

EDITORIAL

Special issue on genome wide association study

Lin Hou¹, Qiongshi Lu², Can Yang³, Hongyu Zhao^{4,5,6,*}

¹ Center for Statistical Science; Department of Industrial Engineering; MOE Key Laboratory of Bioinformatics, School of Life Sciences; Tsinghua University, Beijing 100084, China

² Department of Biostatistics and Medical Informatics; Department of Statistics; Center for Demography of Health and Aging; University of Wisconsin-Madison, WI 53726, USA

³ Department of Mathematics, the Hong Kong University of Science and Technology, Hong Kong, China

⁴ Department of Biostatistics, Yale School of Public Health, New Haven, CT 06510, USA

⁵ Department of Genetics, Yale University, New Haven, CT 06510, USA

⁶ Program of Computational Biology and Bioinformatics, Yale University, New Haven, CT 06510, USA

* Correspondence: hongyu.zhao@yale.edu

Received January 18, 2021

Since the first success of genome wide association study (GWAS) was reported in 2005 by Hoh and colleagues on the identification of a major gene for age related macular degeneration, many thousands of published studies have reported reproducible association signals between hundreds of thousands of genetic variants and thousands of traits and diseases in the past 16 years. There is a growing number of biobanks with up to one million participants developed across the world to facilitate genetic studies of complex traits through GWAS. Furthermore, many consortium projects have been initiated to collect diverse types of data that have proved very valuable for integrated analysis of GWAS data. These rich data present both great opportunities and challenges to scientists on data management, computation, analysis, integration, and interpretation. This special issue of *Quantitative Biology* collects nine papers that cover major progresses, challenges, and future directions for GWAS analysis.

In their review, Tang & He offer a comprehensive overview of advances and challenges for GWAS. One lesson that GWAS researchers have learned is that integrative analysis can be very powerful to both identify additional genes and dissect the genetic architecture of complex diseases. Among different resources generated, the Genotype-Tissue Expression (GTEx) Project stands out to be among the most informative for GWAS. With first-hand experience from participation in the GTEx project, Liao *et al.* summarize key issues in the handling of GTEx data and major discoveries on the regulations of gene expressions across tissues. More detailed discussions of statistical developments of various methods to integrate GTEx data and other data sets with both genotype and expression data, an area called transcriptome wide association study (TWAS), are covered in Xie *et al.* and Zhu & Zhou. In addition to gene expression data, imaging traits have also offered additional information as discussed by Knutson & Pan. As for downstream analysis of GWAS results, Mendelian randomization has been widely adopted for causal inference among traits. This important topic is the focus of the review by Zhu and is also discussed in the context of TWAS. Another major area of GWAS downstream analysis is the development of polygenic risk scores (PRS) that can be used to identify individuals at higher risk for different diseases for more effective monitoring and prevention. Zhao *et al.* review the state-of-the-art methods for PRS construction, and these PRS methods are instrumental to the success of direct-to-consumer genetic testing practices, a topic covered by Kang *et al.* While GWAS have mostly focused on common variants, the availability of whole exome and whole genome sequencing data will be next frontier for human genetics research. Li *et al.* consider the analysis of *de novo* mutations for congenital heart disease.

As many biobank data sets are coupled with electronic medical records, and some participants also have

their imaging, wearable device, and other data available, it is anticipated that much will be learned from these data to identify the genetic basis of complex diseases across different ethnic groups through the development and applications of novel statistical and computational methods that can best use these data sources. The genetic findings will likely lead to novel disease treatments, and make personalized monitoring, prevention, diagnosis, and treatment a reality. We hope this issue will draw interests into this fascinating field.