This paper was commissioned for the Committee on Reproducibility and Replicability in Science, whose work was supported by the National Science Foundation and the Alfred P. Sloan Foundation. Opinions and statements included in the paper are solely those of the individual author, and are not necessarily adopted, endorsed, or verified as accurate by the Committee on Reproducibility and Replicability in Science or the National Academies of Sciences, Engineering, and Medicine.

Reproducibility and Replicability in Large Scale Genetic Studies

Xihong Lin
Department of Biostatistics and Department of Statistics
Harvard University
xlin@hsph.harvard.edu

1. Introduction

Hundreds of Genome-Wide Association Studies (GWAS) in the last decade have successfully led to the discovery of over 12,000 genetic variants that are associated with a wide range of common diseases and traits (Visscher, et al, 2017). GWAS involve analysis of hundreds of thousands to millions of common genetic variants across the genome using data from large cohorts of individuals, commonly from case-control studies and cohort studies, to identify genetic variants associated with a disease or a trait of interest. The published genetic variant-disease association findings have been systematically and expertly curated in the last decade in the NHGRI-EBI GWAS Catalog (www.ebi.ac.uk/gwas), which provides a rich, quality-controlled community resource widely used by many researchers (MacArthur, et al, 2016). Overall, GWAS have represented a sweeping tour de force in human genetics, and paved a road for emerging Whole Genome Sequencing studies and biobank studies to advance precision health.

Through the collective efforts of the scientific community, findings of large scale GWAS have largely been found to be reproducible and replicable. The ability to reproduce data and results is at the heart of science. In recent years, a high-level concern about irreproducible results has been raised by the scientific community (Baker, et al, 2016; Fanelli, 2018), which has called for more strategies to enhance rigor and reproducibility

in research (Collins and Tabak, 2014; Redish, et al, 2018). These calls were followed by a surge of discussions in both the statistical community and the general scientific community on improving the practice of statistical significance using p-values (Amrhein and Greenland, 2018; Benjamin, et al, 2018; Wasserstein and Lazar, 2016). In view of these concerns, the lessons learned from GWAS are particularly valuable to facilitate efforts to develop strategies to improve reproducibility and replicability.

Indeed, in the last decade, GWAS have identified replicable common genetic variants associated with many common human diseases and traits. Examples of common disease genetic discoveries include cardiovascular disease, type 2 diabetes, inflammatory bowel disease, COPD, and cancers. Examples of quantitative traits include lipids, BMI, height, and lung function measures.

The success of GWAS in reproducing and replicating disease-associated variant discoveries is due to multiple factors. We highlight a few of them in this article. Specifically, we will discuss the strategies 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 of standardizing data generation, data processing and protocol development, as well as standardizing analysis pipelines and software, and making data and resource sharing feasible and well-supported. The community involved in these collective efforts includes quantitative and domain science researchers, funding agencies, and publishers.

Several key factors have contributed to the success of reproducibility and replicability in GWAS. They include (i) consistency in data generation and extensive quality control steps to ensure reliability of genotype data; (ii) genotype and phenotype harmonization; (iii) a push for large sample sizes through the establishment of large international disease consortia; (iv) rigorous study design and standardized statistical analysis protocols, including consensus building on controlling for key confounders such as genetic ancestry/population stratification, the use of stringent criteria to account for multiple testing, and the development of norms for conducting independent replication studies and meta-analyzing multiple cohorts; (v) a culture of large-scale international

collaboration, and sharing of data, results and tools, empowered by strong infrastructure support; (vi) an incentive system, which is created to meet scientific needs and is recognized and promoted by funding agencies, journals as well as grant and paper reviewers, for scientists to perform reproducible, replicable and accurate research.

The successful GWAS model used for data generation and analytic practice along with the culture of collaboration and data, results and tool sharing have facilitated the development of strategies for enhancing reproducibility and replicability of emerging state-of-the-art Whole Genome/Exome Sequencing (WGS/WES) studies and biobank studies, such as the two ongoing large national Whole Genome Sequencing programs include the Genome Sequencing Program (GSP) of the National Human Genome Research Institute (http://gsp-hg.org/), and the Trans-Omics for Precision Medicine Program (TOPMed) of the National Heart, Lung and Blood Institute (https://nhlbiwgs.org/). GSP and TOPMed together plan to do whole genome/exome sequencing of a total of 300,000-350,000 individuals, with the goal of studying the genetic variants and genes associated with many different diseases and traits. In this paper, some initial efforts of data reproducibility and analysis reproducibility of GSP and TOPMed will be discussed,

In-depth discussions of issues, challenges, solutions, practices and culture are valuable for addressing reproducibility and replicability in emerging large scale precision health studies, such as the All of Us Program of the National Institute of Health (NIH) (https://allofus.nih.gov/), the rapidly increasing number of biobank studies, as well as studies in other disciplines.

2. Data Reproducibility

Research data are a fundamental building block of science. Indeed, data reproducibility is a foundation for reproducible and replicable science. GWAS have focused on making genotype data reproducible by establishing community standards for genotype data, quality control protocols, and collaborative frameworks. GWAS data involve genotyping tens of millions of genetic variants across the genome from hundreds

of thousands of individuals from many different cohorts in a GWAS study. Genotyping is often performed at large genotyping centers, which have developed a tradition of collaboration in developing open-access variant calling algorithms and pipelines that have been tested and become community standards. Issues such as batch and center effects are commonly investigated and addressed in the variant calling algorithms. Extensive and transparent quality control (QC) steps have been developed to ensure the reliability of genotyping data. These standard QC protocols have been widely adopted by the GWAS community and are disseminated through training modules, online educational materials, short courses, and publications.

As the field transitions to the Whole Genome Sequencing (WGS) era, the GWAS practice of ensuring genotype data reproducibility and accuracy is being advanced in ongoing large scale WGS studies of a wide range of diseases and traits across diverse populations, such as those of GSP and TOPMed. There is a strong interest in jointly analyzing genomes across many cohorts and studies in order to boost power for disease/trait mapping, as well as studying population genetics and genome biology. To facilitate genome aggregation across centers and promote reliable joint analysis, five major US genome sequencing centers forged a collaboration to jointly develop WGS data processing and file format standards by harmonizing upstream data processing steps, while allowing for different variant callers (Regier, et al, 2018). These standards are made publicly available in Github. They provide guidelines for ongoing and future sequencing studies. The genome centers apply center-specific alignment, data processing and variant callers to the same testing data sets to demonstrate "functionally equivalent" (FE) results. This exercise allows FE results produced by different centers to be used for joint variant calling with minimal batch effects, helping lay the foundation for broad data sharing and joint analysis in large scale human genetics studies.

Phenotype quality, standardization and harmonization play an equally critical role in data and analysis reproducibility and result replicability. Examples of phenotype data include disease/trait outcomes, exposures, and treatment information, which are often collected from epidemiological studies, Electronic Medical Records (EMRs) and other sources. Compared to genotype data, phenotype data from sources such as EMRs are

more complex, and their accuracy, harmonization and standard development are more challenging. Substantial efforts have been made by TOPMed on harmonizing phenotypes across cohorts.

Recognizing the importance of data standards, the Global Alliance for Genomics and Health (GA4GH, *https://www.ga4gh.org/*) was formed in 2013 as an international nonprofit alliance to create frameworks and standards to "drive uptake of standards and frameworks for genomic data sharing within the research and healthcare communities." Such an international effort on collaborative framework and data standard building and guideline development strengthens reproducibility and harmonization of both genotype data and phenotype data, and helps integrate research and medical genomes worldwide, while embodying the concept of preproducibility (Stark, 2018), which means "research has been described in adequate detail for others to undertake it." A similar effort on developing data standardization and harmonization is being made by the Global Genomic Medicine Collaborative (G2MC**,** *www.g2mc.org***)**, which aims at recognizing and harnessing activities related to the global implementation of genomic medicine.

3. Roles of Study Design in Reproducibility and Replicability

Rigorous and well-documented study design is of critical importance for ensuring study validity as well as enhancing reproducibility and replicability. Poorly designed studies can cause difficulties in replicating findings in other studies. Researchers need to carefully consider the factors that impact data reproducibility and result replicability in all dimensions of study design. In GWAS and WGS studies, examples of these design consideration factors include genotype data generation to minimize batch and center effects, phenotype data collection to minimize selection bias, as well as inclusion of both a discovery phase and a validation phase, and the procurement of large sample sizes through large international disease consortia.

For genotype data collection, genotyping and sequencing protocols and sample allocation across sequencing centers need to be carefully planned to minimize batch and center effects. For example, there is a need to balance the number of cases and

the number of controls and the ethnicities of cases and controls between batches and centers. It is considered undesirable practice to sequence mainly cases in one center and mainly controls in another center; to sequence mainly Caucasians in one center and African Americans in another center; and or to sequence cases using one platform and controls using another platform, as these kinds of sample allocation can lead to genotyping bias, which may bias analysis results. Genotyping bias can be reduced by joint calling using pooled genotype data from different sequencing center centers followed by a carefully developed QC procedure to reduce remaining batch and center effects.

For phenotype data, commonly used epidemiological and clinical study designs include case-control, cross-sectional, and cohort studies. Sampling schemes of study participants needs to be carefully considered in the design phase and taken into account in the analysis phase as well. Selection bias requires particular attention in large-scale studies. Indeed, although variance is important, bias plays a more important role in studies involving big data.

It is critical to minimize the selection bias at the design phase of a study. For example, in GWAS and Whole Genome Sequencing association studies, differential population structures between cases and controls can result in bias in association analysis. For example, if disease cases use Caucasians in the US, while controls use Caucasians from the open-access UK Biobank (Sudlow, et al, 2015), cases and controls are likely to differ in ancestry, which will confound association analysis results.

A critical factor that contributes to the success of GWAS discovery is the community convention of building both a discovery phase and a replication phase in a GWAS study. Samples from independent studies are used in the replication phase to replicate the findings of the variants with the strongest evidence of association. In addition, a stringent genome-wide statistical significance level for meta-analysis of the combined data is used to correct for a large number of tests across the genome, e.g., using the Bonferroni correction p-value<10^{-8} for a million genetic variants.

One reason the GWAS community has widely adopted this important practice of including both the discovery phase and the replication phase in a GWAS study is the recognition that common disease-associated genetic variant effects are often weak and a large number of hypothesis tests of genetic markers across the genome are likely to result in the very top hits to be false positives. Hence, selection of a set of variants that are pushed for the replication phase often uses a less stringent genome-wide significance criterion at the discovery phase, in consideration of weak disease/trait-associated common variant effects and the fact that some null variants can happen to be observed for a marginal association attaining genome-wide significance by chance.

Another important design-related factor that contributes to the success of replicability of GWAS findings is that GWAS often involves a very large sample size, which is achieved by forming large national and international disease/trait-specific GWAS consortia. This tradition of conducting a large scale GWAS study has been developed because the common genetic variants associated with diseases/traits often have weak effects and large sample sizes are needed to reach a stringent genome-wide significant threshold and to detect weak effects. Candidate gene studies often have small sample sizes, use much higher type I error rates, and lack built-in replication studies. These limitations of candidate gene studies often result in false positives and difficulties in replicating the findings of candidate gene studies.

4. **Analysis Reproducibility**

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, these standards and software have not only contributed to the success of analysis reproducibility and result replicability, but also facilitated national and international collaboration in large GWAS consortia. These standards have been collectively developed and widely adopted by the GWAS community. Even though

sharing genetic and phenotype data might not always be feasible for all study cohorts, with these standardized analysis and commonly used open access software that implement these analyses, researchers of different cohorts of a GWAS study are able to process data and perform analyses consistently and transparently. They can also share cohort-specific analysis summary statistics, which can be easily used for meta-analysis in both the discovery phase and the replication phase. The efforts related to standard building have paid off in GWAS, and set up a good model for data and analysis standard development in emerging Whole Genome Sequencing studies and biobank studies. Consistent and rigorous standards of data analysis and result reporting, and powerful software play an instrumental role in clinical trials.

GWAS analysis protocols are pre-specified and standardized. These include the QC procedure, statistical models and methods, incorporation of the stringent genome-wide significance level, and advance 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. Indeed, it is difficult to publish a GWAS paper without replication studies or meta-analysis in top journals, such as Nature Genetics. It is also difficult to get GWAS grants funded without independent replication, as reviewers often have such an expectation. The scientific need and the community culture of conducting replication studies or meta-analysis to ensure result replicability create strong incentives for national and international collaboration between researchers and the formation of large consortia. This also underlines the importance of the multifaceted efforts by researchers, journals and funding agencies to make reproducibility and replicability of scientific research a real world practice.

GWAS analysis methods are tested and standardized by the community. They include regression analysis using individual variants, careful evaluation of key confounders including the genetic ancestry of the samples, and adjustment for population structure in regression analysis using principal components, as well as other covariates such as age and gender, as well as the use of stringent Bonferroni criteria to

adjust for multiple comparison. All of these are considered to avoid spurious associations and biases in estimated effect sizes. GWAS are increasingly performed through large meta-analyses that combine statistical evidence from multiple cohorts.

Incorporating domain knowledge in study design and data analysis is also essential to enhance replicability of results. For example, disease sub-types play a critical role in precision health. Genetic bases of disease sub-types may be different. A lack of consideration of disease sub-types in GWAS might result in GWAS results not being replicable. For example, lung cancer has several subtypes, such as adenocarcinoma, squamous cell carcinoma, and small cell carcinoma. Inflammation and immune-related genes are mainly associated with the risk of lung squamous cell carcinoma, while not much associated with the risk of lung adenocarcinoma. If the studies of a discovery phase mainly consist of adenocarcinoma cases while a replication phase consists of mainly squamous cell carcinoma cases, the results of the GWAS results from the discovery phase are likely not to be replicated. In addition, careful consideration of ethnicity is needed in GWAS. Ethnic differences between studies in the discovery phase and the replication phase could result in a failure in replicability of findings. In addition, different strategies for handling ethnicities in meta-analysis are desirable, such as performing ethnicity-specific meta-analysis, and across ethnicity meta-analysis.

5. Data, Resource and Tool Sharing

Some of the key impediments to performing reproducible and replicable research in the past included the lack of a culture of data sharing, result sharing, few tools to make it easy to share data and results, and limited sharing of open-access software and code used for performing analyses in published papers. Data sharing in the GWAS community has been a major enabling factor in gene mapping success. GWAS data sharing has become a norm due to the mandates of funding agencies, including the National Institute of Health. Systematic and secured data sharing is made feasible and convenient through dbGAP (*https://www.ncbi.nlm.nih.gov/gap*) that has been built and supported by the National Center for Biotechnology Information (NCBI). Researchers

can file an application for accessing the individual level GWAS data collected by the studies funded by NIH by following the NIH data security policy. Each application is subject to review and approval by the dgGAP. If approved, the downloaded data need to be securely stored by following the NIH guidelines. NCBI provides technical support for data sharing and access through dbGAP.

The NIH Data Commons (*https://commonfund.nih.gov/commons*) is being developed to make broad data sharing possible by the general health science research community and meet the community needs as well as standards for being FAIR – Findable, Accessible, Interoperable, and Reusable. The pilot projects of the NIH Data Commons Pilot Phase Consortium are developing and testing a cloud-based platform using three high profile datasets, including the TOPMed data, where health science investigators can store, share and access data and software. These pilot projects will help setting policies, processes, and architecture for the NIH Data Commons.

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 the field of interest. Such software needs to be well maintained and supported and widely adopted by the target research community. For example, Plink (*http://zzz.bwh.harvard.edu/plink/*) has been widely used by researchers to process and analyze GWAS data. It contains comprehensive from-start-to-end analytic tools needed for GWAS analysis. It reads genotype data that are generated from commonly used genotyping arrays, performs QC, calculates PCs, performs association analysis, and displays results with easy vitalization.

To facilitate open access data and result sharing, substantial efforts have also been made to extract and curate published replicated GWAS findings. The GWAS Catalog, a collaborative effort of the European Bioinformatics Institute (EMBL-EBI) and the National Human Genome Research Institute (NHGRI), provides and maintains a consistent and easy-to-use freely available database of published significant disease/trait-genetic variant associations. For each curated GWAS study, the

association analysis summary statistics for each significant SNP, including regression coefficients, standard errors and p-values, are reported for both the discovery phase and the replication phase, as well as meta-analysis results. Study sample characteristics, such as sample sizes and ethnicities for both the discovery phase and the replication phase are also reported. The Catalog also publishes a regularly updated GWAS diagram of SNP-trait associations, mapped onto the human genome by chromosomal location and displayed on the human karyotype.

In addition, GWAS result summary statistics, such as effect sizes, their standard errors, and p values of millions of SNPs, of an increasing number of large scale GWAS studies of a wide range of diseases and traits, have become rapidly widely available in the public domain in the last few years, e.g., see *https://grasp.nhlbi.nih.gov/FullResults.aspx*. The pubic availability of such SNP-level analysis result summary statistics has enabled discoveries of novel associations, estimation of heritability, quantification of pleiotropy across diseases/traits, construction of polygenetic risk prediction scores, and functional analysis of discovered disease-associated variants, without creating the need to access the original GWAS data.

As we move into the sequencing and precision health era, Whole Genome Sequencing (WGS) and Whole Exome Sequencing (WES) data are starting to become publicly available. For example, the WGS data of the first 8000 individuals sequenced by TOPMed are available in dbGAP and available for the general research community to access. As biobanks have rapidly emerged, large biobank data have also become publicly available for research purpose. For example, the UK Biobank (Sudlow, et, al 2015) has released both GWAS and rich phenotypic data, such as EMRs, life styles, imaging, and treatment data on 500,000 individuals to the international research community.

6. Discussions

Reproducible and replicable research has become increasingly important for the success of scientific discovery, especially in dealing with massive data. Advancing

reproducible and replicable science faces numerous significant challenges. GWAS provide the research community with valuable lessons as the community collectively develops strategic plans for advancing reproducible and replicable research in a wide range of scientific disciplines.

As demonstrated in GWAS, multiple components need to be considered to make scientific research more reproducible and replicable. They include data reproducibility; analysis reproducibility and result replicability. Specifically, first, a rigorous design of a study needs to consider the key factors that affect reproducibility and replicability, such as batch effects and selection bias; building both the discovery phase and the replication phase, with a procurement of a large sample size through forming large international research consortia.

Second, systematic and transparent data generation and processing pipelines and a rigorous statistical analysis protocol, as well as development of field-specific data and analysis standards and collaborative framework, as well as open access analysis tools and software, are pivotal. These include efforts on data generation consistency and harmonization; development of standardized QC and data processing pipelines, and rigorous standardized statistical analysis that properly addresses the key analytic issues in the field, as well stringent statistical inference procedures, such as stringent p-values to adjust for multiple comparison and adjustment for key confounders and design-related matters. These standards need to be empirically evaluated and tested. Efforts need to be organized for building these data and analysis pipelines and standards, as well as ensuring them to be up-to-date and widely followed by the research community. Open access, cohensive and high quality software packages that allow for version control and can be hosted in open access development platforms that are easy for repositories, such as Github, are critical and need to implement these up-to-date standardized data process and statistical analysis protocols.

Third, mandates for data and result sharing by funding agencies play an important role in reproducible and replicable science. These mandates need to be supported by

centralized well-maintained research data infrastructure, such as NIH dbGAP and Data Commons, and meet the desired principles and standards, such as FAIR introduced in the NIH Data Science Strategic Plans (*https://datascience.nih.gov/strategicplan*). Data sharing and data security policies and guidelines need to be developed.

Fourth, it is critical to build research incentives and community culture for reproducible and replicable research jointly by researchers, funding agencies and journals. Indeed, the scientific community in a discipline needs to be partnered by being engaged in collaboratively developing a culture and tradition of standardizing data generation, data processing and protocol development, as well as standardizing analysis pipelines and software; tailoring these towards the discipline of interest, and making data and resource sharing feasible and well-supported. More education for the research community on the practices of successful use cases of reproducible and replicable science and their benefits will be desirable.

Finally, to assure a future of sustainable, reproducible, and replicable science, we need to increase community efforts to have deeper discussions of issues, culture, practices and solutions related to reproducibility and replicability. In addition, strategic multi-faceted actions need to be taken to extend the basic principles and strategies that have been shown to work in successful use cases, such as GWAS, to other fields of research , and tailor them towards individual fields. The pivotal role of statistics and data science in this journey should continue to be emphasized. Quantitative scientists and domain scientists, as well as funding agencies and private sectors, need to work together to encourage and take actions on data sharing and the adoption of best data and analytic practices and available tools. With such joint communicty efforts, we can accelerate open reproducible and replicable science and improve the accuracy of scientific discovery.

Acknowledgement

This research was funded by the National Academies of Sciences, Engineering, and Medicine study on reproducibility, the grants from the National Institute of Health

R35-CA197449, P01-CA134294, U01-HG009088, U19-CA203654, and R01-HL113338. The author thanks the committee for helpful comments.

References

- Amrhein, V., and Greenland, S. (2018) "Remove, Rather Than Redefine, Statistical Significance," *Nature Human Behavior,* 2, p4.

- Baker, M., 2016. REPRODUCIBILITY CRISIS?. *Nature, 533, p.26.*

- Benjamin, D.J., Berger, J.O., Johannesson, M., Nosek, B.A., Wagenmakers, E.J., Berk, R., Bollen, K.A., Brembs, B., Brown, L., Camerer, C. and Cesarini, D., 2018. Redefine statistical significance. *Nature Human Behaviour, 2(1), p.6.*

- Collins, F.S. and Tabak, L.A., 2014. NIH plans to enhance reproducibility. Nature, *505(7485), p.612.*

- Depristo MA, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data, *Nat. Genet. , 2011, vol. 43 (pg. 491-498)*

- 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), pp.2628-2631.*

- MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., Junkins, H., McMahon, A., Milano, A., Morales, J. and Pendlington, Z.M. (2016) The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic acids research, 45(D1), pp.D896-D901.*

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

- Regier, A.A., Farjoun, Y., Larson, D., Krasheninina, O., Kang, H.M., Howrigan, D.P., Chen, B.J., Kher, M., Banks, E., Ames, D.C., English, A.C., Li, H., The NHLBI Trans-Omics for Precision Medicine (TOPMed) Program, Abecasis, G., 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. *bioRxiv, p.269316.*

- Stark, P. (2018). No reproducibility without preproducibility. *Nature, 557, p613.*

- Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M. and Liu, B., 2015. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine,* 12(3), p.e1001779.

- Visscher, P.M., Wray, N.R., Zhang, Q., Sklar, P., McCarthy, M.I., Brown, M.A. and Yang, J. (2017) 10 years of GWAS discovery: biology, function, and translation. *The American Journal of Human Genetics, 101(1), pp.5-22.*

- Wasserstein, R.L. and Lazar, N.A., 2016. The ASA's statement on p-values: context, process, and purpose. *The American Statistician, 70(2), pp.129-133.*