VIEWPOINTS

Multiplexed assays reveal effects of missense variants in MSH2 and cancer predisposition

Sofie V. Nielsen, Rasmus Hartmann-Petersen[ID], Amelie Stein[ID]*, Kresten Lindorff-Larsen[ID]*

Department of Biology, The Linderstrøm-Lang Centre for Protein Science, University of Copenhagen, Copenhagen, Denmark

* amelie.stein@bio.ku.dk (AS); lindorff@bio.ku.dk (KL-L)

DNA sequencing plays an increasingly central role in clinical research and diagnostics. Genome-wide association studies have established many links between genes and disease but do not reveal the effect of most of the many possible variants within each disease-related gene. Thus, while the explosion in sequencing of human genomes has revealed millions of missense variants that change protein sequences, we only understand the phenotypic and clinical consequences of a minute fraction of these. This lack of knowledge has direct consequences for clinical action. Even if a variant is discovered in a known disease-related gene, most variants have the status of "unknown significance" (VUS) simply because they have not been encountered before in the population or been studied in the laboratory.

Lynch Syndrome (LS) is a cancer predisposition syndrome that increases the risk of particularly colorectal and gynecological cancers [1]. LS is generally caused by loss-of-function (LoF) variants in one of several mismatch repair (MMR) genes, including *MSH2* [2]. Identification of pathogenic variants in *MSH2* would be of direct clinical relevance, but many missense variants are of unknown pathogenic significance. While computational methods exist to predict pathogenicity, including methods specific for MMR genes [3], they remain imperfect and are only considered as "supporting evidence" for variant classification [4].

For this reason, a number of experimental approaches have been undertaken to assess whether a specific missense variant in *MSH2* is pathogenic or not [5–7]. Some methods can provide detailed mechanistic understanding, yet they can be time consuming since each variant is handled individually and further, they are most easily applied retrospectively. Thus, most current functional assays are challenging to scale to the almost 18,000 possible single amino acid substitutions in MSH2, making it difficult to assign pathogenicity to any new clinically discovered variant.

In contrast, experiments based on multiplexed assays of variant effects (MAVEs; also sometimes known as deep mutational scanning) can be used to probe the effects of thousands of variants in a single experiment [8,9]. MAVEs combine developments in high-throughput DNA synthesis, functional assays, and rapid sequencing techniques. The first step in a MAVE is to construct a DNA library of variants that can be introduced into cells, e.g., by integration on the chromosome, on a plasmid, or by genome editing. The next step is to separate variants by a property of interest. This is often achieved by applying selective pressure, such that cells carrying a functional variant from the library will have higher growth rates than those with nonfunctional variants or, alternatively, by coupling to observable phenotypes like fluorescence followed by cell sorting. The relative frequencies of the variants in the library change depending on how well they are able to perform under selective conditions and are determined by DNA sequencing of the pool of cells before and after the selection. Finally, each variant's

Competing interests: The authors have declared that no competing interests exist.

change in frequency is used to compute a score (normalised to wild-type fitness) that quantifies the effect of the variant on the property selected for.

Two recent studies have taken advantage of the MAVE technology to investigate LoF variants in MSH2, assaying two different aspects of MSH2 function [10,11]. Impressively, Jia and colleagues score the function of 94.4% of all possible MSH2 variants, with the goal of identifying missense variants that cause LoF [11]. The assay that they use probes the ability of a given MSH2 variant to mediate G2-M arrest and cell death following treatment with 6-thioguanine (6TG) [12] and has previously been used to classify MSH2 VUS in low throughput [5]. Thus, wild-type-like MSH2 variants will be selected against, and as the study aims to clarify which of the many reported VUS in MSH2 that are potentially pathogenic, it is advantageous that the assay selects for nonfunctional variants. This selection strategy may have proven to be particularly important in this case since 89.4% of all tested variants were characterized as neutral, and MSH2 thus appears highly tolerant to single amino acid substitutions. In fact, 510 out of 934 positions tolerated substitution to any amino acid. In contrast, substitutions to proline or any of the charged amino acids appear to be particularly detrimental, and the majority (77%) of detrimental variants are buried within the native MSH2 structure. Finally, Jia and colleagues compare the ability of the functional scores to classify a curated set of pathogenic and benign variants and find that the experimentally obtained score outperformed several commonly used computational pathogenicity predictors.

The approach taken by Ollodart and colleagues is based on a multiplexed version of a canavanine-resistance assay [13,14], which they use to probe the mutation rate in yeast cells expressing one of ca. 200 different MSH2 variants. They validate their multiplexed experimental setup using a set of variants from a previous study [15], which probed mutation rates for 55 MSH2 variants, one at a time, using a similar canavanine-resistance assay. Finally, they measure mutation rates on a curated set of 185 variants from ClinVar and other clinical sources, which includes benign, pathogenic, and VUS, and find that the assay captures most of these pathogenicity classifications.

The experiments reported by Ollodart and colleagues directly quantify the mutation rates in a mixed cell population but, as of yet, does not scale to the same number of variants as the 6TG survival experiments by Jia and colleagues. The stochastic nature of spontaneously arising mutations makes it challenging to assess in a pooled experiment. The 6TG assay is, on the other hand, specific for MMR proteins, whereas mutation rate measurements are more broadly applicable. Further, the microsatellite instability observed in LS likely reflects an increased mutation rate rather than a failure to signal to G2-M arrest [2]. However, despite differences in both organism and assay, the two analyses largely agree on the functional status of the variants that were probed in both assays (Fig 1A), lending strong support for their use in assigning functional consequences to variants. Of the 176 variants that were assessed by both studies, 86 scored wild-type-like in both, 51 scored as LoF in the 6TG assay and increased mutational rate, and for 39 variants (22%) were there discrepancies between the two studies.

We here briefly discuss the four of these discordant variants with classified phenotypes. E198G is currently classified as benign, and the MAVEs show low mutation rate, but also resistance to 6TG. The variant has previously been shown to have low protein levels and cause functional defects [15,23,24], and E198G is not found in the population sequencing aggregation database gnomAD [16]. G827R showed an increased mutation rate, has been seen as a somatic mutation in a tumor with microsatellite instability, and revision by a clinical expert concluded that the histological phenotype is consistent with pathogenicity. Visual inspection of the MSH2:MSH6 complex (PDB ID 2o8e; [19]) shows that G827 is in the protein–protein interface and that larger and charged side chains such as R would likely perturb binding. The
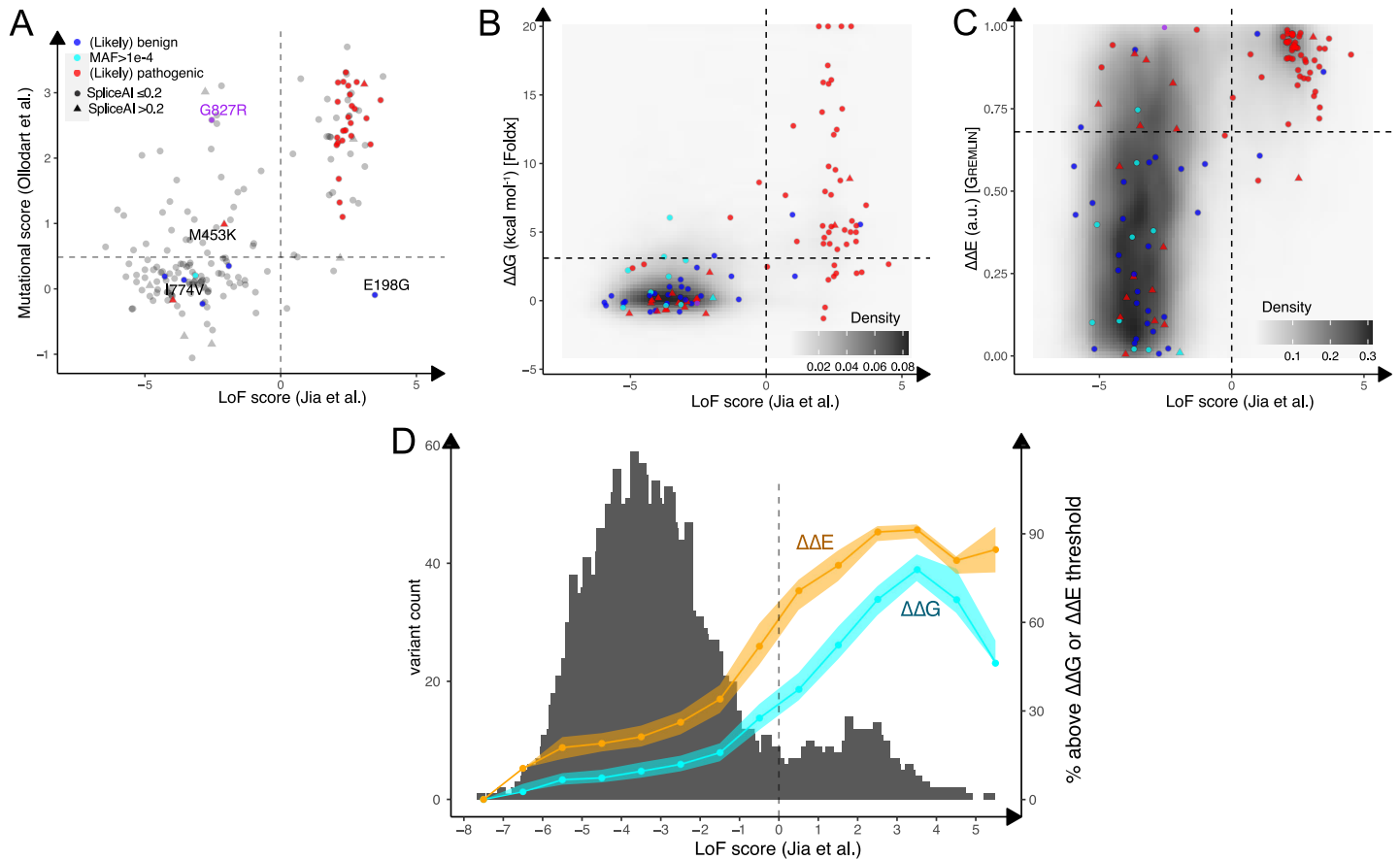
Fig 1. Comparison of MAVE and computational scores for MSH2 variant consequences. (A) LoF scores (on $\log_2$-scales relative to wild type) from the 6TG assay [11] versus mutation rate fold change [10] for 176 variants probed in both assays with known pathogenic variants shown in red and benign variants in blue; VUS with MAF $>10^{-4}$ [16], and therefore more likely to be benign, shown in cyan. The experimental setups cannot capture effects of potential effects on splicing, so we annotate (triangles) variants where the corresponding change in the DNA is predicted [17] to affect splicing (SpliceAI score >0.2). (B) LoF scores [11] versus $\Delta\Delta$Gs as previously calculated [18] on the MSH2 structure (PDB 2o8e [19]) by FoldX [20]. Variant labels are as in panel A. The gray background density shows the effect of all other variants, with the majority being predicted to have a small effect on stability ($\Delta\Delta$G<2 kcal/mol) and to be functional in the MAVE. For visual clarity, extremely destabilizing variants ($\Delta\Delta$Gs >20 kcal mol^{-1}) are visualized at 20. (C) LoF scores [11] versus $\Delta\Delta$Es as previously calculated [18] using Gremlin [21] and illustrated in the same way as in panel B. The dashed horizontal line indicates the threshold dividing pathogenic from nonpathogenic variants calculated by ROC analysis as described in [22]. (D) Distribution of MSH2 variants by LoF score (gray) and fraction of variants in each bin that are above the $\Delta\Delta$G (cyan) and $\Delta\Delta$E (orange) thresholds of 3.1 kcal mol^{-1} and 0.68, respectively. Shaded areas indicate the error of the fractions arising from uncertainty of the threshold. $\Delta\Delta$G and $\Delta\Delta$E values are available at https://github.com/KULL-Centre/PRISM/tree/master/data/proteins/Msh2. 6TG, 6-thioguanine; LoF, loss-of-function; MAF, minor allele frequency; MAVE, multiplexed assays of variant effect; ROC, receiver operating characteristic; VUS, variant of uncertain significance.

https://doi.org/10.1371/journal.pgen.1009496.g001

G827R somatic variant was, however, found in a patient that also carries a S676L variant, which may itself be pathogenic [10,25]. M453K is 6TG sensitive but has a moderately increased mutation rate; it has also been described to affect splicing and may be pathogenic for this reason [26]. Finally, I774V scores as wild-type-like in both MAVEs but is listed as pathogenic in ClinVar and is predicted to affect splicing [17]. Variant pathogenicity due to introduction (or removal) of splice sites hence appears to be a notable limitation of the assays used here.

We have previously studied wild-type MSH2 and 24 missense variants including both pathogenic and benign variants as well as variants with unknown pathogenicity [18]. Our results showed that most of the pathogenic variants were found at low steady-state protein levels because they were rapidly degraded by the proteasome. We also found that structure-based computational predictions of the change in thermodynamic stability could be used to predict

cellular abundance and thereby pathogenicity for many variants. Together with an assessment of sequence conservation, we suggested that most pathogenic missense variants in *MSH2* cause LS by a mechanism that involves loss of protein stability with resulting loss of abundance and function.

Indeed, 76% of the pathogenic variants that were scored as LoF by Jia and colleagues are predicted to be destabilized beyond the value of $\Delta\Delta G = 3.1$ kcal mol^{-1} (where $\Delta\Delta G$ refers to the change in protein stability) that we determined by comparing cellular protein levels to stability calculations using the FoldX software [18] (Fig 1B), and 63% of all variants with LoF scores greater than zero (the cutoff determined by Jia and colleagues) have $\Delta\Delta G > 3.1$ kcal mol^{-1}. This value can be compared to the just 14% of the functional variants (LoF score <0) that are predicted to have $\Delta\Delta G > 3.1$ kcal mol^{-1}. Thus, in line with our previous findings on a small subset of variants, it is likely that a dominant fraction of pathogenic variants has low protein abundance and that this might explain their pathogenicity. In the future, it will be interesting to examine further the 14% of the functional variants that are predicted to be destabilized. Some of these might be explained by inaccuracies in the stability predictions and by the complicated relationship between thermodynamic stability and cellular protein abundance that might mean that the same level of destabilization could result in different cellular abundancies [27]. We also note that there appears to be a continuous transition of LoF and loss of stability, so that while only 7% of the most functional variants (LoF score <-5) have $\Delta\Delta G > 3.1$ kcal mol^{-1}, this number increases to 37% among the variants that are regarded as functional but just below the cutoff (LoF score between -1 and 0) (Fig 1D). Indeed, the analysis suggests a gradual LoF beginning from LoF scores > -2, which also happens to be approximately the highest value observed in control synonymous variants [11].

Variants that are stable and abundant in the cell can still cause LoF, e.g., by removing key interactions in the binding sites for DNA, ATP, or other proteins. The effects of stability loss and other mechanisms for LoF can often be quantified by analyzing conservation patterns observed in multiple sequence alignments, for example, through a score termed $\Delta\Delta E$ where large values correspond to nonconservative substitutions. Almost all 6TG-resistant pathogenic variants with high LoF scores have high $\Delta\Delta E$ scores (Fig 1C), indicating that these substitutions are rare or absent among homologs of MSH2 [28,29]. These observations are in line with similar findings in other genes and diseases which show that joint analyses of protein stability effects and sequence conservation may be used both to predict which variants show LoF and to find those that do so due to loss of stability and resulting low protein abundance [30–32]. We also note that 53% of the variants with LoF score <0 and $\Delta\Delta G > 3.1$ kcal mol^{-1} also have high $\Delta\Delta E$ scores, lending further support to their importance to MSH2 function, possibly in aspects that were not captured in the screen, or because these variants show a mild LoF. As for the analysis of $\Delta\Delta G$ described above, we find that variants with high values of $\Delta\Delta E$ become enriched already at intermediate values of LoF score (Fig 1D).

The new results reported by Jia and colleagues and Ollodart and colleagues provide opportunities for future clinical applications. First, the experimental scores—in particular the comprehensive assessment by the 6TG assay—can be useful information for ascribing pathogenicity to new variants that may be discovered in the clinic and might also warrant reassessment of certain previous classifications. Thus, rather than to wait for the results of new functional assays, clinical geneticists may simply look up the variant effects in these experiments if they become validated for clinical use. Second, when more than one variant is present in a patient, it may be difficult to determine from the clinical data which variant(s) is causative, and the data generated by the MAVEs may help in such assignments. Third, data generated by MAVEs are extremely useful for benchmarking prediction methods [29–34], which may in turn be improved for use in other genes and diseases. Fourth, the data and complementary

computational analyses may be used to help pinpoint the mechanisms by which variants cause LoF, information that might be particularly relevant for developing future therapies. For example, experiments in yeast have shown that it may be possible to restore function of some MSH2 LoF variants that are unstable and degraded in the cell by disrupting the machinery that recognizes and targets the variants for degradation [35].

Acknowledgments

References

1. Dominguez-Valentin M, Sampson JR, Seppälä TT, Ten Broeke SW, Plazzer J-P, Nakken S, et al. Cancer risks by gene, age, and gender in 6350 carriers of pathogenic mismatch repair variants: findings from the Prospective Lynch Syndrome Database. Genet Med. 2020; 22:15–25. https://doi.org/10.1038/s41436-019-0596-9 PMID: 31337882

2. Lynch HT, Snyder CL, Shaw TG, Heinen CD, Hitchins MP. Milestones of Lynch syndrome: 1895–2015. Nat Rev Cancer. 2015; 15:181–94. https://doi.org/10.1038/nrc3878 PMID: 25673086

3. Ali H, Olatubosun A, Vihinen M. Classification of mismatch repair gene missense variants with PON-MMR. Hum Mutat. 2012; 33:642–50. https://doi.org/10.1002/humu.22038 PMID: 22290698

4. Brnich SE, Abou Tayoun AN, Couch FJ, Cutting GR, Greenblatt MS, Heinen CD, et al. Recommendations for application of the functional evidence PS3/BS3 criterion using the ACMG/AMP sequence variant interpretation framework. Genome Med. 2019; 12:3. https://doi.org/10.1186/s13073-019-0690-2 PMID: 31892348

5. Houlleberghs H, Dekker M, Lantermans H, Kleinendorst R, Dubbink HJ, Hofstra RMW, et al. Oligonucleotide-directed mutagenesis screen to identify pathogenic Lynch syndrome-associated MSH2 DNA mismatch repair gene variants. Proc Natl Acad Sci U S A. 2016; 113:4128–33. https://doi.org/10.1073/pnas.1520813113 PMID: 26951660

6. Tricarico R, Kasela M, Mareni C, Thompson BA, Drouet A, Staderini L, et al. Assessment of the InSiGHT Interpretation Criteria for the Clinical Classification of 24 MLH1 and MSH2 Gene Variants. Hum Mutat. 2017; 38:64–77. https://doi.org/10.1002/humu.23117 PMID: 27629256

7. Rasmussen LJ, Heinen CD, Royer-Pokora B, Drost M, Tavtigian S, Hofstra RMW, et al. Pathological assessment of mismatch repair gene variants in Lynch syndrome: past, present, and future. Hum Mutat. 2012; 33:1617–25. https://doi.org/10.1002/humu.22168 PMID: 22833534

8. Fowler DM, Fields S. Deep mutational scanning: a new style of protein science. Nat Methods. 2014; 11:801–7. https://doi.org/10.1038/nmeth.3027 PMID: 25075907

9. Starita LM, Ahituv N, Dunham MJ, Kitzman JO, Roth FP, Seelig G, et al. Variant Interpretation: Functional Assays to the Rescue. Am J Hum Genet. 2017; 101:315–25. https://doi.org/10.1016/j.ajhg.2017.07.014 PMID: 28886340

10. Ollodart AR, Yeh C-LC, Miller AW, Shirts BH, Gordon AS, Dunham MJ. Multiplexing Mutation Rate Assessment: Determining Pathogenicity of Msh2 Variants in S. cerevisiae. Genetics (In press)

11. Jia X, Burugula BB, Chen V, Lemons RM, Jayakody S, Maksutova M, et al. Massively parallel functional testing of MSH2 missense variants conferring Lynch syndrome risk. Am J Hum Genet. 2020. https://doi.org/10.1016/j.ajhg.2020.12.003 PMID: 33357406

12. Brown KD, Rathi A, Kamath R, Beardsley DI, Zhan Q, Mannino JL, et al. The mismatch repair system is required for S-phase checkpoint activation. Nat Genet. 2003; 33:80–4. https://doi.org/10.1038/ng1052 PMID: 12447371

13. Paquin CE, Adams J. Relative fitness can decrease in evolving asexual populations of S. cerevisiae. Nature. 1983; 306:368–70. https://doi.org/10.1038/306368a0 PMID: 16752492

14. Lang GI, Murray AW. Estimating the per-base-pair mutation rate in the yeast Saccharomyces cerevisiae. Genetics. 2008; 178:67–82. https://doi.org/10.1534/genetics.107.071506 PMID: 18202359

15. Gammie AE, Erdeniz N, Beaver J, Devlin B, Nanji A, Rose MD. Functional characterization of pathogenic human MSH2 missense mutations in Saccharomyces cerevisiae. Genetics. 2007; 177:707–21. https://doi.org/10.1534/genetics.107.071084 PMID: 17720936

16. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020; 581:434–43. https://doi.org/10.1038/s41586-020-2308-7 PMID: 32461654

17. Jaganathan K, Kyriazopoulou Panagiotopoulou S, McRae JF, Darbandi SF, Knowles D, Li YI, et al. Predicting Splicing from Primary Sequence with Deep Learning. Cell. 2019; 176:535–548.e24. https://doi.org/10.1016/j.cell.2018.12.015 PMID: 30661751

18. Nielsen SV, Stein A, Dinitzen AB, Papaleo E, Tatham MH, Poulsen EG, et al. Predicting the impact of Lynch syndrome-causing missense mutations from structural calculations. PLoS Genet. 2017; 13: e1006739. https://doi.org/10.1371/journal.pgen.1006739 PMID: 28422960

19. Warren JJ, Pohlhaus TJ, Changela A, Iyer RR, Modrich PL, Beese LS. Structure of the human MutSalpha DNA lesion recognition complex. Mol Cell. 2007; 26:579–92. https://doi.org/10.1016/j.molcel.2007.04.018 PMID: 17531815

20. Guerois R, Nielsen JE, Serrano L. Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. J Mol Biol. 2002; 320:369–87. https://doi.org/10.1016/S0022-2836(02)00442-4 PMID: 12079393

21. Balakrishnan S, Kamisetty H, Carbonell JG, Lee S-I, Langmead CJ. Learning generative models for protein fold families. Proteins. 2011; 79:1061–78. https://doi.org/10.1002/prot.22934 PMID: 21268112

22. Abildgaard AB, Stein A, Nielsen SV, Schultz-Knudsen K, Papaleo E, Shrikhande A, et al. Computational and cellular studies reveal structural destabilization and degradation of MLH1 variants in Lynch syndrome. elife. 2019; 8. https://doi.org/10.7554/eLife.49138 PMID: 31697235

23. Martinez SL, Kolodner RD. Functional analysis of human mismatch repair gene mutations identifies weak alleles and polymorphisms capable of polygenic interactions. Proc Natl Acad Sci U S A. 2010; 107:5070–5. https://doi.org/10.1073/pnas.1000798107 PMID: 20176959

24. Bouvet D, Bodo S, Munier A, Guillerm E, Bertrand R, Colas C, et al. Methylation Tolerance-Based Functional Assay to Assess Variants of Unknown Significance in the MLH1 and MSH2 Genes and Identify Patients With Lynch Syndrome. Gastroenterology. 2019; 157:421–31. https://doi.org/10.1053/j.gastro.2019.03.071 PMID: 30998989

25. Drost M, Lützen A, van Hees S, Ferreira D, Calléja F, Zonneveld JBM, et al. Genetic screens to identify pathogenic gene variants in the common cancer predisposition Lynch syndrome. Proc Natl Acad Sci U S A. 2013; 110:9403–8. https://doi.org/10.1073/pnas.1220537110 PMID: 23690608

26. Bapat BV, Madlensky L, Temple LK, Hiruki T, Redston M, Baron DL, et al. Family history characteristics, tumor microsatellite instability and germline MSH2 and MLH1 mutations in hereditary colorectal cancer. Hum Genet. 1999; 104:167–76. https://doi.org/10.1007/s004390050931 PMID: 10190329

27. Stein A, Fowler DM, Hartmann-Petersen R, Biophysical L-LK. Mechanistic Models for Disease-Causing Protein Variants. Trends Biochem Sci. 2019; 44:575–88. https://doi.org/10.1016/j.tibs.2019.01.003 PMID: 30712981

28. Feinauer C, Weigt M. Context-Aware Prediction of Pathogenicity of Missense Mutations Involved in Human Disease. 2017. https://doi.org/10.1101/103051

29. Hopf TA, Ingraham JB, Poelwijk FJ, Schärfe CPI, Springer M, Sander C, et al. Mutation effects predicted from sequence co-variation. Nat Biotechnol. 2017; 35:128–35. https://doi.org/10.1038/nbt.3769 PMID: 28092658

30. Jepsen MM, Fowler DM, Hartmann-Petersen R, Stein A, Lindorff-Larsen K. Chapter 5—Classifying disease-associated variants using measures of protein activity and stability. In: Pey AL, editor. Protein Homeostasis Diseases. Academic Press; 2020. pp. 91–107.

31. Cagiada M, Johansson KE, Valančiūtė A, Nielsen SV, Hartmann-Petersen R, Yang JJ, et al. Understanding the origins of loss of protein function by analyzing the effects of thousands of variants on activity and abundance. bioRxiv. 2020. p. 2020.09.28.317040. https://doi.org/10.1101/2020.09.28.317040

32. Chiasson MA, Rollins NJ, Stephany JJ, Sitko KA, Matreyek KA, Verby M, et al. Multiplexed measurement of variant abundance and activity reveals VKOR topology, active site and human variant impact. elife. 2020; 9. https://doi.org/10.7554/eLife.58026 PMID: 32870157

33. Frazer J, Notin P, Dias M, Gomez A, Brock K, Gal Y, et al. Large-scale clinical interpretation of genetic variants using evolutionary data and deep learning. bioRxiv. 2020. p. 2020.12.21.423785. https://doi.org/10.1101/2020.12.21.423785

34. Livesey BJ, Marsh JA. Using deep mutational scanning to benchmark variant effect predictors and identify disease mutations. Mol Syst Biol. 2020; 16:e9380. https://doi.org/10.15252/msb.20199380 PMID: 32627955

35. Arlow T, Scott K, Wagenseller A, Gammie A. Proteasome inhibition rescues clinically significant unstable variants of the mismatch repair protein Msh2. Proc Natl Acad Sci U S A. 2013:246–51. https://doi.org/10.1073/pnas.1215510110 PMID: 23248292