frontiers | Frontiers in Genetics

EDITORIAL
published: 14 April 2022
doi: 10.3389/fgene.2022.882611

Editorial: Current Status and Future Challenges of Biobank Data Analysis

Tzu-Pin Lu[1], Yoichiro Kamatani[2], Gillian Belbin[3], Taesung Park[4] and Chuhsing Kate Hsiao[1]*

[1]Department of Public Health, College of Public Health, Institute of Epidemiology and Preventive Medicine, National Taiwan University, Taipei, Taiwan, [2]Department of Computational Biology and Medical Sciences, Graduate School of Frontier Sciences, The University of Tokyo, Tokyo, Japan, [3]Institute of Genomic Health, Icahn School of Medicine at Mount Sinai, New York, NY, United States, [4]Department of Statistics, Seoul National University, Seoul, South Korea

Keywords: biobank data analysis, GWAS, integrative analysis, personalized medicine, statistical genetics

Editorial on the Research Topic

Current Status and Future Challenges of Biobank Data Analysis

Global health care systems are severely strained under the burden of aging populations, which are prone to enhanced incidence of chronic diseases. Biobanks help manage the population-wide disease burden by providing resources for scientific and medical research, allowing improved public health and individual patient care (Harris et al., 2012; Jacobs et al., 2018). Biobanks are the result of the combination of emerging novel technologies, well-catalogued human biological samples, and corresponding clinical and research data (Harris et al., 2012). The last 2 decades have witnessed the emergence of several biobanks around the world. Their primary aims are to uncover the genetic etiology of various chronic illnesses and to study the interaction of ethnicity with the underlying clinico-pathology and genetics of individuals across different populations (Sanner and Frazier, 2007).

A wide spectrum of phenotypic information is available in biobanks, including diagnoses, risk factors, physical and metabolic parameters, clinical information, as well as data on behavioral and social factors (Harris et al., 2012; Jacobs et al., 2018; Coppola et al., 2019). Single nucleotide polymorphism (SNP) genotyping arrays allow genome-wide association studies (GWASs) and polygenic risk score (PRS) development (Jimmy Juang et al., 2020) or phenome-wide association studies (PheWASs) (Liu and Crawford, 2022). Moreover, findings from GWASs can be utilized to conduct genetic tests on cohorts (Yamamoto et al., 2017). Biobanks require strict quality standards, which has led to the discovery of many new diagnostic and prognostic markers and a better pathophysiological understanding of disease development (Jacobs et al., 2018).

Biobanks provide high-dimensional data with large sample sizes, allowing gains in statistical power towards identification of novel genetic findings. However, the sheer amount of data they contain presents a greater computational burden that needs to be dealt with carefully. Another issue associated with traditional GWASs is the existence of linkage disequilibrium that leads to non-causal genetic markers. Additionally, for weak signals (minor allele frequency (MAF) ≤0.01), existing statistical models may not be powerful enough to detect putative rare variants. Hence, before conducting analyses and building models, it becomes imperative to reduce the data dimension while retaining the essential information of the original data (Sakaue et al., 2020). This is especially relevant for studies that aim to account for missing heritability, due to the modest effect of GWAS-reported loci on disease risk of complex traits, through genetic interactions (gene-gene, gene-environment). Prospective interactions exponentially increase with the increase in total genetic loci, leading to a number of challenges associated with high-dimensional data, as otherwise known as the curse of dimensionality (Chattopadhyay and Lu, 2019). Therefore, robust state-of-the-art computational techniques are required in order to identify and validate genetic interactions, eventually leading to a body of evidence that can explain a part of the current gap in heritability of complex traits and diseases.

Edited and reviewed by:
Simon Charles Heath,
Center for Genomic Regulation (CRG),
Spain

***Correspondence:**
Chuhsing Kate Hsiao
ckhsiao@ntu.edu.tw

Specialty section:
This article was submitted to
Statistical Genetics and Methodology,
a section of the journal
Frontiers in Genetics

Received: 24 February 2022
Accepted: 24 March 2022
Published: 14 April 2022

Citation:
Lu T-P, Kamatani Y, Belbin G, Park T and Hsiao CK (2022) Editorial: Current Status and Future Challenges of Biobank Data Analysis.
Front. Genet. 13:882611.
doi: 10.3389/fgene.2022.882611

High-throughput techniques have made available to researchers the genomic information from multiple platforms, such as DNA single nucleotide variants (SNVs), copy number variations (CNVs), and DNA methylation from SNP microarrays and/or short- or long-read next-generation sequencing platforms. Genetic susceptibility to a specific disease can be gauged better if combined multi-omics studies can be conducted for each patient/individual (Curtis et al., 2012; Brucker et al., 2020). Analysis of a single type of -omic data is limited by correlations primarily providing reactive signals rather than real causal ones, while integration of multiple -omics data types potentially may identify real causal changes that elucidate disease etiology or treatment targets (Hasin et al., 2017). Such findings can subsequently be tested for confirmation in molecular studies.

This Research Topic aimed to bring together studies that can showcase the potential and wealth of information that biobanks hold by providing a comprehensive overview of the current advances in scientific and clinical knowledge derived from biobank data analysis. Primarily, it aimed to compile algorithms and statistical models that report novel findings for different phenotypes, perform integrated analyses of multi-omics data in health and disease, and report sophisticated bioinformatics and statistical techniques with the aim of establishing an association between the genetic profiles and biological phenotypes. The second goal was to weigh the various difficulties and challenges posed involved in high-dimensional biobank data analysis. To this end, five studies were published under this Research Topic, as described in the following sections. Four studies were conducted on subjects from the Taiwan Biobank (TWB) (Wei et al., 2021), and the fifth was a review discussing the challenges related to biobank data analysis.

The TWB was established in 2012, with the aim of creating a population-based cohort of 200,000 adults, recruited at regular time intervals, with no cancer diagnosis at the time of enrollment (Juang et al., 2021). The general population of Taiwan constitutes individuals of Han-Chinese ancestry who immigrated from various provinces of China and local Taiwanese aboriginals. For genotype imputation purposes, the TWB additionally offers a reference panel created from whole-genome sequencing data from 1,445 early recruited participants. Furthermore, it offers two customized SNP genotyping arrays. Being the largest publicly available genetic database of individuals with East Asian ancestry, it helps document population-specific risk variants to improve the clinical care of the participants.

A study by Chen et al. described the association of a much-reported variant *ABCG2* rs2231142 and body mass index (BMI) with the risk of incidence of hyperuricemia (HUA), a major risk factor for gout, in East Asian populations. HUA is associated with various comorbidities, including obesity, hypertension, type 2 diabetes, hyperlipidemia, cardiovascular diseases, chronic kidney diseases, stroke, osteoporosis, erectile dysfunction, and obstructive sleep apnea, and has been reported as an independent predictor of premature mortality (Vincent et al., 2017; Singh and Cleveland, 2019; Huang et al., 2020). They conducted their study on 4,228 HUA patients from the TWB and reported a higher risk of HUA in association with either the “risk T” allele of *ABCG2* rs2231142 (TT or TG genotype) or higher BMI for both men and women. They further established the association of strong genetic-environmental (GxE) interaction with very high risk of HUA. Based on their findings, they recommended controlling body weight (i.e., lowering BMI) for patients with high risk of HUA carrying the *ABCG2* rs2231142 risk T allele.

Three studies were published under this Research Topic that conducted genomic analysis through proposed scalable methodologies to improve the relevance, utility and interpretability of the reported findings. One was by Chi et al., who proposed SEAGLE, a scalable exact algorithm for large-scale set-based gene-environment (G × E) tests on continuous traits, and applied it to subjects from the TWB. SEAGLE deploys matrix computations to calculate variance-component test statistics and p-values of G × E interactions. It requires no additional assumptions or approximations and is computationally efficient, with the ability to accommodate sample sizes up to the order of 10^5, thereby eliminating the requirement for high performing computing resources. Extensive simulation studies under different scenarios and assumptions were conducted to establish its scalability and comparable power and type I error rates.

The second study was by Yu et al., who proposed an integrative co-localization (INCO) approach for combining more than one type of -omic data. SNVs and CNVs from the same genomic unit were utilized for obtaining their concurrent effect and dealing with the sparsity of rare variations. Traditional integrated analyses of multi-omics data usually analyze each type of data separately, after which a naïve union or intersection analysis of significant findings is conducted to identify candidate genes. Such approaches may fall behind in identifying causal variants for traits that are a result of the concurrent effect of both the omics levels. INCO is a hybrid approach that conducts a screening procedure at the gene level, followed by modeling a concurrent effect from both levels of data, irrespective of whether each of them has a marginal association with the trait. Finally, it focuses on narrower genetic regions for bypassing the sparsity effect due to rare variants. Yu et al. conducted comprehensive simulations to demonstrate the scalability of their approach under different assumptions and scenarios and then applied their method to the study of subjects from the TWB, specifically their low-density lipoprotein cholesterol and triglyceride levels. They reported a potentially novel association of the *VNN2* gene, which is a protein coding gene involved in cell migration and fatty acid metabolism.

Finally, the third study was reported by Sun et al. and demonstrated that a novel quality control procedure can improve the accuracy of rare variant calling in SNP arrays. Detection of rare variants through genomic association studies still remains a challenge for various reasons—noisy signals and batch effects to name a few—and improvement in the genotyping quality may be an avenue for better clinical applications. Analyzing a custom Axiom array of data consisting of 267,247 rare variants obtained from 43,433 individuals in the TWB, an advanced normalization adjustment was adopted to prevent false calls caused by splitting the cluster, and a rare het adjustment was employed to lower false calls of rare variants. The concordance of the MAF for the called variants was measured by comparing it with that of the allelic frequencies from array data. Finally, genotyping results were used to detect familial hypercholesterolemia, thrombophilia, and maturity-onset diabetes of the young to assess the performance of their proposed procedure in disease screening. All heterozygous calls were verified by Sanger sequencing or qPCR, and the positive predictive value of each step was reported with an increase of up to 100%. Findings from this study demonstrated that correctly conducted genotype calling of rare variants could potentially be a solution for pathogenic variant detection through SNP arrays.

Lastly, a review was published by Bi and Lee, where they discussed challenges in multi-omics data analysis with the aim of aiding statisticians, epidemiologists, and other medical scientists in dealing with biobank-level data. They described in detail the current and future statistical and computational roadblocks that researchers stumble upon while performing GWASs (single point association or multilevel epistatic association) and PheWASs on large-scale biobank data; summarized recently developed scalable and robust regression approaches; and introduced Phewebs and some phenome-wide analysis results at the variant, gene, and pathway levels. They further outlined the need for more advanced methods and tools for handling future challenges and furnished comprehensive information for statisticians to obtain an up-to-date understanding of the tools and technologies at hand.

Based on these published studies in this Research Topic, we discussed some of the difficulties in analyzing massive amounts of genetic data. In addition to the methodological issues, another crucial challenge now is to combine biobanks worldwide for comprehensive studies to improve the equity of obtaining genetics data in human genome research. Since genomic research findings are often translated into genetic testing, disease diagnosis, and therapeutic solutions, especially in the era of personalized medicine, it is important that the scientific conclusions are not drawn from a biased sample to enhance population health globally (Harris and Sulston, 2004; Editorial, 2021). One way to do this would be to have a collective effort from researchers conducting studies with genetic databases such as biobanks around the world, so that diversity can be achieved. Several initiatives have started to reach this goal (The H3Africa Consortium, 2014; GenomeAsia100K Consortium, 2019; Robine and Varmus, 2021). We have not covered this in the current Research Topic, but we hope there will be one focusing on this issue in the near future.

In summary, biobanks play a central role in elucidating disease etiology and promoting public health. Incorporating biological, clinical, and genetic information into multi-omics analysis protocols is crucial to this end. Efficient use of big data by biobanks, both retrospectively and prospectively, would accelerate the implementation of preventive measures, optimized treatments, and personalized healthcare.

AUTHOR CONTRIBUTIONS

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

ACKNOWLEDGMENTS

We appreciate the input received from Dr. Amrita Chattopadhyay, who provided valuable opinions that helped to shape the manuscript.

REFERENCES

Bi, W., and Lee, S. (2021). Scalable and Robust Regression Methods for Phenome-wide Association Analysis on Large-Scale Biobank Data. *Front. Genet.* 12, 682638. doi:10.3389/fgene.2021.682638

Brucker, A., Lu, W., Marceau West, R., Yu, Q.-Y., Hsiao, C. K., Hsiao, T.-H., et al. (2020). Association Test Using Copy Number Profile Curves (CONCUR) Enhances Power in Rare Copy Number Variant Analysis. *Plos Comput. Biol.* 16, e1007797. doi:10.1371/journal.pcbi.1007797

Chattopadhyay, A., and Lu, T. P. (2019). Gene-gene Interaction: the Curse of Dimensionality. *Ann. Transl Med.* 7, 813. doi:10.21037/atm.2019.12.87

Chen, Y. J., Chen, I. C., Lin, H. J., Lin, Y. C., Chang, J. C., Chen, Y. M., et al. (2021). Association of ABCG2 Rs2231142 Allele and BMI with Hyperuricemia in an East Asian Population. *Front. Genet.* 12, 709887. doi:10.3389/fgene.2021.709887

Chi, J. T., Ipsen, I. C. F., Hsiao, T. H., Lin, C. H., Wang, L. S., Lee, W. P., et al. (2021). SEAGLE: A Scalable Exact Algorithm for Large-Scale Set-Based Gene-Environment Interaction Tests in Biobank Data. *Front. Genet.* 12, 710055. doi:10.3389/fgene.2021.710055

Coppola, L., Cianflone, A., Grimaldi, A. M., Incoronato, M., Bevilacqua, P., Messina, F., et al. (2019). Biobanking in Health Care: Evolution and Future Directions. *J. Transl Med.* 17, 172. doi:10.1186/s12967-019-1922-3

Curtis, C., Shah, S. P., Shah, S. P., Chin, S.-F., Turashvili, G., Rueda, O. M., et al. (2012). The Genomic and Transcriptomic Architecture of 2,000 Breast Tumours Reveals Novel Subgroups. *Nature* 486, 346–352. doi:10.1038/nature10983

Editorial (2021). Precision Medicine Needs an Equity Agenda. *Nat. Med.* 27, 737. doi:10.1038/s41591-021-01373-y

GenomeAsia100K Consortium (2019). The GenomeAsia 100K Project Enables Genetic Discoveries across Asia. *Nature* 576, 106–111. doi:10.1038/s41586-019-1793-z

Harris, J. R., Burton, P., Knoppers, B. M., Lindpaintner, K., Bledsoe, M., Brookes, A. J., et al. (2012). Toward a Roadmap in Global Biobanking for Health. *Eur. J. Hum. Genet.* 20, 1105–1111. doi:10.1038/ejhg.2012.96

Harris, J., and Sulston, J. (2004). Genetic Equity. *Nat. Rev. Genet.* 5, 796–800. doi:10.1038/nrg1454

Hasin, Y., Seldin, M., and Lusis, A. (2017). Multi-omics Approaches to Disease. *Genome Biol.* 18, 83–15. doi:10.1186/s13059-017-1215-1

Huang, W.-S., Lin, C.-L., Tsai, C.-H., and Chang, K.-H. (2020). Association of Gout with CAD and Effect of Antigout Therapy on CVD Risk Among Gout Patients. *J. Investig. Med.* 68, 972–979. doi:10.1136/jim-2019-001140

Jacobs, G., Wolf, A., Krawczak, M., and Lieb, W. (2018). Biobanks in the Era of Digital Medicine. *Clin. Pharmacol. Ther.* 103, 761–762. doi:10.1002/cpt.968

Jimmy Juang, J. M., Liu, Y. B., Julius Chen, C. Y., Yu, Q. Y., Chattopadhyay, A., Lin, L. Y., et al. (2020). Validation and Disease Risk Assessment of Previously Reported Genome-wide Genetic Variants Associated with Brugada Syndrome: SADS-TW BrS Registry. *Circ. Genom Precis Med.* 13, e002797. doi:10.1161/CIRCGEN.119.002797

Juang, J.-M. J., Lu, T.-P., Su, M.-W., Lin, C.-W., Yang, J.-H., Chu, H.-W., et al. (2021). Rare Variants Discovery by Extensive Whole-Genome Sequencing of the Han Chinese Population in Taiwan: Applications to Cardiovascular Medicine. *J. Adv. Res.* 30, 147–158. doi:10.1016/j.jare.2020.12.003

Liu, S., and Crawford, D. C. (2022). Maturation and Application of Phenome-wide Association Studies. *Trends Genet.* 38 (4), P353–P363. doi:10.1016/j.tig.2021.12.002

Robine, N., and Varmus, H. (2021). New York's Polyethnic-1000: a Regional Initiative to Understand How Diverse Ancestries Influence the Risk, Progression, and Treatment of Cancers. *Trends Cancer* 8 (4), 269–272. doi:10.1016/j.trecan.2021.11.005

Sakaue, S., Hirata, J., Kanai, M., Suzuki, K., Akiyama, M., Lai Too, C., et al. (2020). Dimensionality Reduction Reveals fine-scale Structure in the Japanese Population with Consequences for Polygenic Risk Prediction. *Nat. Commun.* 11, 1569. doi:10.1038/s41467-020-15194-z

Sanner, J. E., and Frazier, L. (2007). Factors that Influence Characteristics of Genetic Biobanks. *J. Nurs. Scholarship* 39, 25–29. doi:10.1111/j.1547-5069.2007.00139.x

Singh, J. A., and Cleveland, J. D. (2019). Gout Is Associated with a Higher Risk of Chronic Renal Disease in Older Adults: a Retrospective Cohort Study of U.S. Medicare Population. *BMC Nephrol.* 20, 93–97. doi:10.1186/s12882-019-1274-5

Sun, T. H., Shao, Y. J., Mao, C. L., Hung, M. N., Lo, Y. Y., Ko, T. M., et al. (2021). A Novel Quality-Control Procedure to Improve the Accuracy of Rare Variant Calling in SNP Arrays. *Front. Genet.* 12, 736390. doi:10.3389/fgene.2021.736390

The H3Africa Consortium (2014). Enabling the Genomic Revolution in Africa: H3Africa Is Developing Capacity for Health-Related Genomics Research in Africa. *Science* 344, 1346. doi:10.1126/science.1251546

Vincent, Z. L., Gamble, G., House, M., Knight, J., Horne, A., Taylor, W. J., et al. (2017). Predictors of Mortality in People with Recent-Onset Gout: a Prospective Observational Study. *J. Rheumatol.* 44, 368–373. doi:10.3899/jrheum.160596

Wei, C. Y., Yang, J. H., Yeh, E. C., Tsai, M. F., Kao, H. J., Lo, C. Z., et al. (2021). Genetic Profiles of 103,106 Individuals in the Taiwan Biobank Provide Insights into the Health and History of Han Chinese. *NPJ Genom Med.* 6, 10. doi:10.1038/s41525-021-00178-9

Yamamoto, K., Hachiya, T., Fukushima, A., Nakaya, N., Okayama, A., Tanno, K., et al. (2017). Population-based Biobank Participants' Preferences for Receiving Genetic Test Results. *J. Hum. Genet.* 62, 1037–1048. doi:10.1038/jhg.2017.81

Yu, Q.-Y., Lu, T.-P., Hsiao, T.-H., Lin, C.-H., Wu, C.-Y., Tzeng, J.-Y., et al. (2021). An Integrative Co-localization (INCO) Analysis for SNV and CNV Genomic Features with an Application to Taiwan Biobank Data. *Front. Genet.* 12, 709555. doi:10.3389/fgene.2021.709555

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.