Learning Lessons on Reproducibility and Replicability in Large Scale Genome-Wide Association Studies

Xihong Lin^{1,2}

¹Department of Statistics, Faculty of Arts and Sciences, Harvard University, Cambridge, Massachusetts, United States of America,

²Department of Biostatistics, Harvard T. H. Chan School of Public Health, Harvard University, Boston, Massachusetts, United States of America

Published on: Dec 16, 2020

DOI: <https://doi.org/10.1162/99608f92.33703976>

License: [Creative Commons Attribution 4.0 International License \(CC-BY 4.0\)](https://creativecommons.org/licenses/by/4.0/)

ABSTRACT

Reproducibility and replicability play a pivotal role in science. The article reflects on reproducibility and replicability as they figure in large scale genome-wide association studies. Overall, we emphasize the importance of enhancing data reproducibility, analysis reproducibility, and result replicability. We make recommendations pertaining to the development of study designs that address 1) batch effects and selection bias, 2) the incorporation of discrete discovery and replication phases, and 3) the procurement of a large sample size. We emphasize the importance of systematic and transparent data generation, processing, and quality control pipelines, as well as a rigorous field-specific standardized analysis protocol. We offer guidance with respect to collaborative frameworks, open access analysis tools, and software, and the use of supporting mandates, infrastructure, and repositories for data and resource sharing. Finally, we identify the role of incentives and culture in fueling the production of reproducible and replicable research through partnerships of researchers, funding agencies, and journals.

Keywords: analysis reproducibility and standardization, community culture building and collaborative framework, data reproducibility and standardization and harmonization, data repositories, result replicability, multi-phase study design

1. Introduction

Recently, the US National Academy of Sciences, Engineering, and Medicine published a comprehensive report on Reproducibility and Replicability in Science (National Academy of Sciences, 2019). The European Commission also published a scoping report on Reproducibility of Scientific Results in the EU (European Commission, 2020). These two reports remind us of the importance of reproducibility and replicability to the task of ensuring the validity of a new scientific discovery and trust in science.

Just by way of a refresher, reproducibility pertains to obtaining consistent results using the same data input and analytic methods and tools, while replicability pertains to obtaining consistent results across independent studies. In recent years, the scientific community has raised red flags concerning the risks of irreproducible results (Baker et al., 2016; Fanelli, 2018), and called for improvements in the rigor and reproducibility in research (Collins & Tabak, 2014; Redish et al., 2018) and, in particular, the practice of statistical significance using p-values (Wasserstein and Lazar, 2016). The need for concerted action is urgent, since “the lack of reproducibility of scientific studies has caused growing concern over the credibility of claims of new discoveries based on ‘statistically significant’ findings” (Benjamin et al., 2018).

Here, with a view toward addressing this pressing problem, we seek to share the lessons that we have learned about enhancing reproducibility and replicability in large scale Genome-Wide Association Studies (GWAS), and to make a few recommendations as well. We hope that these lessons are useful for advancing reproducible

and replicable science in emerging studies of whole genome sequencing and biobanks, as well as in other disciplines.

GWAS entails the analysis of hundreds of thousands to millions of common genetic variants across the genome, using data from large case-control studies and cohort studies to identify genetic variants associated with diseases and traits (Hirschhorn & Daly, 2005). Hundreds of GWAS in the last decade have led to the discovery of over 10,000 genetic variants associated with a wide range of common diseases and traits (Visscher et al., 2017). The findings in GWAS have much to teach us regarding the development of strategies to improve reproducibility and replicability across sciences.

Here, we discuss lessons learned from GWAS on data reproducibility, analysis reproducibility, and result replicability. We emphasize the importance of engaging the scientific community in collaboratively developing a culture centered around the practices of 1) validating and standardizing data generation, data processing, and protocol development; 2) testing and standardizing open sourcing analysis pipelines and software; 3) building and supporting infrastructure and repositories to allow for convenient and safe data and resource sharing, and 4) engaging researchers, funding agencies, and journals in collective efforts aimed at improving data and resource sharing, with a view toward the larger aim of promoting reproducible and replicable science.

2. Strategies for Enhancing Data Reproducibility

The importance of data reproducibility to reproducible and replicable science cannot be overstated. In GWAS, recognition of this importance informs efforts to generate and call robust genotype data. In the past ten years, the GWAS community has made significant inroads on the task of making genotype data reproducible by establishing community standards for genotype generation and calling, quality control protocols, and phenotype standardization in Electronic Health Record (EHR)-based biobanks, and collaborative frameworks (Thorisson et al., 2009; Laurie et al., 2012).

GWAS data arise from the carefully designed process of genotyping of tens of millions of genetic variants across the genome from hundreds of thousands of individuals who are themselves from many different cohorts. Genotyping is often performed at large genotyping centers, which have developed collaborative practices entailing open-access variant calling algorithms and pipelines that are tested and used to formulate community standards. Important issues, such as batch and center effects, are addressed in the formation of variant calling algorithms and quality control protocols (Laurie et al., 2012). These standardized QC protocols have been widely tested and disseminated and adopted by the GWAS community (Marees et al., 2018).

Phenotyping quality, standardization, and harmonization play a critical role in data and analysis reproducibility and result replicability (Thorisson et al., 2009). Examples of phenotype data include disease/trait outcomes, exposures, and treatment information, which are often collected from epidemiological studies and Electronic Health Records (EHRs). Compared to genotype data, phenotype data from sources such as EHRs are more complex, and pose challenges with respect to accuracy, harmonization and standard development (Pathak et al.,

2013). To address these challenges, substantial efforts have been made to develop community standards using phecode, which aggregates International Classification of Diseases, Ninth and Tenth Revisions (ICD-9 and ICD-10 codes) by clinical phenotypes for phenome-wide association studies (PheWAS) using EHRs (Wu et al., 2019).

Among the efforts to establish community standards for data and sharing, the [Global Alliance for Genomics and Health](#) (GA4GH), which was formed in 2013 as an international nonprofit alliance to “drive uptake of standards and frameworks for genomic data sharing within the research and healthcare communities,” stands out. The GA4GH proposed a principled and practical framework for the responsible sharing of genomic and health-related data by bringing together different stakeholders (Knoppers, 2014). International collaborative efforts to develop a framework for data standards and sharing represent our best chance to improve data exchange and governance, while strengthening the reproducibility and harmonization of both genotype and phenotype data.

3. Roles of Study Design in Reproducible and Replicable Science

Rigorous and well-documented study design is of critical importance for ensuring study validity, as well as enhancing reproducibility and replicability. Poorly designed studies are often to blame when it comes to causing difficulties in the replication of findings by other studies. In GWAS, examples of relevant design consideration factors include genotype data generation to minimize batch and center effects, phenotype data collection to minimize selection bias, the inclusion of distinct discovery and validation phases, and the procurement of large sample sizes through large international disease consortia.

For genotype data collection, the protocols for genotyping and sample allocation across genotyping centers need to be carefully planned to minimize batch and center effects, e.g., to balance the number of cases and the number of controls, as well as the ethnicities of cases and controls between batches and centers. Blocking and randomization work well in this context (Lambert & Black, 2012). Genotyping and batch and center bias can be further reduced by joint calling, using pooling data from different centers, followed by a carefully developed QC procedures (Regier et al., 2019; Taliun et al., 2020).

For phenotype data, the sampling schemes of study participants need to be carefully considered in the design phase, and also taken into account in the analysis phase. Selection bias requires particular attention in large-scale studies (Munafò et al., 2017). Indeed, relative to variance, bias plays a much more important role in studies involving big data.

Candidate gene studies often have small sample sizes and lack built-in replication studies; they use much higher type I error rates in declaring statistical significance. In the context of candidate gene studies, these limitations often result in false positives and difficulties in replicating findings. To address this challenge, and help with improving replicability, GWAS often use a very large sample size, which is achieved by forming

large national and international disease/trait-specific consortia, stringent type I error rates, and a multi-phase design.

A well-established convention of GWAS study design reflects the insight that replicability is enhanced through the use of both a discrete discovery phase and a replication phase. GWAS uses a stringent genome-wide statistical significance level for meta-analysis of the combined data to correct for a large number of tests across the genome, e.g., using the Bonferroni correction (Visscher et al., 2017). Given the large number of tests of genetic variants across the genomes, top hits are likely to be false positive. Hence, replicating these findings in independent samples is critical.

4. Enhancing Analysis Reproducibility and Result Replicability

Data, analysis and result reporting standards, coupled with open-access, well maintained and easy-to-use analysis software that perform standardized statistical and computational analysis in a field, play a critical role in reproducibility and replicability of scientific research. In GWAS, scientists collectively develop, test, and adopt community standards and software, rather than devising idiosyncratic approaches and applying them in a piecemeal fashion. This communal strategy has not only enhanced analysis reproducibility and result replicability, but also facilitated national and international collaboration in large GWAS consortia. Even though sharing genetic and phenotype data might not always be feasible for all study cohorts, with standardized analysis and open access software that implements these analyses, researchers from different cohorts working on a large collaborative GWAS study have demonstrated their ability to process data and perform analyses consistently and transparently. The community standards are tested and improved over years, as empirical evidence evolves. Just as importantly, through years of effort, the community has developed a culture of sharing cohort-specific analysis summary statistics. Adherence to these cultural norms are regulated by [the NIH policy](#) and facilitated by [the repository at the GWAS Catalog](#).

Pre-specified and standardized GWAS analysis protocols (Visscher et al., 2017) include QC procedures, statistical models and methods, incorporation of a stringent genome-wide significance level, and advanced planning of the studies to be used in the discovery phase and the replication phase, as well as a meta-analysis plan. Replication studies in GWAS are built-in, and their inclusion has become a standard practice in the GWAS field through communal efforts over the years. Indeed, at this point, it is difficult to publish a GWAS paper without replication studies or meta-analysis in top journals. It is also difficult to get GWAS grants funded without independent replication studies, as reviewers often expect such studies. The scientific need for replication studies and meta-analysis, and the community culture built around these needs has created strong incentives for national and international collaboration between researchers, as well as the formation of large disease/trait specific consortia. This phenomenon underlines the importance of multifaceted, collaborative efforts by researchers, journals and funding agencies to translate the norms of reproducibility and replicability into real world practices within and across scientific disciplines.

GWAS analysis methods that are empirically tested and standardized by the community include regression analysis using individual variants, evaluation of key confounders such as population structure using ancestry principal components, as well as the use of stringent Bonferroni criteria to adjust for multiple comparisons. All of these aim at reducing the chance of spurious associations and biases in estimated effect sizes. Incorporating domain knowledge into study design and data analysis is essential to enhance replicability of results. For example, ethnic differences between studies in the discovery phase and the replication phase could result in failure with respect to the replicability of findings.

5. The Role of Data, Resource and Tool Sharing, and Repositories in Promoting Reproducibility and Replicability in Science

Data sharing for biomedical research is key to progress in our understanding of human health. Some of the key impediments to performing reproducible and replicable research in the past included the absence of a culture of data and result sharing, as well as the limited sharing of open-access software and code. Data sharing in the GWAS community has been a major—indeed, a normative—enabling factor in numerous gene mapping successes, partially because of the mandates of funding agencies, such as the NIH, using [the NIH platform dbGAP](#). Several NIH Data Commons, such as [ANVIL by the National Human Genome Research Institute](#), and [BioData Catalyst by the National Heart, Lung and Blood Institute](#), have recently been developed to facilitate broad data sharing, while promoting the use of FAIR – Findable, Accessible, Interoperable, and Reusable (Wilkinson et al., 2016) norms for dealing with data.

Also important in this context has been the influence of GA4GH, which proposed a comprehensive framework of ethical governance, consent, privacy, and security (Knoppers, 2014), reflecting the urgent need to ensure individual-level data privacy while promoting data sharing, through regulation and safe federal and organizational repositories. NIH dbGAP, Data Commons, and UK Biobank (Sudlow et al., 2015) provide good secured data repository models. In addition, to facilitate future research projects, the scope of which cannot be specified at the time of biosample collection, blank or broad consent, such as General Research Use GRU and Health/Medical/Biomedical research (HMB) consents, have proved to be most valuable (Lunshof et al., 2008; Dyke et al., 2016).

Reproducible and replicable research relies on the availability of open access, easy-to-use, and comprehensive analysis software that implements standardized analysis protocols and tools in a field (Purcell, et al, 2007). Such software needs to be tested, validated, well-maintained and supported, and widely adopted by the target research community. For example, [Plink](#) has been widely used by researchers to process and analyze GWAS data (Purcell et al., 2017). It contains comprehensive from-start-to-end analytic GWAS tools. It reads genotype data that are generated from commonly used genotyping arrays, performs QC, calculates ancestry PCs, performs association analysis, and vitalizes results.

Substantial efforts have also been made to extract and curate published replicated GWAS findings. The [GWAS Catalog](#) provides and maintains a consistent, easy-to-use, and freely available database of published significant disease/trait-genetic variant associations, including the association analysis, summary statistics for both the discovery and replication phases, as well as genome-wide GWAS summary statistics (MacArthur et al., 2016).

6. Discussions and Recommendations

Reproducible and replicable research is important for the success of scientific discovery, especially in dealing with the special challenges posed by massive data. The history of GWAS provides the research community with several valuable lessons on the strategies for promoting data reproducibility, analysis reproducibility and result replicability.

We highlight several recommendations. First, scientific communities should develop rigorous study designs by considering the key factors that affect reproducibility and replicability, such as 1) batch effects and selection bias; 2) the need to build discrete discovery and replication phases; and 3) the procurement of a large sample size through forming large international research consortia.

Second, scientific communities should develop systematic and transparent data generation and processing pipelines, a rigorous and empirically tested statistical analysis protocol, field-specific community data and analysis standards, and especially, collaborative frameworks, as well as open access analysis tools and software. Examples of such efforts include consistency of data generation, data harmonization, the development of standardized QC and data processing pipelines, and standardized analysis protocols that are empirically evaluated and tested, as well as open access, cohesive, and high quality software packages for standardized analyses.

Third, scientific communities should establish mandates for secured data and resource sharing and regulation by funding agencies that support centralized well-maintained research data infrastructure and repositories, such as the NIH dbGAP and Data Commons, which meet the desired FAIR principles and standards. In addition, scientific communities should develop data sharing, data privacy, security and governance policies and guidelines. Broad consents that do not restrict the scope of research, and allow for the use of biosamples and clinical information in future research projects are to be encouraged.

Fourth, scientific communities should build research incentives and a communal culture for reproducible and replicable research by supporting partnerships between researchers, funding agencies, and journals. This effort should entail collaboratively developing a culture and tradition of standardizing data generation, processing and protocol development, and standardizing analysis pipelines and software in a field, making data and resource sharing easy, and well-supported by regulations and safe repositories.

Finally, to assure a future of sustainable, reproducible, and replicable science, we need to encourage deeper discussions of issues, culture, practices, and solutions related to reproducibility and replicability within and

across our scientific communities. The pivotal role of statistics and data science in this endeavor should be emphasized. Quantitative scientists and domain scientists, as well as funding agencies, journals, academia and private sectors, need to work together to encourage and take actions regarding data sharing and, more broadly, the adoption of best data and analytic practices and available tools. With such joint communal efforts, we can accelerate the progress of open and reproducible and replicable science, and improve the accuracy and the depth of scientific discovery.

Acknowledgments

The author thanks the editor and the reviewers for their helpful comments that have improved the paper.

Disclosure Statement

This research was funded by the grants from the National Institute of Health R35-CA197449, U01-HG009088, U19-CA203654.

References

- Baker, M. (2016). Reproducibility crisis? *Nature*, 533(7604), 452–454. <https://doi.org/10.1038/533452a>
- Benjamin, B. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., Bollen, K. A., Brembs, B., Brown, L., Camerer, C., Cesarini, D., Chambers, C. D., Clyde, M., Cook, T. D., Boeck, P. D., Dienes, Z., Dreber, A., Easwaran, K., Efferson, C., . . . Johnson, V. E. (2018). Redefine statistical significance. *Nature Human Behaviour*, 2(1), 6–10. <https://doi.org/10.1038/s41562-017-0189-z>
- Collins, F. S., & Tabak, L. A. (2014). NIH plans to enhance reproducibility. *Nature*, 505(7485), 612–613. <https://doi.org/10.1038/505612a>
- Dyke, S. O., Philippakis, A. A., Rambla De Argila, J., Paltoo, D. N., Luetkemeier, E. S., Knoppers, B. M., Brookes, A. J., Spalding, J. D., Thompson, M., Roos, M. & Boycott, K. M., Brudno, M., Hurles, M., Rehm, H. L., Matern, A., Fiume, M., & Sherry, S. T. (2016). Consent codes: Upholding standard data use conditions. *PLoS Genetics*, 12(1), Article e1005772. <https://doi.org/10.1371/journal.pgen.1005772>
- European Commission. (2020). *Reproducibility of Scientific Results in the EU: A Scoping Report*. European Union, Luxembourg
- Fanelli, D. (2018). Opinion: Is science really facing a reproducibility crisis, and do we need it to? *Proceedings of the National Academy of Sciences*, 115(11), 2628–2631. <https://doi.org/10.1073/pnas.1708272114>
- Hirschhorn, J. N., & Daly, M. J. (2005). Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics*, 6(2), 95–108. <https://doi.org/10.1038/nrg1521>

- Knoppers, B. M. (2014). Framework for responsible sharing of genomic and health-related data. *The HUGO Journal*, 8(1), Article 3. <https://doi.org/10.1186/s11568-014-0003-1>
- Lambert, C. G. & Black, L. J. (2012). Learning from our GWAS mistakes: From experimental design to scientific method. *Biostatistics*, 13(2), 195–203. <https://doi.org/10.1093/biostatistics/kxr055>
- Laurie, C. C., Doheny, K. F., Mirel, D. B., Pugh, E. W., Bierut, L. J., Bhangale, T., Boehm, F., Caporaso, N. E., Cornelis, M. C., Edenberg, H. J., Gabriel, S. B., Harris, E. L., Hu, F. B., Jacobs, K. B., Kraft, P., Landi, M. T., Lumley, T., Manolio, T. A., McHugh, C., . . . GENEVA Investigators (2010). Quality control and quality assurance in genotypic data for genome-wide association studies. *Genetic Epidemiology*, 34(6), 591–602. <https://doi.org/10.1002/gepi.20516>
- Lunshof, J. E., Chadwick, R., Vorhaus, D. B. & Church, G. M. (2008). From genetic privacy to open consent. *Nature Review Genetics*, 9(5), 406–411. <https://doi.org/10.1038/nrg2360>
- MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., Junkins, H., McMahon, A., Milano, A., Morales, J. & Pendlington, Z. M., Welter, D., Burdett, T., Hindorff, L., Flicek, P., Cunningham, F., & Parkinson, H. (2016). The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Research*, 45(D1), D896–D901. <https://doi.org/10.1093/nar/gkw1133>
- Marees, A. T., de Kluiver, H., Stringer, S., Vorspan, F., Curis, E., Marie-Claire, C., & Derks, E. M. (2018). A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. *International Journal of Methods in Psychiatric Research*, 27(2), Article e1608. <https://doi.org/10.1002/mpr.1608>
- Munafò, M. R., Tilling, K., Taylor, A. E., Evans, D. M., & Davey Smith, G. (2018). Collider scope: When selection bias can substantially influence observed associations. *International Journal of Epidemiology*, 47(1), 226–235. <https://doi.org/10.1093/ije/dyx206>
- National Academies of Sciences, Engineering, and Medicine. (2019). *Reproducibility and replicability in science*. National Academies Press. <https://doi.org/10.17226/25303>
- Pathak, J., Kho, A. N., & Denny, J. C. (2013). Electronic health records-driven phenotyping: Challenges, recent advances, and perspectives. *Journal of the American Medical Informatics Association*, 20(2), e206–e211. <https://doi.org/10.1136/amiajnl-2013-002428>
- Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., De Bakker, P. I., Daly, M. J. & Sham, P. C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 81(3), 559–575. <https://doi.org/10.1086/519795>

Redish, A. D., Kummerfeld, E., Morris, R. L. & Love, A. C. (2018). Opinion: Reproducibility failures are essential to scientific inquiry. *Proceedings of the National Academy of Sciences*, 115(20), 5042–5046.

<https://doi.org/10.1073/pnas.1806370115>

Regier, A. A., Farjoun, Y., Larson, D. E., Krasheninina, O., Kang, H. M., Howrigan, D. P., Chen, B. J., Kher, M., Banks, E., Ames, D. C. & English, A. C., Li, H., Xing, J., Zhang, Y., Matise, T., Abecasis, G. R., Salerno, W., Zody, M. C., Neale, B. M., & Hall, I. M. (2018). Functional equivalence of genome sequencing analysis pipelines enables harmonized variant calling across human genetics projects. *Nature Communications*, 9(1), Article 4038. <https://doi.org/10.1038/s41467-018-06159-4>

Thorisson, G. A., Muilu, J. & Brookes, A. J. (2009). Genotype–phenotype databases: challenges and solutions for the post-genomic era. *Nature Reviews Genetics*, 10(1), 9–18. <https://doi.org/10.1038/nrg2483>

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M. and Liu, B., Matthews, P., Ong, G., Pell J., Silman, A., Young, A., Sprosen T., Peakman, T., & Collins, R. (2015). UK biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med*, 12(3), Article e1001779. <https://doi.org/10.1371/journal.pmed.1001779>

Visscher, P.M., Wray, N.R., Zhang, Q., Sklar, P., McCarthy, M.I., Brown, M.A. & Yang, J. (2017). 10 years of GWAS discovery: Biology, function, and translation. *The American Journal of Human Genetics*, 101(1), 5–22. <https://doi.org/10.1016/j.ajhg.2017.06.005>

Wasserstein, R.L. & Lazar, N.A. (2016). The ASA’s statement on p-values: Context, process, and purpose. *The American Statistician*, 70(2), 129–133. <https://doi.org/10.1080/00031305.2016.1154108>

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J. W., da Silva Santos, L. B., Bourne, P. E. & Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., . . . Mons, B. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 3(1), 1–9. <https://doi.org/10.1038/sdata.2016.18>

Wu, P., Gifford, A., Meng, X., Li, X., Campbell, H., Varley, T., Zhao, J., Carroll, R., Bastarache, L., Denny, J.C. & Theodoratou, E. (2019). Mapping ICD-10 and ICD-10-CM codes to phecodes: workflow development and initial evaluation. *JMIR Medical Informatics*, 7(4), Article e14325. <https://doi.org/10.2196/14325>

©2020 Xihong Lin. This article is licensed under a Creative Commons Attribution (CC BY 4.0) [International license](https://creativecommons.org/licenses/by/4.0/), except where otherwise indicated with respect to particular material included in the article.