**Supplementary information S1 (box): Variant Annotation Algorithms**

Numerous methods have been developed to infer the functional or selective effect of new variants. These methods vary in their statistical approaches and in the type of information they use. Both factors have a strong impact on predicted variant effect, and therefore great attention should be applied when choosing among those methods.

Many methods focus on protein-coding sequences, because coding sequences are thought to be functionally important, and we understand them better. The simplest approach in this category is a binary division of coding variants into synonymous (S) and nonsynonymous (NS), i.e. whether the variant leads to an amino acid change. A more sophisticated method based on a similar approach is Polyphen2[1], which further partitions NS variants based on their likely effect on protein function: it uses a protein structure database to perform a multiple sequence alignment and determine additional characteristic features that will be informative to predict the deleteriousness of the substitution. Some of these features involve the biochemical properties of the new amino acid and how conserved the protein is. A score is then calculated based on a Naïve Bayes classifier. This approach has the advantage of assigning a degree of deleteriousness based on how likely this protein is to be damaging, though the training set ultimately depends on a limited number of variants that were identified to be associated to a disease. Mutation Taster[2] is another algorithm that uses a broader variety of databases to establish the evolutionary conservation of the site and the likelihood that it has an effect on splice site changes or on protein function. It then also uses a Naïve Bayes classifier approach to assess the posterior probability of the disease potential of a given variant. The main novelty of Mutation Taster is that it is able to incorporate indels as well as non-coding substitutions, in addition to single amino-acid changes.

The main limitation of these algorithms is that their results depend on how representative the training sets are. As an example, biases in deleterious variant discovery for different populations, together with the difficulties of identifying small-effect variants, may lead to inaccurate results.

PROVEAN[3] and SIFT[4,5] are other methods limited to coding variants. They also support amino acid insertions and deletions and they differ from the previous methods in that they work directly with the amino acid sequence instead of the DNA sequence, and thus information provided by the nature of the substitution is not taken into account. This method compares the target sequence with other homologous sequences; forms clusters based on similarity, and the most related clusters are then used as a supporting sequence set to compute a deleteriousness score. This method is not affected by biases related with deleterious variant discovery, as does not require a training set beyond the database of homologous sequence. There is a substantial overlap in disease variant annotation between Polyphen2 and PROVEAN (87%) [http://provean.jcvi.org/about.v1.0.php], suggesting similar performance in detecting variants associated with diseases. However, there are discrepancies across methods to annotate variants that have not been previously associated to any disease[6]. Finally, a challenge that affects variant annotation methods is a reliance on reference or putatively ancestral genomic sequence. Polyphen2 was recently shown to have a reference bias, meaning that derived variants in the reference genome tend to be annotated as benign rather than functionally significant[7]. This reference bias has been addressed in the new version, however similar biases may also exist for other algorithms.

So far we have discussed methods that are designed to determine the functional effect of a substitution in a protein coding sequence. Other methods have been developed to predict the impact of any substitution along the genome. An example is SnpEff[8]. In a similar way to Polyphen2 or Mutation Taster, it also draws on databases of gene organization, splice variants and protein structure to determine the functional effect of nucleotide substitutions. It annotates not only NS changes, but also variants across the whole genome. Specifically, the biological effect inferred by SNPEff will depend on the biological unit that is impacted by the substitution. Typically, introns or variants upstream of a gene will have a “modifier effect”, but variants that are in splice sites will have a “high” effect. This method is blind to previous associations of variants to diseases, or conservation across phylogenies. However, it also relies on how well known is the structure of the human genome. Substantial work still needs to be done to unravel the functional units in the non-coding genome, which could lead to changes in the predicted effect.

Other methods that consider substitutions across the whole genome use a phylogenetic approach instead of relying on the biological interpretation of the variant locus. They assess the functional effect of a variant by determining how conserved it is across related species. The main difference between methods in this category compared to the previous ones is that they focus on evolutionary rather than functional effect: variants that cause a dramatic functional change in a protein but do not affect the fitness of the carrier (duplications, alternative metabolic pathways, dominance, trans regulation, etc.) will be annotated as neutral by these methods. GERP[9] and PhyloP[10] are methods that follow this approach. GERP identifies elements that have been conserved across multiple alignments by quantifying substitution deficits compared to what would be expected under a neutral scenario. The

underlying assumption is therefore that the absence of expected polymorphisms is due to selective constraint, and that any changes would have been deleterious. One limitation of this method is that analysis is limited to regions of the genome that can be reliably aligned across typically as much as 35 mammal species. Also, changes in the recombination map may affect how conserved variants are regardless of their sensitivity to purifying selection. PhyloP is a very similar method, combining alignment to predict selective effect, but it additionally annotates regions that display accelerated evolution (regions with more polymorphisms that would be expected under neutrality). This method first computes a null distribution of the number of substitutions expected under neutrality given the tree model, and then obtains p-values or conservation / acceleration scores using a variety of tests.

An alternate approach, FitCons[11], combines both functional annotation based on ENCODE, and associated patterns of polymorphism and divergence to estimate the "fitness consequence" of a point mutation in the genome. Briefly, like other functional genomic methods, fitCons groups genomic positions with similar assigned functional categories to form clusters. These clusters are then classified according to a "fitness consequence" score based on the patterns of polymorphism and genetic differentiation they show, compared to nearby neutral regions and accounting for positive and purifying selection. This method does not rely on the assumption that genomic elements are present at orthologous locations over long periods of time, as conservation-based methods do, and is not limited to previously described variants as many functional annotation methods. However, it depends on the accuracy of ENCODE. In addition to all the above computational prediction approaches, empirical methods that directly quantify the impact of thousands of nonsynonymous variants on protein-protein interactions and protein stability will soon be providing direct experimental evidence for the impact of variants on protein function[12].

All these approaches provide distinct information about putative variant effect. To integrate this information, CADD[13] generates a single prediction from multiple annotation sources, including other variant effect predictors. The prediction uses a support vector machine trained on its ability to distinguish a real dataset of human derived variants from a simulated dataset: the "real" dataset is obtained by counting differences between present-day humans and an inferred ancestral genome, and the "simulated" dataset is generated from the inferred ancestral genome through a *de novo* mutational model. Thus CADD assesses evolutionary importance, but integrates any piece of functional or evolutionary evidence available to reach impressive classification accuracy.

**References:**


1. Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nat Meth* **7,** 248–249 (2010).
2. Schwarz, J. M., Rödelsperger, C., Schuelke, M. & Seelow, D. MutationTaster evaluates disease-causing potential of sequence alterations. *Nat Meth* **7,** 575–576 (2010).
3. Choi, Y., Sims, G. E., Murphy, S., Miller, J. R. & Chan, A. P. Predicting the Functional Effect of Amino Acid Substitutions and Indels. *PLoS ONE* **7,** e46688 (2012).
4. Kumar, P., Henikoff, S. & Ng, P. C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc* **4,** 1073–1081 (2009).
5. Ng, P. C. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Research* **31,** 3812–3814 (2003).
6. Flanagan, S. E., Patch, A.-M. & Ellard, S. Using SIFT and PolyPhen to Predict Loss-of-Function and Gain-of-Function Mutations. *Genetic Testing and Molecular Biomarkers* **14,** 533–537 (2010).
7. Simons, Y. B., Turchin, M. C., Pritchard, J. K. & Sella, G. The deleterious mutation load is insensitive to recent population history. *Nature Genetics* **46,** 220–224 (2014).
8. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)* **6,** 80–92 (2012).
9. Cooper, G. M. *et al.* Distribution and intensity of constraint in mammalian genomic sequence. *Genome Research* **15,** 901–913 (2005).
10. Siepel, A., Pollard, K. S. & Haussler, D. New methods for detecting lineage-specific selection. Research in Computational Molecular Biology. Springer Berlin Heidelberg, 190–205 (2006).
11. Gulko, B., Hubisz, M. J., Gronau, I. & Siepel, A. A method for calculating probabilities of fitness consequences for point mutations across the human genome. *Nature Genetics* **47**, 276-283 (2015).
12. Wei, X. *et al.* A massively parallel pipeline to clone DNA variants and examine molecular phenotypes of human disease mutations. *PLoS Genet* **10,** e1004819 (2014).
13. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics* **46,** 310–315 (2014).