Human Genetics (2022) 141:1545–1548
https://doi.org/10.1007/s00439-022-02483-4

EDITORIAL

Computational interpretation of human genetic variation

Yana Bromberg[1] · **Predrag Radivojac**[2]

Published online: 23 September 2022

Computational interpretation of human genetic variants comprises the development and application of analysis and prediction techniques aimed at elucidating the impact of variants in an individual's genome on different organismal, cellular, and molecular phenotypes. A distinguishing characteristic of this interdisciplinary field is a remarkable breadth of the phenotypes of interest and their genetic architectures, biological and environmental contexts, data modalities and data generating platforms, as well as computational techniques developed to make sense of all available data. Even further contributing to the complexity of the field are the issues of safe deployment of the newly developed tools and mitigation of ethical challenges necessary for societal acceptance of both the research process and clinical application (Szabo 2019; McInnes et al. 2021).

In the clinic, computational tools often incorporate patient, and sometimes family, genome variation data to identify individual or groups of variants with diagnostic (i.e., variants that cause the observed symptoms), predictive (i.e., variants that stratify individuals based on the risk of disease development), or pharmacogenomic (i.e., variants useful for tailoring the treatment) relevance (Rehm et al. 2015; Adhikari et al. 2020; McInnes et al. 2021). And while a useful computational tool or pipeline need not precisely describe disease mechanisms in its predictions, it benefits from doing so (Rost et al. 2016) and is expected to work across diverse human populations or explicitly state such limitations (Popejoy and Fullerton 2016; Schwartz et al. 2021). Beyond the clinic, these tools could further contribute to the mechanistic understanding of non-disease phenotypes, evolutionary processes, and would ideally be also useful to healthy individuals; e.g., in tailoring diets that will lead to longer and more fulfilling lives.

Existing computational analysis techniques and predictive models draw from different research communities. Relevant methods range from biophysics techniques and molecular dynamics to systems modeling to modern machine learning (Hu et al. 2019). Recent emergence of deep learning has spurred the latest wave of powerful and data-hungry approaches capable of learning arbitrary concepts given enough data and compute power (LeCun et al. 2015). However, biological and clinical data are complex, often defying modeling assumptions and thus requiring sophisticated and careful integration across different modalities (Žitnik et al. 2019), from -omics data to language-based records to images. Finally, data protections due to privacy concerns and competitive or international restrictions bring further difficulties and promote the need for distributed learning, learning from encrypted data, and model-to-data approaches (Wang et al. 2017; Yan et al. 2021).

Evaluation of method performance presents additional challenges (The Critical Assessment of Genome Interpretation Consortium 2022). Clinicians are generally interested in knowing whether a variant is pathogenic to potentially recommend treatment and surveillance strategies (Rehm et al. 2015) and, consequently, cannot tolerate large fraction of false positive predictions (Pejaver et al. 2022). On the other hand, biomedical researchers may be more interested in prioritizing variants for experimental studies and so a method's ability to identify significant variants and its false discovery rate are differently balanced. The complexity of the field and the differences in expertise, combined with method availability, are also prone to encourage evaluation that leads to method misuse. This is common even within the research community; e.g., predictors aiming to identify variants with structural (Schymkowitz et al. 2005; Capriotti et al. 2005; Li et al. 2021) or functional (Bromberg and Rost 2007) effect can be suboptimally used as pathogenicity predictors.

In this special issue of *Human Genetics*, we present 11 studies that advance the field of computational interpretation

✉ Yana Bromberg
yana@bromberglab.org

✉ Predrag Radivojac
predrag@northeastern.edu

[1] Department of Biochemistry and Microbiology, Rutgers University, New Brunswick, NJ 08901, USA

[2] Khoury College of Computer Sciences, Northeastern University, Boston, MA 02115, USA

of human genetic variation. We encouraged submissions that would review various aspects of the field, offer perspectives on the current state of the art, include original method descriptions, or perform comprehensive method evaluation. Human variation and its downstream effects were considered in the broad sense, including germline and somatic, coding and noncoding, and single and multi-nucleotide variants. Effects on structure and function of single biological macromolecules (DNA, RNA, protein) as well as whole-genome structure and function were of interest. At the same time, methods that address broad disease areas such as cancer, neurodevelopmental disorders, or those that consider disease through the lens of evolution, were invited. We received 25 submissions of which 11 were accepted for publication, and are briefly summarized below.

In a thorough review of the state of the art of coding variant interpretation, Katsonis et al. (2022) highlight the importance of computational methods for exploration of disease, if not patient, specifics. The authors find that available methods are non-generically applicable and should be used in annotation of appropriate cohorts, with a full understanding of their limitations. When applied in an informed fashion and evaluated rigorously (Pejaver et al. 2022; The Critical Assessment of Genome Interpretation Consortium 2022), the predictors can be considered essential tools for genome interpretation.

Sarquella-Brugada et al. (2021) investigate variant reclassification (Harrison and Rehm 2019) using a patient cohort with inherited cardiac channelopathies. Their study suggests that a timely genome reanalysis—at no more than five-year intervals—may help ensure more accurate variant annotation. Curiously, in their re-analysis of a cohort of previously treated patients, they find that most of the reclassification is due to the changes in reported variant frequencies over time and further identify a significant contribution from the less studied channelopathy-associated genes.

Sun et al. (2022a) describe a collection of modifier genes and variants that, through epistasis, significantly contribute to, rather than directly cause, observed phenotypes. They identify distinct properties of modifier variants (e.g., higher allele frequency) and present observations that are in line with both the differential functionalities of individual protein residues and with the polygenic nature of many disorders. The authors' findings highlight the need to further explore causality pathways to enable high-resolution diagnostics and develop novel treatments. This work also opens new avenues for the studies of variable expressivity, penetrance, and other epistatic outcomes.

While multiple tools exist for the analysis of variant impact on protein function, it is worth re-emphasizing that proteins carry out their function within molecular pathways. The work by Ozturk and Carter (2022) investigates protein interaction interfaces and suggests that network-based assessments of variant effect carry previously uncaptured signal. They find that the effects of somatic variants are better described via this approach than germline effects, suggesting different modes of variant activity and constraints.

Jiang et al. (2021) integrate spatio-temporal gene expression data and protein–protein interactions with variant data to prioritize de novo exonic variants from patients with autism spectrum disorder. The data were integrated based on the principles of positive-unlabeled learning that enable score calibration based on the probability of a gene to be involved in autism as well as the probability of a variant to disrupt gene function. They demonstrate that such an approach discriminates among de novo variants between cases and controls and experimentally validate disruption of protein–protein interactions in their top-scoring missense mutation located in the *ATP1A3* gene.

Similarly focused on a single phenotype, Sun et al. (2022b) demonstrate that assessing coding variants in the mother's genome for their impact on the corresponding protein functionality is sufficiently informative of the likelihood of embryonic aneuploidy. This work is in line with similar findings of machine learning applied to exonic variants to summarize traits or pathogenesis mechanisms in Crohn's disease (Wang et al. 2019) as well as *A. thaliana* phenotypes (Raimondi et al. 2022).

Recent breakthroughs in biological sequence and structure analysis have been facilitated by the use of deep learning techniques (Jumper et al. 2021). In their work for this special issue, Marquet et al. (2021) demonstrate that high-dimensional embeddings of protein sequences, combined with generalized amino acid substitution scores such as BLOSUM62 (Henikoff and Henikoff 1992), are able to predict effects of amino acid substitutions on protein function as well as methods that use multiple sequence alignments; i.e., evolutionary conservation.

This finding is particularly relevant in light of the work by Capriotti and Fariselli (2022), which asks an important question about the value of conservation in describing variant impact. In comparison with a range of existing computational methods that use a variety of input features, the authors find that sequence evolutionary parameters are necessary and, alone, nearly fully sufficient to accurately identify pathogenic variants. These two studies combine to suggest that sequence alone, or together with generic conservation scores, may be sufficient to identify functional and pathogenic protein variants.

In a focus away from proteins, Waldern et al. (2021) review the molecular mechanisms of disease at the transcriptome level. The authors confirm the importance of post-transcriptional regulatory motifs, but also note that many disease-causing variants fall outside and even far away from well-defined regulatory motifs. They review evidence that these outside regions alter function via mRNA structure

disruption and highlight the examples of riboSNitches, disease-causing single nucleotide variants that affect regulation, dysregulating binding affinity to RNA-binding proteins and micro RNA (miRNA) as well as splice site accessibility, via changes to mRNA structure.

Villegas-Miron et al. (2022) study human miRNA variants and their ability to differentiate human populations. The authors show that these variants affect gene expression and that the excessive complexity of miRNA regulatory networks makes its variants culprits in changes of multiple pathways simultaneously. Importantly, these subsequently tend to be linked to a broad range of pathogenic processes.

Finally, Xia and Yanai (2022) study gene expression patterns during spermatogenesis to investigate the possibility of expression-based control of germline mutation rates via an interplay of transcription-coupled DNA repair (promoting lower mutation rate with increased gene expression) and transcription-coupled damage (promoting higher mutation rate with increased gene expression). They present results in support of the transcriptional scanning hypothesis, a process leading to mutation rate reduction through transcription-coupled DNA repair, as a more influential mechanism. Their results also contribute to a mechanistic explanation of higher rates of somatic variants in low-expression genes (Lawrence et al. 2013), thus adding another dimension to our understanding of mutation rates (Eyre-Walker and Keightley 2007).

Acknowledgements Profs. Bromberg and Radivojac offer their gratitude to the *Human Genetics* journal and numerous reviewers who contributed to the selection of the high-quality articles.

Funding This work was supported in part by the National Institutes of Health awards GM115486 (YB), HG012022 (PR), and HD101246 (PR).

Declarations

Conflict of interest The authors have not disclosed any competing interests.

References

Adhikari AN, Gallagher RC, Wang Y, Currier RJ, Amatuni G, Bassaganyas L, Chen F, Kundu K, Kvale M, Mooney SD, Nussbaum RL, Randi SS, Sanford J, Shieh JT, Srinivasan R, Sunderam U, Tang H, Vaka D, Zou Y, Koenig BA, Kwok PY, Risch N, Puck JM, Brenner SE (2020) The role of exome sequencing in newborn screening for inborn errors of metabolism. Nat Med 26(9):1392–1397

Bromberg Y, Rost B (2007) SNAP: predict effect of non-synonymous polymorphisms on function. Nucleic Acids Res 35(11):3823–3835

Capriotti E, Fariselli P, Casadio R (2005) I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. Nucleic Acids Res 33(Web Server issue):W306–W310

Capriotti E, Fariselli P (2022) Evaluating the relevance of sequence conservation in the prediction of pathogenic missense variants. Hum Genet 141(10)

Eyre-Walker A, Keightley PD (2007) The distribution of fitness effects of new mutations. Nat Rev Genet 8(8):610–618

Harrison SM, Rehm HL (2019) Is 'likely pathogenic' really 90% likely? Reclassification data in ClinVar. Genome Med 11(1):72

Henikoff S, Henikoff JG (1992) Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci USA 89(22):10915–10919

Hu Z, Yu C, Furutsuki M, Andreoletti G, Ly M, Hoskins R, Adhikari AN, Brenner SE (2019) VIPdb, a genetic variant impact predictor database. Hum Mutat 40(9):1202–1214

Jiang Y, Urresti J, Pagel KA, Pramod AB, Iakoucheva LM, Radivojac P (2022) Prioritizing de novo autism risk variants with calibrated gene- and variant-scoring models. Hum Genet 141(10)

Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Zidek A, Potapenko A, Bridgland A, Meyer C, Kohl SAA, Ballard AJ, Cowie A, Romera-Paredes B, Nikolov S, Jain R, Adler J, Back T, Petersen S, Reiman D, Clancy E, Zielinski M, Steinegger M, Pacholska M, Berghammer T, Bodenstein S, Silver D, Vinyals O, Senior AW, Kavukcuoglu K, Kohli P, Hassabis D (2021) Highly accurate protein structure prediction with AlphaFold. Nature 596(7873):583–589

Katsonis P, Wilhelm K, Williams A, Lichtarge O (2022) Genome interpretation using in silico predictors of variant impact. Hum Genet 141(10)

Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, Carter SL, Stewart C, Mermel CH, Roberts SA, Kiezun A, Hammerman PS, McKenna A, Drier Y, Zou L, Ramos AH, Pugh TJ, Stransky N, Helman E, Kim J, Sougnez C, Ambrogio L, Nickerson E, Shefler E, Cortes ML, Auclair D, Saksena G, Voet D, Noble M, DiCara D, Lin P, Lichtenstein L, Heiman DI, Fennell T, Imielinski M, Hernandez B, Hodis E, Baca S, Dulak AM, Lohr J, Landau DA, Wu CJ, Melendez-Zajgla J, Hidalgo-Miranda A, Koren A, McCarroll SA, Mora J, Crompton B, Onofrio R, Parkin M, Winckler W, Ardlie K, Gabriel SB, Roberts CWM, Biegel JA, Stegmaier K, Bass AJ, Garraway LA, Meyerson M, Golub TR, Gordenin DA, Sunyaev S, Lander ES, Getz G (2013) Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature 499(7457):214–218

LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521(7553):436–444

Li G, Panday SK, Alexov E (2021) SAAFEC-SEQ: a sequence-based method for predicting the effect of single point mutations on protein thermodynamic stability. Int J Mol Sci 22(2):606

Marquet C, Heinzinger M, Olenyi T, Dallago C, Erckert K, Bernhofer M, Nechaev D, Rost B (2022) Embeddings from protein language models predict conservation and variant effects. Hum Genet 141(10)

McInnes G, Sharo AG, Koleske ML, Brown JEH, Norstad M, Adhikari AN, Wang S, Brenner SE, Halpern J, Koenig BA, Magnus DC, Gallagher RC, Giacomini KM, Altman RB (2021) Opportunities and challenges for the computational interpretation of rare variation in clinically important genes. Am J Hum Genet 108(4):535–548

Ozturk K, Carter H (2022) Predicting functional consequences of mutations using molecular interaction network features. Hum Genet 141(6):1195–1210

Pejaver V, Byrne AB, Feng BJ, Pagel KA, Mooney SD, Karchin R, O'Donnell-Luria A, Harrison SM, Tavtigian SV, Greenblatt MS, Biesecker LG, Radivojac P, Brenner SE, ClinGen Sequence Variant Interpretation Working Group (2022). Evidence-based calibration of computational tools for missense variant pathogenicity

classification and ClinGen recommendations for clinical use of PP3/BP4 criteria. bioRxiv 2022.03.17.484479

Popejoy AB, Fullerton SM (2016) Genomics is failing on diversity. Nature 538(7624):161–164

Raimondi D, Corso M, Fariselli P, Moreau Y (2022) From genotype to phenotype in *Arabidopsis thaliana*: in-silico genome interpretation predicts 288 phenotypes from sequencing data. Nucleic Acids Res 50(3):e16

Rehm H. L, Berg J. S, Brooks L. D., Bustamante C. D., Evans J. P., Landrum M. J., Ledbetter D. H., Maglott D. R., Martin C. L., Nussbaum R. L., Plon S. E, Ramos E. M., Sherry S. T., Watson M. S, ClinGen (2015) ClinGen—the clinical genome resource. N Engl J Med 372(23):2235–2242

Rost B, Radivojac P, Bromberg Y (2016) Protein function in precision medicine: deep understanding with machine learning. FEBS Lett 590(15):2327–2341

Sarquella-Brugada G, Fernandez-Falgueras A, Cesar S, Arbelo E, Coll M, Perez-Serra A, Puigmule M, Iglesias A, Alcalde M, Vallverdu-Prats M, Fiol V, Ferrer-Costa C, Del Olmo B, Pico F, Lopez L, Garcia-Alvarez A, Jorda P, Tiron de Llano C, Toro R, Grassi S, Oliva A, Brugada J, Brugada R, Campuzano O (2022) Clinical impact of rare variants associated with inherited channelopathies: a 5-year update. Hum Genet 141(10)

Schwartz R, Down L, Jonas A, Tabassi E (2021) A proposal for identifying and managing bias in artificial intelligence. Draft NIST Special Publication 1270

Schymkowitz J, Borg J, Stricher F, Nys R, Rousseau F, Serrano L (2005) The FoldX web server: an online force field. Nucleic Acids Res 33(Web Server issue):W382–W388

Sun H, Lan X, Ma L, Zhou J (2022a) Revealing modifier variations characterizations for elucidating the genetic basis of human phenotypic variations. Hum Genet 141(6):1223–1233

Sun S, Miller M, Wang Y, Tyc KM, Cao X, Scott Jr, RT, Tao X, Bromberg Y, Schindler K, Xing J (2022b) Predicting embryonic aneuploidy rate in IVF patients using whole-exome sequencing. Hum Genet 141(10)

Szabo L (2019) Artificial intelligence is rushing into patient care—and could raise risks. Sci Am 12(24). https://www.scientificamerican.com/article/artificial-intelligence-is-rushing-into-patient-care-and-could-raise-risks/

The Critical Assessment of Genome Interpretation Consortium (2022) CAGI, the Critical Assessment of Genome Interpretation, establishes progress and prospects for computational genetic variant interpretation methods. arXiv preprint arXiv:2205.05897

Yan Y, Schaffter T, Bergquist T, Yu T, Prosser J, Aydin Z, Jabeer A, Brugere I, Gao J, Chen G, Causey J, Yao Y, Bryson K, Long DR, Jarvik JG, Lee CI, Wilcox A, Guinney J, Mooney S, The DREAM Challenge Consortium (2021) A continuously benchmarked and crowdsourced challenge for rapid development and evaluation of models to predict COVID-19 diagnosis and hospitalization. JAMA Netw Open 4(10):e2124946

Villegas-Miron P, Gallego A, Bertranpetit J, Laayouni H, Espinosa-Parrilla Y (2022) Signatures of genetic variation in human microRNAs point to processes of positive selection and population-specific disease risks. Hum Genet 141(10)

Waldern JM, Kumar J, Laederach A (2022) Disease-associated human genetic variation through the lens of precursor and mature RNA structure. Hum Genet 141(10)

Wang S, Jiang X, Singh S, Marmor R, Bonomi L, Fox D, Dow M, Ohno-Machado L (2017) Genome privacy: challenges, technical approaches to mitigate risk, and ethical considerations in the United States. Ann N Y Acad Sci 1387(1):73–83

Wang Y, Miller M, Astrakhan Y, Petersen BS, Schreiber S, Franke A, Bromberg Y (2019) Identifying Crohn's disease signal from variome analysis. Genome Med 11(1):59

Xia B, Yanai I (2022) Gene expression levels modulate germline mutation rates through the compound effects of transcription-coupled repair and damage. Hum Genet 141(6):1211–1222

Žitnik M, Nguyen F, Wang B, Leskovec J, Goldenberg A, Hoffman MM (2019) Machine learning for integrating data in biology and medicine: principles, practice, and opportunities. Inf Fusion 50:71–91