

From prioritisation to understanding: mechanistic predictions of variant effects

Greg Slodkowicz  & M Madan Babu 

The widespread application of sequencing technologies, used for example to obtain data from healthy individuals or patient cohorts, has led to the identification of numerous mutations, the effect of which remains largely unclear. Therefore, developing approaches allowing accurate *in-silico* prediction of mutation effects is becoming increasingly important. In their recent study, Beltrao and colleagues (Wagih *et al.*, 2018) describe an integrative approach for determining the effects of mutations from the perspective of protein structure, conservation and transcription factor binding. This allows for predicting the mechanisms underlying the most impactful variants rather than just identifying these variants.

Mol Syst Biol. (2018) 14: e8741

See also: O Wagih *et al.* (December 2018)

New variants are now routinely being discovered by large-scale sequencing of both healthy and patient cohorts, but understanding their phenotypic consequences remains a challenge. Existing methods can prioritise variants according to their potential to disrupt protein function, but these methods cannot uncover the mechanistic details underlying these disruptions. As a result, our understanding of the mechanisms underlying genetic diseases is lagging behind, despite the large number of existing prediction tools. In their recent work, Wagih *et al.* (2018) present Mutfunc, a new resource that provides mechanistic predictions of mutational impact on function by combining sequence conservation with analysis of mutational effects on protein stability, interaction interfaces, post-translational modifications and linear motifs. Additionally,

Mutfunc predicts disruptions to start and stop codons, and quantifies the impact of non-coding mutations, by analysing their effects on transcription factor binding sites, allowing predictions of gene expression deregulation. Mutfunc provides pre-computed mechanistic annotations for mutations in three model organisms: *Homo sapiens*, *Saccharomyces cerevisiae* and *Escherichia coli* (Fig 1).

Protein-coding mutations can lead to a range of outcomes, from having no deleterious effect to causing diseases of moderate or high severity. In clinical studies, well-described mutations are frequently outnumbered by uncharacterised variants discovered by high-throughput sequencing of patient cohorts. Mechanistic predictions from Mutfunc can be used to identify variants with similar effects to those of already characterised variants and thus assign effects to variants with unknown clinical significance. For instance, known pathogenic variants in glyoxylate and hydroxypyruvate reductase (GRHPR) reduce its catalytic activity by disrupting homodimerisation. Applying Mutfunc to GRHPR variants with uncertain clinical significance reveals a subset of mutations that affect the homodimerisation interface in a manner similar to the confirmed disease mutations. Similarly, Mutfunc identifies variants in P_{ARK2}, a protein implicated in Parkinson's disease, which have a destabilising effect similar to that of known pathogenic variants. These variants are predicted to cause a disease phenotype. Wagih *et al.* further validated Mutfunc by demonstrating a correlation between predicted functional impact and phenotype in 166 *S. cerevisiae* strains grown in 43 conditions. Combined per-gene burden scores

based on Mutfunc predictions were calculated, allowing the identification of a large number of significant associations between genotypes and growth rates.

These two types of validation highlight the usefulness of this approach in both focused studies aimed at separating neutral from deleterious variants, as well as in broader studies aimed at understanding the mechanistic details underlying previously identified genetic associations (e.g. in GWAS cohorts). Mutfunc can also be used to augment other genetic approaches, including polygenic risk scores which combine the predicted effects of many variants (Khera *et al.*, 2018) without providing insights into the underlying disease mechanisms.

Historically, a major impediment to adopting methods based on protein structure was the significant shortage of solved structures. Mutfunc addresses this by providing high-quality homology models, pre-computed using the ModPipe software, greatly increasing structural coverage. Importantly, predicting mutational impact is particularly challenging within disordered regions. These regions lack a fixed conformation and generally do not appear in crystal structures. Currently, predictions are based on sequence conservation, disruptions in known linear motifs and post-translational modification sites. Improving the ability to identify specific mechanisms by which mutations in disordered regions affect protein function, e.g. disorder-to-order transitions, would greatly increase the utility of mechanistic methods. In this context, experimental developments in large-scale mutagenesis (Staller *et al.*, 2018), phage display (Davey *et al.*, 2017; Sundell *et al.*, 2018), ligand foot-printing and mass spectrometry

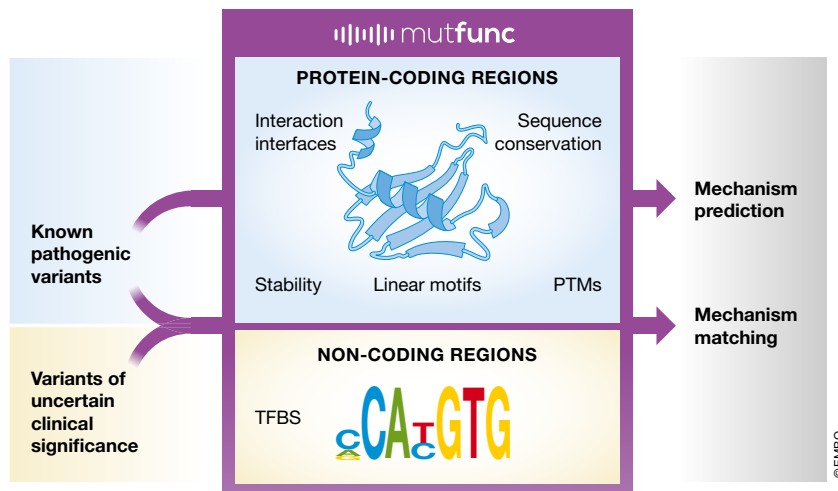


Figure 1. Mutfunc allows mechanistic predictions of variant effect for known pathogenic variants as well as variants of uncertain significance.

PTM, post-translational modification; TFBS, transcription factor binding site.

(preprint: Parker *et al.*, 2018) are beginning to provide insights into functional elements within disordered regions. Alongside these experimental approaches, hybrid experimental–computational approaches such as IDR-Screen that exploit machine learning (Ravarani *et al.*, 2018) will allow for new ways of uncovering mechanistic interpretation of mutations. Lastly, a more principled construction of multiple sequence alignments from which conservation is inferred can also improve predictions. In particular, distinguishing between orthologs and paralogs—which can be subject to very different evolutionary constraints—has been shown to improve variant effect prediction in disease context (Adebali *et al.*, 2016), as well as the identification of specificity-determining residues whose functional importance may be different among paralogs (Flock *et al.*, 2017).

On the basis of the analysed structural features, the set of variant effect predictions from Mutfunc allowed Wagih *et al.* to demonstrate that important functional regions of proteins are depleted in naturally occurring variants. Understanding the constraints imposed by protein structure is a fundamental question in evolutionary biology, and these insights are important both for theoretical and practical reasons. Aggregation of single nucleotide polymorphism (SNP) data can be used to identify constrained regions in proteins (preprint: Samochoa *et al.* 2017), and recent work on the structural placement of variants demonstrated that,

while naturally occurring SNPs tend to be dispersed on protein structures, pathogenic mutations are more likely to be clustered (Sivley *et al.*, 2018). Approaches based on clustering in sequence or in space can thus be used to identify new structural and functional features that can then be integrated into the methods predicting the mechanistic impact of mutations.

Most existing variant effect prediction methods fail to distinguish between loss- and gain-of-function mutations. Mechanistically distinguishing between these two effects would be a significant advancement and would also help identify dominant-negative or haploinsufficiency effects that can cause variants to have different penetrance. The effect of deleterious mutations can also be masked by compensatory effects (epistasis) and allele-specific expression or it can manifest upon environmental influences, including lifestyle choices. Currently, neither mechanistic nor other kinds of prediction tools can directly account for incomplete penetrance of mutations. Overcoming these limitations will require integrating predictions and findings across different levels of complexity, from mutational signatures and their mechanistic effects, to molecular phenotypes, cellular interaction networks and environmental influences.

We are entering a particularly exciting time as, in addition to sequence variation, other data types such as single-cell gene expression, structural, interaction, epigenetics, metabolomics, transcriptomics and

proteomics data are becoming increasingly available. Integration of high-quality unbiased data spanning different levels of biological organisation will be key for advancing our understanding of genotype-to-phenotype relationships, including disease mechanisms.

Acknowledgements

We thank Andrew Kleist for comments on the manuscript. The authors were supported by the Medical Research Council (MC_U105185859) and the LMB AstraZeneca Blue Sky project (BSF17).

References

- Adebali O, Reznik AO, Ory DS, Zhulin IB (2016) Establishing the precise evolutionary history of a gene improves prediction of disease-causing missense mutations. *Genet Med* 18: 1029–1036
- Davey NE, Seo MH, Yadav VK, Jeon J, Nim S, Krystkowiak I, Blikstad C, Dong D, Markova N, Kim PM, Ivarsson Y (2017) Discovery of short linear motif-mediated interactions through phage display of intrinsically disordered regions of the human proteome. *FEBS J* 284: 485–498
- Flock T, Hauser AS, Lund N, Gloriam DE, Balaji S, Babu MM (2017) Selectivity determinants of GPCR–G-protein binding. *Nature* 545: 317
- Khera AV, Chaffin M, Aragam KG, Haas ME, Roselli C, Choi SH, Natarajan P, Lander ES, Lubitz SA, Ellinor PT, Kathiresan S (2018) Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat Genet* 50: 1219
- Parker B, Goncz E, Krist DT, Statsyuk A, Nesvizhskii AI, Weiss E (2018) LiF-MS: mapping unstructured peptide-protein interactions using Ligand-Footprinting Mass Spectrometry. *bioRxiv* <https://doi.org/10.1101/361857> [PREPRINT]
- Ravarani CN, Erkina TY, De Baets G, Dudman DC, Erkin AM, Babu MM (2018) High-throughput discovery of functional disordered regions: investigation of transactivation domains. *Mol Syst Biol* 14: e8190
- Samochoa KE, Kosmicki JA, Karczewski KJ, O'Donnell-Luria AH, Pierce-Hoffman E, MacArthur DG, Neale BM, Daly MJ (2017) Regional missense constraint improves variant deleteriousness prediction. *bioRxiv* <https://doi.org/10.1101/148353> [PREPRINT]
- Sivley RM, Dou X, Meiler J, Bush WS, Capra JA (2018) Comprehensive analysis of constraint on the spatial distribution of missense variants in human protein structures. *Am J Hum Genet* 102: 415–426

Staller MV, Holehouse AS, Swain-Lenz D, Das RK, Pappu RV, Cohen BA (2018) A high-throughput mutational scan of an intrinsically disordered acidic transcriptional activation domain. *Cell Syst* 6: 444–455

Sundell GN, Arnold R, Ali M, Naksukpaiboon P, Orts J, Güntert P, Chi CN, Ivarsson Y (2018) Proteome-

wide analysis of phospho-regulated PDZ domain interactions. *Mol Syst Biol* 14: e8129

Wagih O, Galardini M, Busby B, Memon D, Typas A, Beltrao P (2018) A resource of variant effect predictions of single nucleotide variants in model organisms. *Mol Syst Biol* 14: e8430



License: This is an open access article under the terms of the Creative Commons Attribution 4.0 License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.