentage of genes in the pathway that are covered by at least one SNV and the deleteriousness of the pathway itself ($r = -0.06$). Neither is there a bias towards pathway size (see Supplementary Fig. 8). The log-odd scores learned from Humsavar11 and INDVAr are available as Supplementary Material.

### 3.5 Essential genes in mouse

The combination of gene knock-out experiments in mice on human ortholog genes with human sequencing data have produced information about which genes should be considered *essential* (Georgi *et al.*, 2013): an organism is not viable without them because they operate key functions. Almost a third of the genes in the mouse genome are essential, and given the common architecture mouse and human share for many genetic disorders, it has been possible to characterize a set of 2472 human orthologs of known essential mouse genes (Georgi *et al.*, 2013). We retrieved these gene annotations at the protein level from dbNFSPv2.8 (Liu *et al.*, 2011), characterizing 6283 proteins. We refer to this feature as ESS in the rest of the paper. Also in this case we investigate the possible bias due to the preferential study of human health-relevant proteins: the correlation between the number of known SNVs mapped on a particular gene and its ESS score is $r = 0.011$ (see Supplementary Fig. 6).

### 3.6 Incremental contributions of features in SNV predictions

Based on the variant-oriented and new protein-oriented features, we evaluated the feature contributions within DEOGEN by cross-validating it on the Humsavar11 dataset, which contain slightly less than 60 000 SNVs annotated with the corresponding phenotypic effect. We chose this dataset because it is widely used and allow direct comparison of DEOGEN with the published performance of many different predictors.

DEOGEN integrates the variant-oriented and protein-oriented information in a RF model, where the heterogeneous data are processed and represented in the standardized form of numerical vectors, called *feature vectors*, on which the RF can perform inference and predictions. Since ML approaches are highly sensitive to data bias, we ensured as fair as possible an assessment of the prediction performances by applying a 10-fold stratified cross-validation, where the proteins in each fold have been stratified in order to share less than 30% sequence similarity with the proteins in the other folds (see Methods).

Table 1 shows the relative contribution of every feature to the predictive performance using the Humsavar11 dataset, adding each feature incrementally. Of particular interest are the BAC and Matthews Correlation Coefficient (MCC) scores, which are less influenced by the unbalance between the number of positive and negative instances in a training set: more polymorphisms than deleterious variants are known. We chose PROV as the baseline because it is in itself an advanced predictor that can natively deal with both SNVs and in-frame INDELs. The PROV performance is equivalent to what was reported before (Choi *et al.*, 2012) (Table 1). Adding

**Table 1.** DEOGEN performances on Humsavar 2011 as a function of different features

| Features | Sen | Spe | BAC | Pre | MCC | AUC |
|---|---|---|---|---|---|---|
| PROV | 70.1 | 84.2 | 77.1 | 70.7 | 54.1 | 84.9 |
| PROV+CI | 71.4 | 85.6 | 78.5 | 73.0 | 57.0 | 86.9 |
| PROV+CI+LOR | 73.3 | 87.3 | 80.3 | 75.9 | 60.8 | 89.2 |
| PROV+CI+LOR+PATH | 73.2 | 89.9 | 81.5 | 79.6 | 64.2 | 91.1 |
| PROV+CI+LOR+PATH+DGR | 73.2 | 90.7 | 81.9 | 81.0 | 65.4 | 91.8 |
| PROV+CI+LOR+PATH+DGR+REC | 76.3 | 91.6 | 84.0 | 83.2 | 69.2 | 93.3 |
| PROV+CI+LOR+PATH+DGR+REC+ESS | 77.4 | 91.9 | 84.7 | 84.0 | 70.6 | 93.7 |

the *column-wise* evolutionary conservation scores CI and LOR improves the prediction performance (+12% of MCC), showing that they provide information orthogonal to PROV. These additional features also improve the specificity and sensitivity of the predictor.

We then added the four *protein-oriented* features incrementally to the standard *variant-oriented* effects (see also Methods for a detailed explanation). The information provided by the PATH feature, which indicates the sensitivity to disease of the pathways the protein is part of, gives a 5.6% improvement in terms of MCC. We then added subsequently the number of interaction partners (DGR) in the protein-protein interaction network, the recessiveness (REC) index, which estimates the level of redundancy of the gene, and the ESS score, which quantifies how essential a gene is as evidenced by knock-out experiments in mice. Whereas every feature gives an incremental improvement of the BAC and MCC scores, all four protein-oriented features provide a combined MCC improvement of slightly more than 16% on top of the *variant-oriented* features. These results show that our multi-level contextualisation provides essential information and significantly increases the quality of the predictions.

### 3.7 Comparison with other methods

We compared the cross-validated performances of our predictor with five state-of-the-art predictors based on the Humsavar11 dataset (Table 2) and with six predictors on the Humsavar13 dataset (Table 3). For Humsavar11, the scores were taken from (Choi *et al.*, 2012), with the following specifications: the PROVEAN scores were obtained with the published suggested threshold of -2.282 (Choi *et al.*, 2012), Mutation Assessor scores were computed using the best performing threshold of 1.9 (Reva *et al.*, 2011), the webserver version of PolyPhen2 trained on HumDiv was used in order to avoid the over-estimation of the results due to the overlap between HumVar and Humsavar11 datasets (Adzhubei *et al.*, 2010), and SIFT predictions were obtained by running it on August 2011 NCBInr database (Ng and Henikoff, 2001). CADD (Kircher *et al.*, 2014) scores have been retrieved from dbNFSPv2.8 (Liu *et al.*, 2013). The optimal threshold for discriminating between deleterious and neutral SNVs has been inferred from the ROC curve and is 15.42. Table 2 shows that the different sources of information integrated by our method DEOGEN are effective, since they yield to a 26% improvement of the MCC with respect to PROVEAN, on which DEOGEN is based. DEOGEN performs roughly 36% better than SIFT and HumDiv-trained PolyPhen2, and 30% better than Mutation Assessor. DEOGEN also performs 28.6% better than CADD raw score in terms of MCC and +7.3% in terms of BAC.

The comparison on Humsavar13 is based on (Zeng *et al.*, 2014), where the predictions of SIFT were obtained through the webserver, the scores for GERP++ (Davydov *et al.*, 2010), PhyloP (Cooper *et al.*, 2005), MutationTaster (Schwarz *et al.*, 2010), and CADD (Kircher *et al.*, 2014) were extracted from dbNFSPv2.8 (Liu *et al.*, 2013), and the EFIN scores were obtained from the published 10 fold cross-validation (Zeng *et al.*, 2014). Table 3 shows that DEOGEN performs 18% better in terms of MCC and 10% better in terms of AUC with respect to Mutation Taster. DEOGEN AUC is also 10.8% better than the AUC obtained by CADD (Kircher *et al.*, 2014) raw score. DEOGEN MCC and AUC are also respectively 10% and 3.3% higher than the one obtained by EFIN, the most recent variant effect predictor in this benchmark. Although many other variant prediction methods like SNAP2 [Hecht *et al* (2015)] are available, we here compare only to the most widely adopted methods and refer to Dong *et al.* (2015) for an exhaustive comparison between methods on a blind test set.

To show the applicability of DEOGEN on real clinical cases, we describe in Supplementary Material its performance on three well studied genes with high impact for human health: TP53, F8, and BRCA1. We show that in the first two cases DEOGEN would be really useful in clinical studies since it has very high sensitivities (respectively, 98 and 79%), thus correctly selecting nearly all of the possibly causative SNVs. In the case of BRCA1 predictions are less satisfying because although the *protein-oriented* features highlight the relevance of this gene, the *variant-oriented* features are mostly undecided on their outcome, leading to a poor sensitivity (30%) at 71% specificity.

### 3.8 Contextualisation also improves in-frame INDELs prediction

The majority of the variant-effect predictors published to date are focused on the analysis of SNVs. They are the simplest and most common form of variants in the genome, are responsible for a large share of the observed human variability and disease susceptibility (Abecasis *et al.*, 2012; Cooper and Shendure, 2011; Studer *et al.*, 2013; Tennessen *et al.*, 2012), and account for roughly 67% of known pathological mutations (Stenson *et al.*, 2003; Zhao *et al.*, 2013). On the other hand, WES studies are also identifying a growing amount of INDEL variants (Mills *et al.*, 2006, 2011), which is considered the second largest class of genetic variants in our genome. In particular, short INDELs ranging from 2 to 20 base pairs appear to be responsible for 22% of known pathological mutations (Ball *et al.*, 2005; Zhao *et al.*, 2013). INDELs can be either frame-shifting or in-frame: frame-shifting INDELs cause the insertion/deletion of a number of nucleotides that are not a multiple of 3, thus causing the alteration of the reading frame and the consequential complete change of the gene product or early termination of the transcription. In-frame short INDELs, on the other hand, alter a multiple of three nucleotides, causing the insertion or deletion of 1–6 amino acids at the protein level (Zhao *et al.*, 2013). INDELs are

**Table 2.** Comparison of the predictors on Humsavar 2011

| Method | Missing (%) | Sen | Spe | Pre | BAC | MCC |
|---|---|---|---|---|---|---|
| PROVEAN[a] | 0.0 | 78.4 | 79.1 | 67.9 | 78.6 | 56.0 |
| SIFT[a] | 2.0 | 85.0 | 68.9 | 60.8 | 76.9 | 51.9 |
| Mutation Assessor[a] | 0.6 | 85.3 | 71.0 | 62.5 | 78.2 | 54.1 |
| PolyPhen2[a] | 4.0 | 88.7 | 62.5 | 57.2 | 75.6 | 49.5 |
| CADD[b] | 7.0 | 81.7 | 74.9 | 66.1 | 78.3 | 54.9 |
| DEOGEN | 0.6 | 77.4 | 91.9 | 84.7 | 84.0 | 70.6 |

[a]Results reported from (Choi *et al.*, 2012).

[b]CADD (Kircher *et al.*, 2014) raw score for functional prediction of SNVs as extracted from dbNFSPv2.8 (Liu *et al.*, 2013).

**Table 3.** Comparison of the predictors on Humsavar13

| Method | Missing (%) | Sen | Spe | Pre | BAC | MCC | AUC |
|---|---|---|---|---|---|---|---|
| EFIN[a] | 0.0 | 86.4 | 79.5 | 86.7 | 82.9 | 64.2 | 90.7 |
| GERP++[a] | 20.7 | 96.8 | 24.4 | 45.3 | 60.6 | 28.1 | 76.1 |
| PhyloP[a] | 20.7 | 96.5 | 27.3 | 46.2 | 61.9 | 30.3 | 76.3 |
| Mutation Taster[a] | 20.7 | 86.4 | 74.9 | 69.1 | 80.7 | 59.9 | 85.4 |
| CADD[b] | 15.6 | 81.7 | 74.9 | 66.1 | 78.3 | 54.9 | 84.6 |
| DEOGEN | 4.4 | 77.4 | 91.9 | 84.7 | 84.0 | 70.6 | 93.7 |

[a]Results reported from (Zeng *et al.*, 2014).

[b]CADD (Kircher *et al.*, 2014) raw score for functional prediction of SNVs as extracted from dbNFSPv2.8 (Liu *et al.*, 2013).

generally more disruptive than SNVs, but the effect of purifying selection greatly distinguishes between frameshifting (94% of them are eliminated) and in-frame INDELs (48% eliminated) (Studer *et al.*, 2013), leading to a frequency of one in-frame INDEL every 7 SNVs (Ng *et al.*, 2008). Polymorphic INDELs mutations are in fact also involved in changes of function between the different superfamily domains (Studer *et al.*, 2013; Reeves *et al.*, 2006).

The existing bioinformatics tools developed to predict the pathogenicity of short in-frame INDELs (Choi *et al.*, 2012; Hu and Ng, 2013; Zhao *et al.*, 2013) focus only on *variant-oriented* features. In particular, PROVEAN (Choi *et al.*, 2012) uses an unsupervised *sequence-wise* evolutionary score, while SIFTindel (Hu and Ng, 2013) and DDIG-in (Zhao *et al.*, 2013) combine different features such as secondary structure, domain annotation, disorder prediction, and relative solvent accessibility with ML methods. Here we compared our performances only with PROVEAN because SIFTindel and DDIG-in do not provide prediction performances on a publicly available dataset or comparisons with other methods, although the authors argue that these three methods have similar performances (Hu and Ng, 2013). We thereby show that the SNV conclusions hold for the pathogenicity prediction of INDELs, which also benefits from the multilevel contextualization.

#### 3.8.1 Incremental contributions of features in INDELs predictions

To activate DEOGEN for the prediction of the deleteriousness of in-frame INDELs, we trained a RF model on the INDVAR dataset, which contains 1038 short in-frame INDELs. Analogously to the SNVs prediction (Table 1), we show that incrementally adding the protein context features also leads to better predictions in the INDELs case (Table 4). The same stratified cross-validation procedure was used as with the SNVs case. The PROV score again serves as the baseline feature. Adding the CI, our column-wise score of evolutionary conservation now averaged over the positions involved in the INDEL, provides a 3.5% improvement of the MCC. The pathway-derived log-odd scores also appear to be relevant also for INDELs and provide a +9% improvement of the MCC. In order to identify possible over-fitting caused by the small number of INDELs available in INDVAR, we differentiated between the pathway log-odd scores learned on INDVAR alone ($PATH^{i}$) or on INDVAR plus Humsavar11 ($PATH^{i+s}$) (Table 4). The MCC and AUC scores obtained by learning the odds on INDVAR or INDVAR plus Humsavar11 are very similar, indicating that the information extracted via the log-odd scores is general enough to be transferred from SNVs to INDELs. This is possible despite INDELs being generally more deleterious, and despite slightly different 'most sensitive' pathways being derived when analyzing these different datasets (see Supplementary Figs 9 and 10). The DGR feature causes a slight decrease of MCC but a positive effect (+1.9%) on the AUC, while adding the REC score yields the best prediction performances reached by DEOGEN. ESS appears to provide unrelevant contributions and was not included in the final predictor. The last row of Table 4 shows that the MCC and AUC we obtain for the INDELs prediction are similar to the ones obtained in the SNVs case, respectively around 0.7 and 0.92. If we compare our best MCC on INDELs with the one obtained by PROVEAN (Choi *et al.*, 2012) on the same dataset, it appears that our method gives a 17.4% improvement. The performances of DEOGEN INDELs predictor in function of the type of INDEL (insertions, deletions and replacements) are shown in Supplementary Material Table 1.

**Table 4.** DEOGEN performances on INDVAR as a function of different features

| Method | Sen | Spe | BAC | Pre | MCC | AUC |
|---|---|---|---|---|---|---|
| PROV | 96.1 | 60.8 | 78.4 | 91.6 | 59.7 | 88.0 |
| PROV + CI | 97.0 | 59.7 | 78.3 | 91.5 | 61.8 | 90.1 |
| PROV + CI + $PATH^{s+i}$ | 97.6 | 63.5 | 80.5 | 92.2 | 67.1 | 91.0 |
| PROV + CI + $PATH^{i}$ | 97.4 | 64.8 | 81.1 | 92.5 | 67.4 | 90.5 |
| PROV + CI + $PATH^{i}$ + DGR | 97.5 | 64.0 | 80.7 | 92.3 | 66.8 | 92.2 |
| PROV + CI + $PATH^{i}$ + DGR + REC | 97.4 | 68.8 | 83.1 | 93.2 | 70.0 | 93.1 |
| PROV + CI + $PATH^{i}$ + DGR + REC + ESS | 97.4 | 68.6 | 83.0 | 92.9 | 70.1 | 92.1 |

$^{s+i}$pathway log-odds scores learned considering both Humsavar11 and INDVAR.

$^{i}$pathway log-odds scores learned considering only INDVAR.

## 4 Conclusion

The novel variant-effect predictor we present here integrates heterogeneous sources of biological information in order to improve the contextualization of both the variant and the affected protein. We do this by merging relevant aspects to the molecular phenotype, the gene and the biological pathways involved. We also show that this approach is general enough to be applicable to both SNVs and INDELs, even if in the latter case the average impact at the molecular levels is much more disruptive. This functional contextualization of the protein enables DEOGEN to outperform nine state of the art predictors, indicating that careful integration of data from relevant sources is an excellent approach to improve not only predictions, but also our understanding of the mechanisms underlying the genotype-to-phenotype relationship. Future developments will dig deeper into the interpretation of the single predictions, aiming at understanding how a particular variant influences cell machinery and its functioning.

## Acknowledgement

D.R. thanks Anna Laura Mascagni for support and helpful discussions.

## Funding

D.R. is funded by the Agency for Innovation by Science and Technology in Flanders (IWT). W.F.V. is funded by the Brussels Institute for Research and Innovation (Innoviris) grant BB2B 2010-1-12. M.R. is Research Director at the Fund for Scientific Research - FNRS. A.G. and T.L. are supported by a regional ARC project entitled 'Deciphering Oligo- and Polygenic Genetic Architecture in Brain Developmental Disorders' and the project 'BridgeIris' funded by the Brussels Institute for Research and Innovation (Innoviris)

*Conflict of Interest*: none declared.

## References

Abecasis,G.R. *et al.* (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature*, **491**, 56–65.

Adzhubei,I.A. *et al.* (2010) A method and server for predicting damaging missense mutations. *Nat. Methods*, **7**, 248–249.

Ball,E.V. *et al.* (2005) Microdeletions and microinsertions causing human genetic disease: common mechanisms of mutagenesis and the role of local DNA sequence complexity. *Hum. Mutat.*, **26**, 205–213.

Bamshad,M.J. *et al.* (2011) Exome sequencing as a tool for Mendelian disease gene discovery. *Nat. Rev. Genet.*, **12**, 745–755.

Boycott,K.M. *et al.* (2013) Rare-disease genetics in the era of next-generation sequencing: discovery to translation. *Nat. Rev. Genet.*, **14**, 681–691.

Breiman,L. (2001) Random forests. *Mach. Learn.*, **45**, 5–32.

Calabrese,R. *et al.* (2009) Functional annotations improve the predictive score of human disease related mutations in proteins. *Hum. Mutat.*, **30**, 1237–1244.
Choi,Y. *et al.* (2012) Predicting the functional effect of amino acid substitutions and indels. e46688.
Cooper, D.N. *et al.* (1995) The nature and mechanisms of human gene mutation. In: Scriver, C. *et al.*, (eds.) *The Metabolic and Molecular Bases of Inherited Disease*. McGraw-Hill, NY. 259291.
Cooper,G.M. and Shendure, J. (2011) Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nat. Rev. Genet.*, **12**, 628–640.
Cooper,G.M. *et al.* (2005) Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res.*, **15**, 901–913.
Das,J. *et al.* (2014) Elucidating common structural features of human pathogenic variations using large scale atomic resolution protein networks. *Hum. Mutat.*, **35**, 585–593.
Davydov,E.V. *et al.* (2010) Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput. Biol.*, **6**, e1001025.
De Baets,G. *et al.* (2011) SNPeffect 4.0: on-line prediction of molecular and structural effects of protein-coding variants. *Nucleic Acids Res.*, **40**, D935–D939.
Dehouck,Y. *et al.* (2009) Fast and accurate predictions of protein stability changes upon mutations using statistical potentials and neural networks: PoPMuSiC-2.0. *Bioinformatics*, **25**, 2537–2543.
Dong,C. *et al.* (2015) Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Hum. Mol. Genet.*, **24**, 2125–2137.
Eddy,S.R. (2011) Accelerated profile HMM searches. *PLoS Comput. Biol.*, **7**, e1002195.
Georgi,B. *et al.* (2013) From mouse to human: evolutionary genomics analysis of human orthologs of essential genes. e1003484.
Hecht,M. *et al.* (2015) Better prediction of functional effects for sequence variants. *BMC Genomics*, **16**, S1.
Hu,J. and Ng,P.C. (2013) SIFT Indel: predictions for the functional effects of amino acid insertions/deletions in proteins. e77940.
Kamburov,A. *et al.* (2011) ConsensusPathDB: toward a more complete picture of cell biology. *Nucleic Acids Res.*, **39**, D712–D717.
Kircher,M. *et al.* (2014) A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.*, **46**, 310–315.
Liu,X. *et al.* (2011) dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. *Hum. Mutat.*, **32**, 894–899.
Liu,X. *et al.* (2013) dbNSFP v2.0: a database of human non synonymous SNVs and their functional predictions and annotations. *Hum. Mutat.*, **34**, E2393–E2402.
MacArthur,D.G. *et al.* (2012) A systematic survey of loss-of-function variants in human protein-coding genes. *Science*, **335**, 823–828.
Mills,R.E. *et al.* (2006) An initial map of insertion and deletion (INDEL) variation in the human genome. *Genome Res.*, **16**, 1182–1190.
Mills,R.E. *et al.* (2011) Natural genetic variation caused by small insertions and deletions in the human genome. *Genome Res.*, **21**, 830–839.
Moreau,Y. and Tranchevent,L.-C. (2012) Computational tools for prioritizing candidate genes: boosting disease gene discovery. *Nat. Rev. Genet.*, **13**, 523–536.
Ng,S.B. *et al.* (2009) Targeted capture and massively parallel sequencing of 12 human exomes. *Nature*, **461**, 272–276.
Ng,P.C. and Henikoff, S. (2001) Predicting deleterious amino acid substitutions. *Genome Res.*, **11**, 863–874.
Ng,P.C. *et al.* (2008) Genetic variation in an individual human exome. *PLoS Genet.*, **4**, e1000160.
Pedregosa,F. *et al.* (2011) Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.*, **12**, 2825–2830.
Reeves,G.A. *et al.* (2006) Structural diversity of domain superfamilies in the CATH database. *J. Mol. Biol.*, **360**, 725–741.
Reumers,J. *et al.* (2009) Using structural bioinformatics to investigate the impact of non synonymous SNPs and disease mutations: scope and limitations. *BMC Bioinformatics*, **10**, S9.
Reva,B.Y. *et al.* (2011) Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res.*, **39**, e118.
Robinson,P.N. *et al.* (2014) Improved exome prioritization of disease genes through cross-species phenotype comparison. *Genome Res.*, **24**, 340–348.
Sahni,N. *et al.* (2015) Widespread macromolecular interaction perturbations in human genetic disorders. *Cell*, **161**, 647–660.
Schwarz,J.M. *et al.* (2010) MutationTaster evaluates disease-causing potential of sequence alterations. *Nature Methods*, **7**, 575–576.
Sherry,S.T. *et al.* (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.
Sifrim,A. *et al.* (2013) eXtasy: variant prioritization by genomic data fusion. *Nat. Methods*, **10**, 1083–1084.
Stenson,P.D. *et al.* (2003) Human gene mutation database (HGMD): 2003 update. *Hum. Mutat.*, **21**, 577–581.
Stone,E.A. and Sidow,A. (2005) Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity. *Genome Res.*, **15**, 978–986.
Studer,R.A.H. *et al.* (2013) Residue mutations and their impact on protein structure and function: detecting beneficial and pathogenic changes. *Biochem. J.*, **449**, 581–594.
Tavtigian,S.V. *et al.* (2008) In silico analysis of missense substitutions using sequence-alignment based methods. *Hum. Mutat.*, **29**, 1327.
Tennessen,J.A. *et al.* (2012) Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science*, **337**, 64–69.
van den Berg,B.A. *et al.* (2015) Insight into neutral and disease-associated human genetic variants through interpretable predictors. *PloS One*, **10**.
Yates,C.M. and Sternberg,M.J.E. (2013) The effects of non-synonymous single nucleotide polymorphisms (nsSNPs) on protein protein interactions. *J. Mol. Biol.*, **425**, 3949–3963.
Zeng,S. *et al.* (2014) EFIN: predicting the functional impact of nonsynonymous single nucleotide polymorphisms in human genome. *BMC Genomics*, **15**, 455.
Zhao,H. *et al.* (2013) DDIG-in: discriminating between disease-associated and neutral non-frameshifting micro-indels. *Genome Biol.*, **14**, R23.