



OPEN ACCESS

EDITED AND REVIEWED BY
Joao Carlos Setubal,
University of São Paulo, Brazil

*CORRESPONDENCE
Yves Moreau,
yves.moreau@kuleuven.be

[†]These authors have contributed equally
to this work

SPECIALTY SECTION
This article was submitted
to Genomic Analysis,
a section of the journal
Frontiers in Bioinformatics

RECEIVED 15 November 2022
ACCEPTED 17 November 2022
PUBLISHED 30 November 2022

CITATION
Raimondi D, Orlando G, Verplaetse N,
Fariselli P and Moreau Y (2022), Editorial:
Towards genome interpretation:
Computational methods to model the
genotype-phenotype relationship.
Front. Bioinform. 2:1098941.
doi: 10.3389/fbinf.2022.1098941

COPYRIGHT
© 2022 Raimondi, Orlando, Verplaetse,
Fariselli and Moreau. This is an open-
access article distributed under the
terms of the [Creative Commons
Attribution License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use,
distribution or reproduction in other
forums is permitted, provided the
original author(s) and the copyright
owner(s) are credited and that the
original publication in this journal is
cited, in accordance with accepted
academic practice. No use, distribution
or reproduction is permitted which does
not comply with these terms.

Editorial: Towards genome interpretation: Computational methods to model the genotype-phenotype relationship

Daniele Raimondi^{1†}, Gabriele Orlando^{2†}, Nora Verplaetse¹,
Piero Fariselli³ and Yves Moreau^{1*}

¹ESAT-STADIUS, KU Leuven Leuven, Belgium, ²Switch Laboratory, KU Leuven Leuven, Belgium,
³Department of Medical Sciences, University of Torino, Torino, Italy

Editorial on the Research Topic

[Towards genome interpretation: Computational methods to model the genotype-phenotype relationship](#)

Genome Interpretation (GI) is an umbrella term for the scientific efforts oriented towards modelling and understanding the relationship between genotype and phenotype in living organisms (Daneshjou et al., 2017; Andreoletti et al., 2019; Raimondi et al., 2022a). Even temporarily setting epigenetic and environmental effects aside, untangling the complex relation between the complete set of genetic material of an individual organism (be it a human, other animals, plants, or microorganisms) and its observed phenotypes is an extremely ambitious and challenging endeavor, in particular for non-Mendelian traits. Being able to reliably model this genotype-phenotype relationship could revolutionize many aspects of genetics, biology, and medicine (Daneshjou et al., 2017; Fröhlich et al., 2018). For example, it could warn us about late-onset genetic disorders, helping their prevention (Weedon et al., 2006; Morrison et al., 2007). It could also lead to the design of medications and treatments tailored to each patient's genome, complementing environmental and medical-history data to improve patient prognosis (Fröhlich et al., 2018). Applied to cancer, it could bring a novel understanding of cancer development, helping devise highly specific cocktails of drugs and discover novel molecules to target each unique tumor (Li et al., 2019). Such personalized approaches to medicine, called Precision Medicine, are still largely out of our reach in many clinical settings (Daneshjou et al., 2017; Fröhlich et al., 2018).

In the last decade, the avalanche of scientific results brought by Next Generation Sequencing (NGS) and big data technologies seemed almost unstoppable, and at times it seemed that finally *cracking* the genotype-phenotype problem was within reach. Ten years later, notwithstanding the vast amounts of data collected and numerous advances in genetics (Moreau and Tranchevent, 2012; Boycott et al., 2013; Erwin et al., 2014; Goodwin et al., 2016), including the discovery of the causative variants for many Mendelian

disorders (Bamshad et al., 2011), our genome is still hiding most of its secrets. When it comes to oligogenic and polygenic diseases (i.e., diseases involving respectively few and many genes (Gazzo et al., 2017)), the bottleneck has indeed mostly just shifted from a problem of data availability to one of data interpretation, since the classical approaches used in genetics have shown important shortcomings in uncovering complex disease mechanisms (Manolio et al., 2009; Gibson, 2012; Francisco and Bustamante, 2018; Wald and Robert, 2019).

The advent of NGS technologies was nonetheless invaluable, since they almost brought us at the doorstep of a new era where the scarcity of genomics data will be less and less of a bottleneck. This will make the application of *data hungry* cutting-edge Machine Learning (ML) and Deep Learning (DL) methods to this endeavor finally possible, eventually reproducing the astounding successes that methods such AlphaFold (Jumper et al., 2021; Chowdhury et al., 2022) obtained in structural biology in the realm of Genome Interpretation.

However, data abundance alone will not do the trick, for such a complex problem. The actual implementation of ML/DL methods for GI requires the development of tailor-made algorithms that can deal with the unique issues presented by genomic and phenomics (Houle et al., 2010) data. For example, Whole Exome or Genome Sequencing samples (WES, WGS) can be extremely large, sparse, and noisy (Ng et al., 2008). Moreover, they also pose privacy and ethical issues in their management, storage, and processing (Rieke et al., 2020). Finally, to apply GI to Precision Medicine, models must ensure *accountability* of their predictions, for example by providing means for their interpretability and explainability, following the Explainable AI (XAI) paradigm (Bach et al., 2015; Smilkov et al., 2017; Lapuschkin et al., 2019; Raimondi et al., 2020a).

In the last decade, the bioinformatics community has addressed various specific aspects related to the Genome Interpretation (GI) problem, developing variant-effect predictors (Kircher et al., 2014; Dong et al., 2015; Ioannidis et al., 2016; Jagadeesh et al., 2016; Niroula and Vihinen, 2016; Raimondi et al., 2016; Raimondi et al., 2017), variant-prioritization (Sifrim et al., 2013; Wu et al., 2014; Cipriani et al., 2020) and gene-prioritization tools (Aerts et al., 2006; Guala and Sonnhammer, 2017), also trying to model digenic disease (Gazzo et al., 2017; Papadimitriou et al., 2019) or the protein-level molecular phenotype caused by a variant (Dehouck et al., 2011; Pucci et al., 2020; Raimondi et al., 2022b). Other widespread approaches in this sense include Genome Wide Association Studies (GWAS) (Uffelmann et al., 2021) and Polygenic Risk Scores (PRS) (Wei et al., 2013; Ali et al., 2018; Ala-Korpela and Holmes, 2020; Badré et al., 2021). In the context of plant and animal sciences, genetic marker-based methods for the Genomic Prediction for plants and animal breeding (e.g., BLUP) have been widely used (Daetwyler et al., 2013; Hickey et al., 2017; Wray et al., 2019; Maldonado et al., 2020).

These methods are the most relevant examples of how GI has been tackled so far. Few of them aim at directly modeling the genotype-phenotype relationship, while most focus instead on simpler subproblems, such as predicting the neutral/deleterious

effect of variants or just finding associations between phenotypes and genomic regions.

The growing availability of genomics data will soon enable the application of the latest ML/DL algorithms to GI, attempting to directly model the phenotypes produced by a given genome or exome, following a “genomes in/phenotypes out” paradigm (Raimondi et al., 2020b; Raimondi et al., 2022a). Early examples of such an approach, although on limited data, are methods for the case-controls discrimination of Crohn’s Disease (Wang et al., 2019; Raimondi et al., 2020b), Bipolar Disorder (Lakshman et al., 2017), the multi-phenotypic prediction of *A. thaliana* (Raimondi et al., 2022a) and yeast quantitative traits (Grinberg et al., 2020). We can imagine these methods as framed within a *spectrum of complexity*: at the *narrow end* of the spectrum we have methods aiming at the binary prediction or regression of the presence/absence of a certain phenotype (e.g., in cases/control studies) (Pal et al., 2017; Raimondi et al., 2020b), while at the *broad end* of the spectrum we have methods that perform a multiphenotypic prediction given a certain type of genotype measurement (e.g., WES, WGS or SNP array data) (Grinberg et al., 2020; Raimondi et al., 2022a).

In this Research Topic we collect papers that develop computational and ML methods addressing the challenges posed by this new paradigm of ML-based GI. These studies range from the application of GI to prokaryotes, with methods for the identification of putative cellulolytic anaerobes and for the identification of microsatellites that could act as biomarkers to differentiate *C. pseudotuberculosis* genomes, to yeast, with a Sparse Bayesian method for the prediction of *S. cerevisiae* growth in 46 different environmental conditions. Finally, regarding the development of strategies to apply DL methods to GI in the future, we propose a study investigating the possibility of encoding human genotype data as images, thus making it suitable for the application of DL techniques such as Convolutional Neural Networks for case/control classification.

While this Research Topic is by no means conclusive for a complex and long-term problem such as GI, we hope it can help focus more debate and research efforts on this *flavor* of ML/DL based GI, paving the way for full-fledged applications of this paradigm once large-scale genomic and phenomic data become widely available.

Author contributions

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated

organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

- Aerts, Stein, Lambrechts, Diether, Maity, Sunit, Van Loo, Peter, Coessens, Bert, De Smet, Frederik, et al. (2006). Gene prioritization through genomic data fusion. *Nat. Biotechnol.* 24 (5), 537–544. doi:10.1038/nbt1203
- Ala-Korpela, Mika, and Holmes, Michael V. (2020). Polygenic risk scores and the prediction of common diseases. *Int. J. Epidemiol.* 49 (1), 1–3. doi:10.1093/ije/dyz254
- Ali, Torkamani, Wineinger, Nathan E., and Eric, J. Topol (2018). The personal and clinical utility of polygenic risk scores. *Nat. Rev. Genet.* 19 (9), 581–590. doi:10.1038/s41576-018-0018-x
- Andreoletti, Gaia, Pal, Lipika R., Moul, John, and Brenner, Steven E. (2019). Reports from the fifth edition of cagi: The critical assessment of genome interpretation. *Hum. Mutat.* 40 (9), 1197–1201. doi:10.1002/humu.23876
- Bach, Sebastian, Binder, Alexander, Montavon, Grégoire, Klauschen, Frederick, Müller, Klaus-Robert, and Samek, Wojciech (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS one* 10 (7), e0130140. doi:10.1371/journal.pone.0130140
- Badré, Adrien, Zhang, Li, Muchero, Wellington, Reynolds, Justin C., and Pan, Chongle (2021). Deep neural network improves the estimation of polygenic risk scores for breast cancer. *J. Hum. Genet.* 66 (4), 359–369. doi:10.1038/s10038-020-00832-7
- Bamshad, Michael J., Ng, Sarah B., Bigham, Abigail W., Tabor, Holly K., Emond, Mary J., Nickerson, Deborah A., et al. (2011). Exome sequencing as a tool for mendelian disease gene discovery. *Nat. Rev. Genet.* 12 (11), 745–755. doi:10.1038/nrg3031
- Boycott, Kym M., Vanstone, Megan R., E Bulman, Dennis, and MacKenzie, Alex E. (2013). Rare-disease genetics in the era of next-generation sequencing: Discovery to translation. *Nat. Rev. Genet.* 14 (10), 681–691. doi:10.1038/nrg3555
- Chowdhury, Ratul, Bouatta, Nazim, Biswas, Surojit, Floristean, Christina, Kharkare, Anant, Roye, Koushik, et al. (2022). Single-sequence protein structure prediction using a language model and deep learning. *Nat. Biotechnol.* 40, 1617–1623. doi:10.1038/s41587-022-01432-w
- Cipriani, V., Pontikos, N., Gavin, A., Sergouniotis, P. I., Lenassi, E., Thawong, P., et al. (2020). An improved phenotype-driven tool for rare mendelian variant prioritization: Benchmarking exomiser on real patient whole-exome data. *Genes* 11 (4), 460. doi:10.3390/genes11040460
- Daetwyler, Hans D., Calus, Mario P. L., Pong-Wong, Ricardo, Campos, Gustavo de Los, and Hickey, John M. (2013). Genomic prediction in animals and plants: Simulation of data, validation, reporting, and benchmarking. *Genetics* 193 (2), 347–365. doi:10.1534/genetics.112.147983
- Daneshjoo, Roxana, Wang, Yanran, Bromberg, Yana, Bovo, Samuele, Martelli, Pier L., Babbi, Giulia, et al. (2017). Working toward precision medicine: Predicting phenotypes from exomes in the critical assessment of genome interpretation (cagi) challenges. *Hum. Mutat.* 38 (9), 1182–1192. doi:10.1002/humu.23280
- Dehouck, Yves, Kwasigroch, Jean Marc, Gilis, Dimitri, and Rooman, Marianne (2011). Popmusic 2.1: A web server for the estimation of protein stability changes upon mutation and sequence optimality. *BMC Bioinforma.* 12 (1), 151–212. doi:10.1186/1471-2105-12-151
- Dong, Chengliang, Peng, Wei, Jian, Xueqiu, Gibbs, Richard, Boerwinkle, Eric, Wang, Kai, et al. (2015). Comparison and integration of deleteriousness prediction methods for nonsynonymous snvs in whole exome sequencing studies. *Hum. Mol. Genet.* 24 (8), 2125–2137. doi:10.1093/hmg/ddu733
- Erwin, L., Dijk, V., Auger, H., Yan, J., and Thermes, C. (2014). Ten years of next-generation sequencing technology. *Trends Genet.* 30 (9), 418–426. doi:10.1016/j.tig.2014.07.001
- Francisco, M., and Bustamante, Carlos D. (2018). Polygenic risk scores: A biased prediction? *Genome Med.* 10 (1), 100–103. doi:10.1186/s13073-018-0610-x
- Fröhlich, Holger, Balling, Rudi, Beerenwinkel, Niko, Kohlbacher, Oliver, Kumar, Santosh, Lengauer, Thomas, et al. (2018). From hype to reality: Data science enabling personalized medicine. *BMC Med.* 16 (1), 150–215. doi:10.1186/s12916-018-1122-7
- Gazzo, Andrea, Raimondi, Daniele, Daneels, Dorien, Moreau, Yves, Smits, Guillaume, Van Dooren, Sonia, et al. (2017). Understanding mutational effects in digenic diseases. *Nucleic acids Res.* 45 (15), e140. doi:10.1093/nar/gkx557
- Gibson, Greg (2012). Rare and common variants: Twenty arguments. *Nat. Rev. Genet.* 13 (2), 135–145. doi:10.1038/nrg3118
- Goodwin, Sara, McPherson, John D., and McCombie, W. Richard (2016). Coming of age: Ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* 17 (6), 333–351. doi:10.1038/nrg.2016.49
- Grinberg, Nastasiya F., Orhobor, Oghenejokpeme I., and King, Ross D. (2020). An evaluation of machine-learning for predicting phenotype: Studies in yeast, rice, and wheat. *Mach. Learn.* 109 (2), 251–277. doi:10.1007/s10994-019-05848-5
- Guala, Dimitri, and Sonnhammer, Erik L. (2017). A large-scale benchmark of gene prioritization methods. *Sci. Rep.* 7 (1), 46598–46610. doi:10.1038/srep46598
- Hickey, John M., Chiurugwi, Tinashe, Mackay, Ian, and Powell, Wayne (2017). Genomic prediction unifies animal and plant breeding programs to form platforms for biological discovery. *Nat. Genet.* 49 (9), 1297–1303. doi:10.1038/ng.3920
- Houle, David, R Govindaraju, Diddahally, and Omholt, Stig (2010). Phenomics: The next challenge. *Nat. Rev. Genet.* 11 (12), 855–866. doi:10.1038/nrg2897
- Ioannidis, Nilah M., Rothstein, Joseph H., Pejaver, Vikas, Middha, Sumit, McDonnell, Shannon K., Baheti, Saurabh, et al. (2016). Revel: An ensemble method for predicting the pathogenicity of rare missense variants. *Am. J. Hum. Genet.* 99 (4), 877–885. doi:10.1016/j.ajhg.2016.08.016
- Jagadeesh, K. A., Wenger, A. M., Berger, M. J., Guturu, H., Stenson, P. D., Cooper, D. N., et al. (2016). M-cap eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. *Nat. Genet.* 48 (12), 1581–1586. doi:10.1038/ng.3703
- Jumper, John, Evans, Richard, Alexander, Pritzel, Green, Tim, Figurnov, Michael, Ronneberger, Olaf, et al. (2021). Highly accurate protein structure prediction with alphafold. *Nature* 596 (7873), 583–589. doi:10.1038/s41586-021-03819-2
- Kircher, Martin, Daniela, M. Witten, Jain, Preti, J O'roak, Brian, Cooper, Gregory M., and Shendure, Jay (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* 46 (3), 310–315. doi:10.1038/ng.2892
- Lakshman, Sundaram, Bhat, Rajendra Rana, Viswanath, Vivek, and Li, Xiaolin (2017). Deepbipolar: Identifying genomic mutations for bipolar disorder via deep learning. *Hum. Mutat.* 38 (9), 1217–1224. doi:10.1002/humu.23272
- Lapuschkin, Sebastian, Wäldchen, Stephan, Binder, Alexander, Montavon, Grégoire, Samek, Wojciech, and Müller, Klaus-Robert (2019). Unmasking clever hans predictors and assessing what machines really learn. *Nat. Commun.* 10 (1), 1096–1098. doi:10.1038/s41467-019-08987-4
- Li, Min, Wang, Yake, Zheng, Ruiqing, Shi, Xinghua, Li, Yaohang, Wu, Fang-Xiang, et al. (2019). Deepdsc: A deep learning method to predict drug sensitivity of cancer cell lines. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 18 (2), 575–582. doi:10.1109/tcbb.2019.2919581
- Maldonado, Carlos, Mora-Poblete, Freddy, Contreras-Soto, Rodrigo Iván, Ahmar, Sunny, Chen, Jen-Tsung, Teixeira, António, et al. (2020). Genome-wide prediction of complex traits in two outcrossing plant species through deep learning and bayesian regularized neural network. *Front. Plant Sci.* 11, 593897. doi:10.3389/fpls.2020.593897
- Manolio, Teri A., Collins, Francis S., Cox, Nancy J., Goldstein, David B., Hindorf, Lucia A., Hunter, David J., et al. (2009). Finding the missing heritability of complex diseases. *Nature* 461 (7265), 747–753. doi:10.1038/nature08494
- Moreau, Yves, and Tranchevent, Léon-Charles (2012). Computational tools for prioritizing candidate genes: Boosting disease gene discovery. *Nat. Rev. Genet.* 13 (8), 523–536. doi:10.1038/nrg3253
- Morrison, Alanna C., Bare, Lance A., Chambless, Lloyd E., Ellis, Stephen G., Malloy, Mary, Kane, John P., et al. (2007). Prediction of coronary heart disease risk using a genetic risk score: The atherosclerosis risk in communities study. *Am. J. Epidemiol.* 166 (1), 28–35. doi:10.1093/aje/kwm060
- Ng, Pauline C., Levy, Samuel, Huang, Jiaqi, Stockwell, Timothy B., Walenz, Brian P., Li, Kelvin, et al. (2008). Genetic variation in an individual human exome. *PLoS Genet.* 4 (8), e1000160. doi:10.1371/journal.pgen.1000160

- Niroula, Abhishek, and Vihinen, Mauno (2016). Variation interpretation predictors: Principles, types, performance, and choice. *Hum. Mutat.* 37 (6), 579–597. doi:10.1002/humu.22987
- Pal, Lipika R., Kundu, Kunal, Yin, Yizhou, and Moul, John (2017). Cagi4 crohn's exome challenge: Marker snp versus exome variant models for assigning risk of crohn disease. *Hum. Mutat.* 38 (9), 1225–1234. doi:10.1002/humu.23256
- Papadimitriou, Sofia, Gazzo, Andrea, Versbraegen, Nassim, Nachtegael, Charlotte, Aerts, Jan, Moreau, Yves, et al. (2019). Predicting disease-causing variant combinations. *Proc. Natl. Acad. Sci. U. S. A.* 116 (24), 11878–11887. doi:10.1073/pnas.1815601116
- Pucci, Fabrizio, Kwasigroch, Jean Marc, and Rooman, Marianne (2020). "Protein thermal stability engineering using hotmusic," in *Structural bioinformatics* (Berlin, Germany: Springer), 59–73.
- Raimondi, Daniele, Codicè, Francesco, Orlando, Gabriele, Schymkowitz, Joost, Rousseau, Frederic, and Moreau, Yves (2022). Hpmdb: A machine learning-ready database of protein molecular phenotypes associated to human missense variants. *Curr. Res. Struct. Biol.* 4, 167–174. doi:10.1016/j.crstbi.2022.04.004
- Raimondi, Daniele, Corso, Massimiliano, Fariselli, Piero, and Moreau, Yves (2022). From genotype to phenotype in arabidopsis thaliana: In-silico genome interpretation predicts 288 phenotypes from sequencing data. *Nucleic acids Res.* 50 (3), e16. doi:10.1093/nar/gkab1099
- Raimondi, Daniele, Gazzo, Andrea M., Rooman, Marianne, Lenaerts, Tom, and Vranken, Wim F. (2016). Multilevel biological characterization of exomic variants at the protein level significantly improves the identification of their deleterious effects. *Bioinformatics* 32 (12), 1797–1804. doi:10.1093/bioinformatics/btw094
- Raimondi, Daniele, Orlando, Gabriele, Fariselli, Piero, and Moreau, Yves (2020). Insight into the protein solubility driving forces with neural attention. *PLoS Comput. Biol.* 16 (4), e1007722. doi:10.1371/journal.pcbi.1007722
- Raimondi, Daniele, Simm, Jaak, Adam, Arany, Fariselli, Piero, Cleyne, Isabelle, and Moreau, Yves (2020). An interpretable low-complexity machine learning framework for robust exome-based *in-silico* diagnosis of crohn's disease patients. *Nar. Genom. Bioinform.* 2 (1), lqaa011. doi:10.1093/nargab/lqaa011
- Raimondi, Daniele, Tanyalcin, Ibrahim, Ferté, Julien, Gazzo, Andrea, Orlando, Gabriele, Lenaerts, Tom, et al. (2017). Deongen2: Prediction and interactive visualization of single amino acid variant deleteriousness in human proteins. *Nucleic acids Res.* 45 (W1), W201–W206. doi:10.1093/nar/gkx390
- Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H. R., Albarqouni, S., et al. (2020). The future of digital health with federated learning. *npj Digit. Med.* 3 (1), 119–127. doi:10.1038/s41746-020-00323-1
- Sifrim, A., Popovic, D., Tranchevent, L.-C., Amin, A., Sakai, R., Konings, P., et al. (2013). extasy: variant prioritization by genomic data fusion. *Nat. Methods* 10 (11), 1083–1084. doi:10.1038/nmeth.2656
- Smilkov, D., Thorat, Nikhil, Kim, Been, Viégas, Fernanda, and Martin, Wattenberg (2017). Smoothgrad: Removing noise by adding noise. Available at: <http://arXiv.org/abs/1706.03825>.
- Uffelmann, Emil, Huang, Qin Qin, Munung, Nchangwi Syntia, de Vries, Jantina, Okada, Yukinori, Martin, Alicia R., et al. (2021). Genome-wide association studies. *Nat. Rev. Methods Prim.* 1 (1), 59. doi:10.1038/s43586-021-00056-9
- Wald, Nicholas J., and Robert, Old (2019). The illusion of polygenic disease risk prediction. *Genet. Med.* 21 (8), 1705–1707. doi:10.1038/s41436-018-0418-5
- Wang, Yanran, Miller, Maximilian, Astrakhan, Yuri, Petersen, Britt-Sabina, Schreiber, Stefan, Franke, Andre, et al. (2019). Identifying crohn's disease signal from variome analysis. *Genome Med.* 11 (1), 59. doi:10.1186/s13073-019-0670-6
- Weedon, Michael N., McCarthy, Mark L., Graham, Hitman, Walker, Mark, Groves, Christopher J., Zeggini, Eleftheria., et al. (2006). Combining information from common type 2 diabetes risk polymorphisms improves disease prediction. *PLoS Med.* 3 (10), e374. doi:10.1371/journal.pmed.0030374
- Wei, Zhi, Wang, Wei, Bradfield, Jonathan, Jin, Li, Cardinale, Christopher, Frackelton, Edward., et al. (2013). Large sample size, wide variant spectrum, and advanced machine-learning technique boost risk prediction for inflammatory bowel disease. *Am. J. Hum. Genet.* 92 (6), 1008–1012. doi:10.1016/j.ajhg.2013.05.002
- Wray, Naomi R., Kemper, Kathryn E., Hayes, Benjamin J., Goddard, Michael E., and Visscher, Peter M. (2019). Complex trait prediction from genome data: Contrasting EBV in livestock to PRS in humans. *Genetics* 211 (4), 1131–1141. doi:10.1534/genetics.119.301859
- Wu, Jiaxin, Li, Yanda, and Jiang, Rui (2014). Integrating multiple genomic data to predict disease-causing nonsynonymous single nucleotide variants in exome sequencing studies. *PLoS Genet.* 10 (3), e1004237. doi:10.1371/journal.pgen.1004237