Quantitative Biology 2021, 9(4): 371–377
https://doi.org/10.15302/J-QB-021-0283

PERSPECTIVE

Roles of statistical modeling in characterizing the genetic basis of human diseases and traits

Hongyu Zhao*

Department of Biostatistics, Yale School of Public Health, New Haven, CT 06510, USA
* Correspondence: hongyu.zhao@yale.edu

Received October 9, 2021; Revised October 11, 2021; Accepted October 12, 2021

One major goal of human genetics is to identify variants, genes, and pathways causing diseases, both common and rare. Statistical thinking and approaches have played a major role in advancing human genetics research both historically and in recent years after the completion of the Human Genome Project. There has been a paradigm shift from family-based linkage studies to population-based association studies in the past 15 years thanks to the Human Genome Project, and many genomic regions have been implicated for human diseases and traits. The first success of genome wide association study (GWAS) was published in 2005 by Josephine Hoh and her colleagues [1] who were able to use only 96 cases and 50 controls and a low-resolution array to identify complement factor H as a major gene for age related macular degeneration. Since then, numerous GWAS have been performed for thousands of traits, with results catalogued at the GWAS Catalog (https://www.ebi.ac.uk/gwas/) and other web sites. Most of the identified associations have been for single nucleotide polymorphisms (SNPs), one of the most common genetic variations in the human genome. The number of SNPs that have been associated with a complex trait ranges from a few (*e.g.*, posttraumatic stress disorder) to a few thousands (*e.g.*, height [2]). These successes have benefited greatly from the developments and applications statistical methods to gather, model, analyze, and interpret GWAS data. In the following, we discuss several areas that statistical thinking and modeling have led to significant insights on the design and analysis of human genetic studies, and conclude with future challenges that demand further statistical developments.

This article is dedicated to the feature in 20th anniversary of Human Genome (Eds. Michael Q. Zhang and Xuegong Zhang).

CHOICE OF STUDY DESIGNS TO IDENTIFY COMPLEX DISEASE GENES

Compared to Mendelian diseases, which are caused by the mutations in single genes, complex diseases are the result of perturbation of many genes and pathways and it proved difficult to identify complex disease genes through family studies with a few exceptions that have genes with large effects, such as BRAC1 for breast cancer [3]. The basic idea for family-based study is to track co-segregation of inherited chromosomal segments with disease phenotypes among family members, and these are called linkage analyses. There are two major categories of linkage analyses, parametric and allele-sharing methods. Parametric linkage methods need to make explicit assumptions of disease models, *e.g.*, disease-causing allele frequencies and penetrance, that are often difficult to estimate. In contrast, allele sharing methods compared the observed allele sharing at candidate markers between affected relatives versus that expected by chance when the candidate markers are not in linkage to disease loci. As allele-sharing methods do not rely on specific disease model assumptions, these methods are also called non-parametric. However, very few loci were implicated with either parametric or allele-sharing methods for complex traits. In a seminal

paper published in 1996, Risch and Merikangas [4] showed that the allele-sharing methods, *e.g*., affected sib-pair methods, have very limited statistical power to identify genetic variants conferring moderate relative disease risks. For example, it may require hundreds of thousands of affected sib pairs to have adequate power. In contrast, study designs that look for association signals, *e.g*., case control studies, across individuals in the population (instead of co-segregation signals within families) between candidate markers and disease phenotypes are much more efficient. However, historically researchers were concerned about false positives due to population stratification and other confounding factors. To minimize false positives due to population stratification, Ewens and colleagues [5] proposed a family-based association design that treats the alleles transmitted from the parents to the affected offspring as cases and those untransmitted alleles from the parents to the affected offspring as controls for association analysis. It was shown that this design is robust to population stratification. In fact, this family-bases association design was considered by Risch and Merikangas [4] as a much more powerful alternative to the affected sib-pair design. It is interesting to note that the statistical significance level of 5×10^{-8} was used in [4] to make multiple comparison adjustment, which is the same as the significance threshold commonly used for GWAS. However, this level was adopted to adjust for 10^5 genes with each gene having 10 alleles, as the number of genes was estimated to be 10^5 in the human genome at that time. One limitation of the family-based association study was the need to collect probands and their parents and a number of methods were proposed to deal with the case of having one or no parents. Despite these developments, the design would still be considered less efficient than that based on standard case control studies where the subjects are unrelated individuals. Moreover, for late-onset diseases, it is challenging to collect biospecimen from affected individuals and their parents. Around this time, researchers realized the possibility of using genomic markers to infer and control population stratifications in the samples with a number of methods developed, including the now commonly used principal component analysis [6,7] to control for population stratification. We note that the statistical power comparison of the affected sib-pair design and family-based association design by Risch and Merikangas [4] provided theoretical foundation for the launch of the HapMap project [8], and therefore the revolution brought on by genome wide association studies.

MISSING HERITABILITY

After the initial GWAS successes, it was found that the significant SNPs identified from GWAS only explained a small proportion of the genetic contributions as inferred from twin and family studies. This so-called "missing heritability" [9] led to different hypotheses on the limitations of the GWAS approach to identifying disease causing genes, such as contributions from rare variants and possible gene-gene interactions that could not be captured by GWAS analysis. However, it came as a somewhat surprise to many human geneticists that Peter Visscher and colleagues demonstrated in a 2010 seminal paper [10] that the SNPs captured by genotyping arrays could indeed account for a major proportion of the heritability although only a limited number of them could be found to be statistically significant. To discuss the basic idea of their work, we need to go back to a very basic quantitative genetics model introduced by R. A. Fisher in his 1918 paper [11] that laid the foundation for quantitative genetics. In his paper, Fisher assumed that a quantitative trait is the result of the additive effects from many genetic variants. Based on this model, the overall genetic contribution to the trait variation, *i.e.*, heritability both in the broad and narrow sense, can be defined and the overall genetic contribution can be further partitioned into additive and dominance components. Fisher showed that the phenotypic correlation between a pair of relatives is the weighted sum of the additive and dominance variances with weights determined by the specific relationships between individuals. This formulation facilitates the estimation of additive and dominance variance from a collection of relative sets, without knowing any genetic marker information. For a binary trait, an unobserved liability can be assumed to be underlying the binary trait and a person will manifest the disease if the liability score is above a threshold. These models have been used to estimate heritability from pedigree data using the observed phenotypes and reported relationships. In fact, the "missing heritability" refers to the significant gap between the phenotype variations that can be explained by the genetic markers identified from GWAS and the heritability estimated from family and twin studies.

Instead of considering only genome wide significant markers, Visscher *et al*. [10] started from the Fisher model and assumed that the effect sizes of all the markers that can be analyzed (either observed or imputed) follow the same normal distribution, where the genotyping scores for each marker are normalized to have mean 0 and the variance 1. This genotype standardization effectively assumes that markers with lower minor allele frequencies tend to have larger effect sizes. Based on this assumption, the covariance of trait values between two individuals will depend on the overall genetic similarity across all the markers in the genome, not just the markers with genome wide statistical

significance. Based on this framework, Visscher *et al*. demonstrated that a substantial amount of heritability inferred from family studies can be indeed explained by this so-called chip-based heritability using genetic similarities between individuals. It is called chip-based heritability because only markers that can be inferred from genotyping arrays are considered. This paper shows the critical importance of using appropriate statistical models and inferential procedures to interpret GWAS results, and the random effects model considered in this paper has become the standard for GWAS modeling and analysis.

One limitation of the random effects model used by Visscher *et al*. is that it assumes that all the markers have phenotypic effects and this unlikely holds in practice. It is more likely only a small proportion of the markers can affect a trait where the others are not associated with the trait at all. Fortunately, it was shown that the statistical inference approach for the random effects model used by Visscher *et al*. still provides unbiased estimate of heritability when only a limited proportion of the markers are trait associated [12].

In addition to the simple random effects model where all the markers are assumed to make equal contribution to the trait variation, other models have also been proposed to characterize the relationship between the effect sizes of the SNPs and their properties, *e.g*., allele frequencies and linkage disequilibrium (LD) patterns [13].

SUMMARY STATISTICS-BASED METHODS

Most classical statistical methods assume the availability of individual level data for inference. In the context of GWAS, both individual level genotype and phenotype data are needed. However, due to both privacy and other concerns, individual level data may not be readily shared across studies or with the research community. Instead, summary statistics are often provided from published GWAS. This has stimulated many methods that only require input from summary statistics, including SNP allele frequencies, effect sizes and their standard errors [14]. For the purpose of estimating chip-based heritability, LD score regression [15] represents a major advancement and is commonly used for heritability estimate from summary statistics although it was initially proposed to investigate whether there is any evidence of population stratification in the GWAS results. We note that there are often parallel developments of methods in the GWAS literature for both individual level data and summary statistics.

ENRICHMENT ANALYSIS

Although the random effects model considered by Visscher *et al*. aimed to answer the question encompassing all the markers in the genome, this model can also be applied to a subset of SNPs, *e.g*., those in a pathway or on a specific chromosome, to estimate the heritability explained by this subset of SNPs. In fact, it was found that the variance of phenotype explained by the markers on different chromosomes was for the most part proportional to the chromosomal length with a few exceptions. Based on these estimates, we can identify pathways and marker sets that are enriched for GWAS signals. For example, if the SNPs in a pathway account for 10% of the overall SNPs in the genome but can explain 30% of the chip-based heritability, this represents a three-fold enrichment suggesting the importance of this pathway. The SNP set can be constructed based on other information. For example, we can use the genetic variants annotated to be functional in a specific tissue or cell type and ask the question whether there is enrichment for certain tissues or cell types. Such analyses have implicated the importance of the immune system for Alzheimer disease and Parkinson disease [16]. Both individual level data and summary statistics have been considered for enrichment analysis [17,18].

PLEIOTROPY

Pleiotropy refers to the phenomenon that a genetic variant can affect more than one trait, and it is commonly observed in human genetics. Although pleiotropy can be studied for individual SNPs, genetic correlation has been introduced to quantify the degree of genetic sharing between two traits across the genome. This is formulated by simultaneously considering two random effects model for the two traits, with where genetic correlation is defined by the correlation of the SNP effects on the two traits. Again, both individual data and summary statistics can be used to infer genetic correlation [19–21]. This line of work has led to many methodology developments, including the partitioning of the overall covariance into different annotations [22] and the analysis of local genetic correlation [23]. Although the random effects model is useful for quantifying genetic correlations for the effect sizes for common variants, it is more challenging for rare variants due to the large errors in estimating their effect sizes. Alternative methods are needed to quantify concordance of association signals for rare and *de novo* variants between traits [24]. With shared genetics between traits, we can leverage this to better identify disease genes [25,26].

GENETIC RISK PREDICTION

With the identifications of many SNPs associated with various traits through GWAS, it is natural to translate

these results into useful risk prediction models. The initial risk prediction models were built from SNPs attaining genome-wide statistical significance. However, it was noted early on that there is information from those SNPs without genome wide significance, and including those SNPs may improve prediction accuracy [27]. Built from the same random effects model, improved effect size estimates can be obtained through borrowing information across markers, leading to more accurate prediction models. When only summary statistics are available, a number of methods have been developed that are based on different assumptions of the effect size distributions and statistical inference procedures. The most accurate models are based on Bayesian formulations where different priors are assumed for the proportions of SNPs having non-zero effects on the traits and the distribution for the effect sizes. For example, LD-Pred [28] assumed a spike-slab prior with a point mass at 0 and normal distribution for the other markers. Efforts have also been made to incorporate the annotation information [29] and pleiotropic information [30,31] to improve risk prediction. More recent developments have considered other prior distributions [32,33], including very general non-parametric priors [34].

IDENTIFICATIONS OF FUNCTIONAL GENES AND VARIANTS

Although GWAS have identified many chromosomal regions associated with complex traits, it has been challenging to infer the disease-causing genes and variants due to the presence of LD which refers to statistical dependence of nearby markers. Although LD was critical in designing genotyping arrays and imputations for untyped markers, its presence makes it difficult to distinguish SNPs that are functional from those that are associated with the traits simply due to their associations with the true SNPs through LD. Another difficulty in identifying disease-causing genes and variants is that most significant SNPs are not in the coding regions, suggesting the importance of non-coding regions for complex traits. Yet not much was known about the non-coding regions. A number of national and international programs have been developed, mostly notably ENCODE [35], Roadmap Epigenomics [36], and GTEx [37] to understand the regulatory regions of the human genome. The data gathered from these programs have stimulated a number of data integration methods to identify disease genes, as highlighted by transcriptome wide association study [38–41]. These data have also motivated many functional annotations of the human genome [42], with annotations done at the organism, tissue, and cell type level, allowing more refined analysis. These annotations and biological network information can be used to prioritize disease genes [43–46]. Fine mapping can also be performed based on the integration of these annotations [47,48]. Data from recent projects, such as the Human Cell Atlas project [49], the 4D Nucleome project [50], and other NIH Common Fund projects also offer rich information to identify functional genes and variants.

FUTURE CHALLENGES

As summarized above, remarkable progress has been made to identify SNPs and other markers associated with many human diseases and traits. It is an understatement that GWAS has changed the landscape of human genetics research. From knowing almost no genetic factors for most common diseases, researchers have accumulated many hit regions in the human genome with reproducible statistical association signals yet little understanding the disease-causing genes and variants, and mechanisms. Looking forward in the next several years in human genetics research, there is going to be continued accumulation of genomic loci associated with complex traits, and the identifications of disease-causing genes and variants as a result of ever-increasing sample sizes in different populations, including many biobanks; sequencing of whole exomes and whole genomes for study participants; access to comprehensive medical records, behavioral and exposure information; diverse molecular phenotypes; and large-scale functional assays of genetic variants. Statistical modeling and inference will continue to play important roles to move human genetics forward to ask and answer key questions.

First, causal inference is at the center of many scientific inquiries. Mendelian randomization (MR) uses SNPs as instrumental variables to infer causal relationships among traits [51,52]. There are three key assumptions for the validity of the MR approach, including the absence of horizontal pleiotropy and the absence of association between instrumental variables (SNPs) and confounding factors. Many methods have been developed to make the general MR more robust when these assumptions are violated, and the applications of different methods to the same data often lead to conflicting results [53]. New methods are needed to more effectively combine genetic, environmental, and other information together to consider both unrelated and related subjects in causal inference. Given the large sample size for many genetics studies these days, which may be in the order of hundreds of thousands to millions, much care is needed to not misinterpret associations due to unmeasured subtle confounding

effects as causal.

Second, there is a need to better understand the interplay between common and rare variants on disease onset and progression [54,55]. The analysis of rare variants has been hindered by the available sample sizes to date, but this will likely change in the near future. These can be all be done in the context of diverse data types, such as gene expression [56,57], protein expression, and epigenetic information collected at different scales.

Third, genetic risk prediction efforts have primarily focused on the genetic contributions with different degrees of considerations of other variables. Although many methods have been proposed, the performance is in general very close to each other. To significantly improve prediction accuracy, additional data, such as trajectory information from medical records and molecular data, and more sophisticated modeling, such as those from deep learning and incorporation of non-linear and non-additive effects, may be needed. Ideally, a person's disease risk at a certain point in life should be based on all available information up to that point, not just baseline and genetic information. A much-coordinated effort is needed as this will require much more comprehensive data as well as consideration of data privacy.

Fourth, in addition to predicting disease risks, for many individuals already affected with diseases especially those with congenital diseases, the primary interest is the identification of disease-causing pathways/genes/variants for a specific patient. The computational tools are lacking to sift through many candidate genes and variants for effective prioritizations. Existing methods are not able to leverage the detailed phenotypes from patients and the knowledge gained in the past 15 years from GWAS and other studies yet. This is an area that will likely see rapid development in the near future due to its importance and the maturity and availability of many data collection tools.

Fifth, it is expected that genes and variants identified from GWAS may inform drug developments, either repositioning existing drugs for new indications or selecting compounds from a large number of candidates [58]. Although some progress has been made in this regard, much more can be done in combination with other resources.

Last but not least, we need to recognize the similarity and difference of the genetic contributions to diseases and traits across different populations. For example, most genetic risk predictions have focused on individuals of European ancestry, partly because most subjects in existing GWAS are of European ancestry. It has been shown that the PRS developed from European samples may not perform well for other populations [59]. As a result, clinical use of the PRS thus developed may exacerbate health disparities [60]. Therefore, there is a great need to collect and analyze data from diverse populations to understand biological pathways and predict disease risks.

Rigorous statistical modeling and analysis will be indispensable for all these aspects of future developments so that diverse and rich data sources can be more coherently integrated. It is also likely that new statistical methods may be developed motivated from answering these questions just like what have happened since the first publication of the human genome sequences. Compared to 20 years ago, we have many more statistical tools in our arsenal and much more computing power. Coupled with many more data sources, the next decade will be even more exciting and rewarding for statistical genetics and genomics.

ACKNOWLEDGMENTS

This work was supported in part by NSF grant DMS 1902903 and NIH grants R03HD100883, R03OD030609, and R01GM134005.

COMPLIANCE WITH ETHICS GUIDELINES

The author Hongyu Zhao declares he has no conflict of interests.

OPEN ACCESS

REFERENCES

1. Klein, R. J., Zeiss, C., Chew, E. Y., Tsai, J. Y., Sackler, R. S., Haynes, C., Henning, A. K., SanGiovanni, J. P., Mane, S. M., Mayne, S. T., *et al*. (2005) Complement factor H polymorphism in age-related macular degeneration. Science, 308, 385−389
2. Yengo, L., Sidorenko, J., Kemper, K. E., Zheng, Z., Wood, A. R., Weedon, M. N., Frayling, T. M., Hirschhorn, J., Yang, J., Visscher, P. M., *et al*. (2018) Meta-analysis of genome-wide association studies for height and body mass index in ~700000 individuals of European ancestry. Hum. Mol. Genet., 27, 3641−3649
3. Hall, J. M., Lee, M. K., Newman, B., Morrow, J. E., Anderson,

L. A., Huey, B. and King, M. C. (1990) Linkage of early-onset familial breast cancer to chromosome 17q21. Science, 250, 1684−1689
4. Risch, N. and Merikangas, K. (1996) The future of genetic studies of complex human diseases. Science, 273, 1516−1517
5. Spielman, R. S., McGinnis, R. E. and Ewens, W. J. (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am. J. Hum. Genet., 52, 506−516
6. Zhang, S., Zhu, X. and Zhao, H. (2003) On a semiparametric test to detect associations between quantitative traits and candidate genes using unrelated individuals. Genet. Epidemiol., 24, 44−56
7. Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A. and Reich, D. (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet., 38, 904−909
8. International HapMap Consortium. (2003) The International HapMap Project. Nature, 426, 789−796
9. Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., *et al*. (2009) Finding the missing heritability of complex diseases. Nature, 461, 747−753
10. Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., Madden, P. A., Heath, A. C., Martin, N. G., Montgomery, G. W., *et al*. (2010) Common SNPs explain a large proportion of the heritability for human height. Nat. Genet., 42, 565−569
11. Fisher, R. A. (1918) The correlation between relatives on the supposition of Mendelian inheritance. Trans. R. Soc. Edinb., 52, 399−433
12. Jiang, J., Li, C., Paul, D., Yang, C. and Zhao, H. (2016) On high-dimensional misspecified mixed model analysis in genome-wide association study. Ann. Stat., 44, 2127−2160
13. Speed, D., Cai, N., Johnson, M. R., Nejentsev, S. and Balding, D. J., and the UCLEB Consortium. (2017) Reevaluation of SNP heritability in complex human traits. Nat. Genet., 49, 986−992
14. Pasaniuc, B. and Price, A. L. (2017) Dissecting the genetics of complex traits using summary association statistics. Nat. Rev. Genet., 18, 117−127
15. Bulik-Sullivan, B. K., Loh, P. R., Finucane, H. K., Ripke, S., Yang, J., Patterson, N., Daly, M. J., Price, A. L., Neale, B. M., and the Schizophrenia Working Group of the Psychiatric Genomics Consortium. (2015) LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat. Genet., 47, 291−295
16. Lu, Q., Powles, R. L., Abdallah, S., Ou, D., Wang, Q., Hu, Y., Lu, Y., Liu, W., Li, B., Mukherjee, S., *et al*. (2017) Systematic tissue-specific functional annotation of the human genome highlights immune-related DNA elements for late-onset Alzheimer's disease. PLoS Genet., 13, e1006933
17. Finucane, H. K., Bulik-Sullivan, B., Gusev, A., Trynka, G., Reshef, Y., Loh, P. R., Anttila, V., Xu, H., Zang, C., Farh, K., *et al*. (2015) Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat. Genet., 47, 1228−1235
18. Lu, Q., Powles, R. L., Wang, Q., He, B. J. and Zhao, H. (2016) Integrative tissue-specific functional annotations in the human genome provide novel insights on many complex traits and improve signal prioritization in genome wide association studies. PLoS Genet., 12, e1005947
19. Lee, S. H., Yang, J., Goddard, M. E., Visscher, P. M. and Wray, N. R. (2012) Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. Bioinformatics, 28, 2540−2542
20. Bulik-Sullivan, B., Finucane, H. K., Anttila, V., Gusev, A., Day, F. R., Loh, P. R., Duncan, L., Perry, J. R., Patterson, N., Robinson, E. B., *et al*. (2015) An atlas of genetic correlations across human diseases and traits. Nat. Genet., 47, 1236−1241
21. van Rheenen, W., Peyrot, W. J., Schork, A. J., Lee, S. H. and Wray, N. R. (2019) Genetic correlations of polygenic disease traits: from theory to practice. Nat. Rev. Genet., 20, 567−581
22. Lu, Q., Li, B., Ou, D., Erlendsdottir, M., Powles, R. L., Jiang, T., Hu, Y., Chang, D., Jin, C., Dai, W., *et al*. (2017) A powerful approach to estimating annotation-stratified genetic covariance via GWAS summary statistics. Am. J. Hum. Genet., 101, 939−964
23. Zhang, Y., Lu, Q., Ye, Y., Huang, K., Liu, W., Wu, Y., Zhong, X., Li, B., Yu, Z., Travers, B. G., *et al*. (2021) SUPERGNOVA: local genetic correlation analysis reveals heterogeneous etiologic sharing of complex traits. Genome Biol., 22, 262
24. Guo, H. , Hou, L. , Shi, Y. , Jin, S. C. , Zeng, X. , Li, B. , Lifton, R. P. , Brueckner, M. , Zhao, H. , Lu Q. (2021) Quantifying concordant genetic effects of *de novo* mutations on multiple disorders. bioRxiv, 2021.06.13.448234
25. Turley, P., Walters, R. K., Maghzian, O., Okbay, A., Lee, J. J., Fontana, M. A., Nguyen-Viet, T. A., Wedow, R., Zacher, M., Furlotte, N. A., *et al*. (2018) Multi-trait analysis of genome-wide association summary statistics using MTAG. Nat. Genet., 50, 229−237
26. Xie, Y., Li, M., Dong, W., Jiang, W. and Zhao, H. (2021) M-DATA: A statistical approach to jointly analyzing *de novo* mutations for multiple traits. medRxive,
27. Kang, J., Cho, J. and Zhao, H. (2010) Practical issues in building risk-predicting models for complex diseases. J. Biopharm. Stat., 20, 415−440
28. Vilhjálmsson, B. J., Yang, J., Finucane, H. K., Gusev, A., Lindström, S., Ripke, S., Genovese, G., Loh, P. R., Bhatia, G., Do, R., *et al*. (2015) Modeling linkage disequilibrium increases accuracy of polygenic risk scores. Am. J. Hum. Genet., 97, 576−592
29. Hu, Y., Lu, Q., Powles, R., Yao, X., Yang, C., Fang, F., Xu, X. and Zhao, H. (2017) Leveraging functional annotations in genetic risk prediction for human complex diseases. PLoS Comput. Biol., 13, e1005589
30. Li, C., Yang, C., Gelernter, J. and Zhao, H. (2014) Improving genetic risk prediction by leveraging pleiotropy. Hum. Genet., 133, 639−650
31. Hu, Y., Lu, Q., Liu, W., Zhang, Y., Li, M. and Zhao, H. (2017) Joint modeling of genetically correlated diseases and functional

annotations increases accuracy of polygenic risk prediction. PLoS Genet., 13, e1006836
32. Ge, T., Chen, C. Y., Ni, Y., Feng, Y. A. and Smoller, J. W. (2019) Polygenic prediction via Bayesian regression and continuous shrinkage priors. Nat. Commun., 10, 1776
33. Song, S., Jiang, W., Hou, L. and Zhao, H. (2020) Leveraging effect size distributions to improve polygenic risk scores derived from summary statistics of genome-wide association studies. PLoS Comput. Biol., 16, e1007565
34. Zhou, G. and Zhao, H. (2021) A fast and robust Bayesian nonparametric method for prediction of complex traits using summary statistics. PLoS Genet., 17, e1009697
35. Consortium, E. P., and the ENCODE Project Consortium (2004) The ENCODE (ENCyclopedia Of DNA Elements) Project. Science, 306, 636–640
36. Bernstein, B. E., Stamatoyannopoulos, J. A., Costello, J. F., Ren, B., Milosavljevic, A., Meissner, A., Kellis, M., Marra, M. A., Beaudet, A. L., Ecker, J. R., *et al*. (2010) The NIH roadmap epigenomics mapping consortium. Nat. Biotechnol., 28, 1045−1048
37. Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., *et al*. (2013) The genotype-tissue expression (GTEx) project. Nat. Genet., 45, 580−585
38. Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B. W., Jansen, R., de Geus, E. J., Boomsma, D. I., Wright, F. A., *et al*. (2016) Integrative approaches for large-scale transcriptome-wide association studies. Nat. Genet., 48, 245−252
39. Gamazon, E. R., Wheeler, H. E., Shah, K. P., Mozaffari, S. V., Aquino-Michaels, K., Carroll, R. J., Eyler, A. E., Denny, J. C., Nicolae, D. L., Cox, N. J., *et al*. (2015) A gene-based association method for mapping traits using reference transcriptome data. Nat. Genet., 47, 1091−1098
40. Hu, Y., Li, M., Lu, Q., Weng, H., Wang, J., Zekavat, S. M., Yu, Z., Li, B., Gu, J., Muchnik, S., *et al*. (2019) A statistical framework for cross-tissue transcriptome-wide association analysis. Nat. Genet., 51, 568−576
41. Liu, W., Li, M., Zhang, W., Zhou, G., Wu, X., Wang, J., Lu, Q. and Zhao, H. (2020) Leveraging functional annotation to identify genes associated with complex diseases. PLoS Comput. Biol., 16, e1008315
42. Li, X., Li, Z., Zhou, H., Gaynor, S. M., Liu, Y., Chen, H., Sun, R., Dey, R., Arnett, D. K., Aslibekyan, S., *et al*. (2020) Dynamic incorporation of multiple *in silico* functional annotations empowers rare variant association analysis of large whole-genome sequencing studies at scale. Nat. Genet., 52, 969−983
43. Chung, D., Yang, C., Li, C., Gelernter, J. and Zhao, H. (2014) GPA: a statistical approach to prioritizing GWAS results by integrating pleiotropy and annotation. PLoS Genet., 10, e1004787
44. Chen, M., Cho, J. and Zhao, H. (2011) Incorporating biological pathways via a Markov random field model in genome-wide association studies. PLoS Genet., 7, e1001353
45. Hou, L., Chen, M., Zhang, C. K., Cho, J. and Zhao, H. (2014) Guilt by rewiring: gene prioritization through network rewiring in genome wide association studies. Hum. Mol. Genet., 23, 2780−2790
46. Lu, Q., Yao, X., Hu, Y. and Zhao, H. (2016) GenoWAP: GWAS signal prioritization through integrated analysis of genomic functional annotation. Bioinformatics, 32, 542−548
47. Weissbrod, O., Hormozdiari, F., Benner, C., Cui, R., Ulirsch, J., Gazal, S., Schoech, A. P., van de Geijn, B., Reshef, Y., Márquez-Luna, C., *et al*. (2020) Functionally informed fine-mapping and polygenic localization of complex trait heritability. Nat. Genet., 52, 1355−1363
48. Hutchinson, A., Asimit, J. and Wallace, C. (2020) Fine-mapping genetic associations. Hum. Mol. Genet., 29, R81−R88
49. Rozenblatt-Rosen, O., Stubbington, M. J. T., Regev, A. and Teichmann, S. A. (2017) The Human Cell Atlas: from vision to reality. Nature, 550, 451−453
50. Dekker, J., Belmont, A. S., Guttman, M., Leshyk, V. O., Lis, J. T., Lomvardas, S., Mirny, L. A., O'Shea, C. C., Park, P. J., Ren, B., *et al*. (2017) The 4D nucleome project. Nature, 549, 219−226
51. Katan, M. B. (1986) Apolipoprotein E isoforms, serum cholesterol, and cancer. Lancet, 327, 507−508
52. Smith, G. D. and Ebrahim, S. (2003) 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? Int J. Epidemiol., 32, 1−22
53. Slob, E. A. W. and Burgess, S. (2020) A comparison of robust Mendelian randomization methods using summary data. Genet. Epidemiol., 44, 313−329
54. Bodmer, W. and Bonilla, C. (2008) Common and rare variants in multifactorial susceptibility to common diseases. Nat. Genet., 40, 695−701
55. Crouch, D. J. M. and Bodmer, W. F. (2020) Polygenic inheritance, GWAS, polygenic risk scores, and the search for functional variants. Proc. Natl. Acad. Sci. USA, 117, 18924−18933
56. Zhao, J., Akinsanmi, I., Arafat, D., Cradick, T. J., Lee, C. M., Banskota, S., Marigorta, U. M., Bao, G. and Gibson, G. (2016) A burden of rare variants associated with extremes of gene expression in human peripheral blood. Am. J. Hum. Genet., 98, 299−309
57. Li, X., Kim, Y., Tsang, E. K., Davis, J. R., Damani, F. N., Chiang, C., Hess, G. T., Zappala, Z., Strober, B. J., Scott, A. J., *et al*. (2017) The impact of rare variation on gene expression across tissues. Nature, 550, 239−243
58. So, H. C., Chau, C. K., Chiu, W. T., Ho, K. S., Lo, C. P., Yim, S. H. and Sham, P. C. (2017) Analysis of genome-wide association data highlights candidates for drug repositioning in psychiatry. Nat. Neurosci., 20, 1342−1349
59. Duncan, L., Shen, H., Gelaye, B., Meijsen, J., Ressler, K., Feldman, M., Peterson, R. and Domingue, B. (2019) Analysis of polygenic risk score usage and performance in diverse human populations. Nat. Commun., 10, 3328
60. Martin, A. R., Kanai, M., Kamatani, Y., Okada, Y., Neale, B. M. and Daly, M. J. (2019) Clinical use of current polygenic risk scores may exacerbate health disparities. Nat. Genet., 51, 584−591

