Statistical learning of large-scale genetic data: How to run a genome-wide association study of gene-expression data using the 1000 Genomes Project data

Anton Sugolov[#]

Department of Mathematics, Faculty of Arts and Sciences, University of Toronto;

Eric Emmenegger[#]

Department of Mechanical and Industrial Engineering,University of Toronto;

Andrew D. Paterson[*]

Program in Genetics & Genome Biology

The Hospital for Sick Children, Toronto, Ontario, Canada;

Dalla Lana School of Public Health, University of Toronto;

Lei Sun[*]

Department of Statistical Sciences, Faculty of Arts and Sciences;

Dalla Lana School of Public Health, University of Toronto

September 2, 2022

[#]These authors contributed equally to this work

[*]Corresponding authors: sun@utstat.toronto.edu; andrew.paterson@sickkids.ca

Abstract

Teaching statistics through engaging applications to contemporary large-scale datasets is essential to attracting students to the field. To this end, we developed a hands-on, week-long workshop for senior high-school or junior undergraduate students, without prior knowledge in statistical genetics but with some basic knowledge in data science, to conduct their own genome-wide association studies (GWAS). The GWAS was performed for open source gene expression data, using publicly-available human genetics data. Assisted by a detailed instruction manual, students were able to obtain $\sim$1.4 million p-values from a real scientific study, within several days. This early motivation kept students engaged in learning the theories that support their results, including regression, data visualization, results interpretation, and large-scale multiple hypothesis testing. To further their learning motivation by emphasizing the personal connection to this type of data analysis, students were encouraged to make short presentations about how GWAS has provided insights into the genetic basis of diseases that are present in their friends and/or families. The appended open source, step-by-step instruction manual includes descriptions of the datasets used, the software needed, and results from the workshop. Additionally, scripts used in the workshop are archived on Zenodo to further enhance reproducible research and training.

Keywords: 1000 Genomes Project, Data Visualization, Genome-wide Association Study, Gene Expression, Hands-on Experience, Large-scale Data Analysis, Multiple Hypothesis Testing, Open Resource, Reproducible Research.

1 Introduction

The overarching goal of this project is providing an example of engaging education in statistics to attract senior high-school or undergraduate students to the field, who will eventually grow and mature as competent data scientists. To achieve this goal, we designed a week-long workshop that provides students contextual, immersed, and hands-on learning experience in data science, using publicly-available, contemporary datasets.

We chose genetic data as the domain knowledge because they are complex, large-scale, high-dimensional, and practically important (Risch & Merikangas 1996). Although we do not expect nor want all students to continue their studies in statistical genetics, at the end of the workshop we expect students to (a) know about the variations in the human genome and the structure of the human population, (b) put into use their statistical knowledge by working with the 1000 Genomes Project (1KG) data (Auton et al. 2015), and (c) deepen their statistical understanding in areas including confounding (Cummiskey et al. 2020, Hu & Ziv 2008), heterogeneity (Higgins & Thompson 2002, Gordon et al. 2020), using principle component analysis to capture population structure (Price et al. 2006, Reich et al. 2008, Abdi & Williams 2010), multiple hypothesis testing (Shaffer 1995, Dudbridge & Gusnanto 2008), results interpretation and data visualization (Chanock et al. 2007, Hudiburgh & Garbinsky 2020), and reproducible research (Dragicevic et al. 2019, Ostblom & Timbers 2022).

In the last 15 years, genome-wide association studies (GWAS) have become a highly efficient way to identify genetic variants associated with traits and diseases (Manolio 2010, Visscher et al. 2017, Lappalainen & MacArthur 2021, Tan & Timpson 2022). The typical method involves testing millions of bi-allelic single nucleotide polymorphisms (SNPs), one-at-a-time for association with an outcome (e.g. the continuous blood pressure or the binary trait of high blood pressure) using either linear or logistic multivariate regression, and more recently generalized linear mixed-effect models (Cordell & Clayton 2005, Zhou & Stephens 2012). Although the commonly used statistical models are relatively simple for each SNP, the main challenge relates to the size of the human genome and the number of SNPs. For example, in imputed genetic data from the UK Biobank (Bycroft et al. 2018, Das et al. 2016), about 10 million SNPs are typically analyzed. Additionally, prior to association testing, several (domain-specific) quality control (QC) steps are

necessary to restrict the analysis to SNPs and individuals with high quality data (Marees et al. 2018).

Most individual-level genome-wide SNP data is not publicly available due to privacy (Lunshof et al. 2008). We chose to illustrate GWAS using publicly available trait and genetic data from the 1000 Genomes project, in which participants consented to their data being made freely available (Clarke et al. 2012). Due to the small sample size available (about 1000 in total with 88 Yoruban and 102 Utah individuals, small relative to 1,344,840, the number of SNPs analyzed), we chose a trait that is known to be strongly associated with some SNPs with large genetic effects. This way, there would be sufficient power to detect the association with the small sample size; the remaining SNPs serve as negative controls and demonstrate issues pertinent to large-scale multiple hypothesis testing.

There is a wide variation in the level of gene expression in a specific tissue or cell, and much of this variation is influenced by SNPs near to a specific gene, the so-called cis-eQTL (Võsa et al. 2021). We used an example from earlier literature to illustrate the identification of genetic variation associated with the level of expression of the gene named Endoplasmic Reticulum Aminopeptidase 2 (*ERAP2*) (Cheung et al. 2005). The ERAP2 gene expression levels were measured in peripheral blood B cell lines in Utah residents with European ancestry, and Yoruba people from Ibadan, Nigeria from the 1000 Genomes Project (Roslin et al. 2016). The project is a publicly available catalogue of individual-level human genetic variation[1], constructed by measuring genetic variation with an array of technologies in multiple populations around the world (Auton et al. 2015).

The workshop is designed to be executed with a 4-5 day period. The mornings can be used for the more traditional teaching modus operandi via (interactive) lectures, while the afternoons may be dedicated to the hands-on component with sufficient Teaching Assistant (TA) support. The student-TA ratio could range from 1-5 to 1-10, depending on the readiness of the student cohort. The last 2-3 hour of the workshop is recommended for a general discussion and obtaining feedback from the students, and ideally including short student presentations; see Section 2.5.

[1] `https://www.internationalgenome.org/about`

2 Methods

First and foremost, the workshop provides extensive hands-on experience in conducting, summarizing and interpreting a genome-wide association study to senior high-school students or junior undergraduate students with basic knowledge in data science. The hands-on experience includes using R (Maindonald 2008), running PLINK v1.9 (Purcell et al. n.d.) which is specific to the GWAS domain, and working with large-scale data. A detailed manual is attached as an Appendix, and the most updated version is openly accessible[2].

Additionally, the workshop has the more traditional teaching/learning component through (interactive) lectures, covering complementary topics in genetics and statistics. We have made the lecture notes openly accessible[3].

2.1 Datasets

In total, 190 individuals and 1,344,840 bi-allelic SNPs from the 1000 Genomes Project (Auton et al. 2015, Roslin et al. 2016) passing quality control from The Centre for Applied Genomic (TCAG)[4] were used for the genome-wide association study.

Quality control is a significant component of conducting a proper GWAS (Marees et al. 2018). However, in-depth QC is domain-specific and time-consuming, not suitable for the purpose of this workshop. We thus provides a set of good quality data while emphasizing the importance of QC, so that the participating students could successfully carry out a preliminary GWAS within the first two days of the workshop and obtain $\sim$1.4 million p-values from a real scientific study. We note that this early success is critical to keeping the students engaged and motivated to learn the theories that support their empirical results.

Cheung et al. (2005) identified that the expression of the gene *ERAP2* had strong genetic association in HapMap 3 individuals (International HapMap Consortium 2007), many of which overlapped with the 1KG individuals. Gene expressions of *ERAP2* measured in peripheral blood B cell lines were first extracted from Array Express (Montgomery et al. 2010, Stranger et al. 2012), then matched to the IDs of 1KG individuals, and finally formatted for PLINK; see Appendix A.

[2] https://github.com/sugolov/GWAS-Workshop

[3] https://github.com/LeiSunUofT/How-to-Run-a-GWAS

[4] https://tcag.ca/tools/1000genomes.html;https://www.internationalgenome.org/

The two largest 1KG sub-populations are Yoruban individuals in Ibadan, Nigeria (YRI), and Utah residents (CEPH, Centre d'Etude du Polymorphisme Humain) with Northern and Western European ancestry (CEU). In total, 91 YRI individuals and 104 CEU individuals matched between the 1KG and HapMap 3 datasets, and they were used for the workshop purpose.

Using principal component analysis (PCA) of PLINK v1.9 (Purcell et al. n.d.), three and two outliers were removed, respectively from the YRI and CEU samples. Thus, the final GWAS analysis was restricted to 88 YRI individuals and 102 CEU individuals, and their genetic data of 1,344,840, bi-allelic SNPs. The basic PCA analysis pipeline is provided in the appended manual and could be part of the workshop if time permits.

2.2 Software

An introduction to PLINK (v1.90 beta 6.24) (Chang et al. 2015, Purcell & Chang 2021) is necessary for the purpose of this GWAS workshop. Depending on the readiness of the student cohort (and length of the workshop), a brief introduction to using R (v4.1.0) (R Core Team 2021) could be also part of the workshop; open-resource R introduction materials abound[5].

PLINK is a command line toolkit for performing the GWAS computation efficiently, giving students hands-on experience with the most popular software used in the ongoing GWAS research. Additionally, the installation and use of R packages such as "qqman" (Turner 2018), "ggplot2" (Wickham 2016) and "hexbin" (Carr et al. 2021) introduce students to effective data visualization, a core component of interpreting GWAS results. Included in the open source manual is also a brief introduction to an (optional) use of the UNIX environment.

2.3 Overview of the Workshop Content

We summarize the main steps of running a GWAS of the gene expression data of *ERAP2*, using the 88 YRI individuals and their 1,344,840 SNP data of the 1000 Genomes Project (i.e. the YRI GWAS). We refer the readers to the open source manual and scripts, which include further analyses (i.e. the CEU GWAS) that could be reproduced using the step-by-step instructions.

Since trait distribution and SNP frequency may differ between populations, GWAS is often performed separately for each population (Price et al. 2006). In the analyzed sample, additional

[5] `https://cran.r-project.org/doc/manuals/r-release/R-intro.pdf`

PCA may be conducted to capture fine-scale population structure (Reich et al. 2008); see Section 3 of the appended manual on population stratification.

1. **Prepare the datasets.** Extract the cleaned 1KG SNP data into a separate analysis-specific directory.

   First, students should specify the phenotype of interest and remove individuals who are not needed for the YRI GWAS. Students achieve these with the `--pheno` and `--prune` PLINK commands respectively; for additional details see the section named 'Standard data input' of the PLINK documentation[6].

   Second, students remove rare SNPs (e.g. with a minor allele frequency (MAF) less than 5%) and the sex chromosomes from the analysis using the `--maf 0.05` and `--chr 1-22` flags, respectively [7]. (The 1000 Genomes data quality control performed by Roslin et al. (2016) does not include a MAF-based QC step.) Students should only analyze the autosomal common SNPs, as identifying associations on the sex chromosomes (Chen et al. 2021, Wang et al. 2022) and analyzing rare SNPs (Derkach et al. 2014) requires more intricate methods beyond the scope of the workshop.

   Lastly, for computational reasons, students create binary files from this dataset with `--make-bed`. The `.bim`, `.bed`, `.fam` file types should be generated and named after `ERAP2_YRI`. Students should verify that the parameters they have entered are correct by viewing the `.log` file.

2. **Run the association analysis.** Since gene expression data is continuous, students should specify a linear regression, with PLINK command `--linear`. This evaluates the association between the gene expression and each SNP, also known as the expression quantitative trait loci (eQTL) analysis.

   Association analysis often includes covariates to avoid spurious associations from confounding. The sexes of the individuals are included in the dataset, so students may include this covariate in the eQTL GWAS analysis by using `--linear sex`.

3. **Post-association analysis and results interpretation.** The association results can be sorted with `sort.R`, which also generates a file with the top 50 most significant SNPs.

[6] `https://www.cog-genomics.org/plink/`
[7] `https://www.cog-genomics.org/plink/1.9/filter`

The genome-wide results may be plotted and interpreted, which we explain with examples in the next section; also see the appended manual for additional details. Using appropriate QC steps, including the MAF filtering, prevents `NA` results in the output in principle. However, to be cautious the `NA_removal.R` script may be used to identify and remove `NA` results from the follow-up data visualization analyses.

2.4 A Highlight: Multiple Hypothesis Testing and Data Visualization

During the workshop, students are introduced to the multiple testing problem in GWAS through the morning lectures. Although the concept of multiple hypothesis testing, and its (theoretical) connection with 'p-values being Unif(0,1) distributed under the null', is covered in most introduction courses to statistics, student's understanding and appreciation of this concept is often lacking, in part due to the traditional emphasis on identifying variables with p-values meeting some significance criterion, as opposed to exploring the whole distribution. This, in part, is a result from a lack of hands-on experience with large-scale real data analysis,

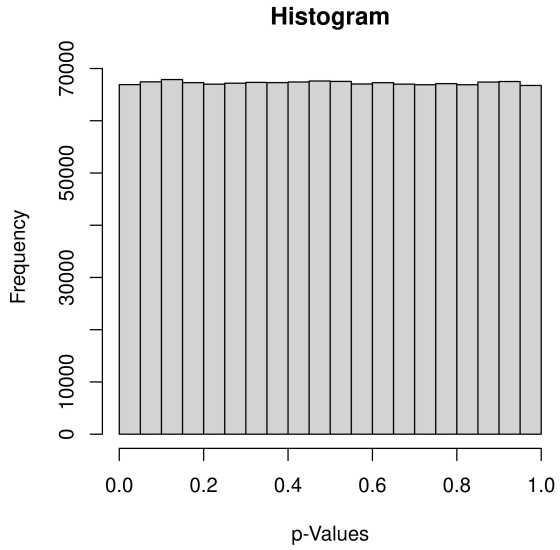

Figure 1: Histogram of the 1,344,840 p-values from the YRI GWAS of the gene expression of *ERAP2*, obtained using the workshop materials. The histogram is close to Unif(0, 1), the expected distribution of p-values under the null of no association.

With close to 1.4 million p-values obtained from a real GWAS, students realize that many

SNPs (close to 70,000 in fact) are 'significant' if the traditional $\alpha = 0.05$ type I error threshold were used. However, the histogram of p-values in Figure 1 shows an empirical distribution close to Unif(0,1), the distribution expected under the null hypothesis of no association. This is expected for a typical GWAS, as unless the trait is polygenic (i.e. with a large number of contributing SNPs) *and* the sample size is very large, most of the SNPs are not expected to be associated with the trait or their associations are not detectable (Devlin et al. 2001, Yang et al. 2011).

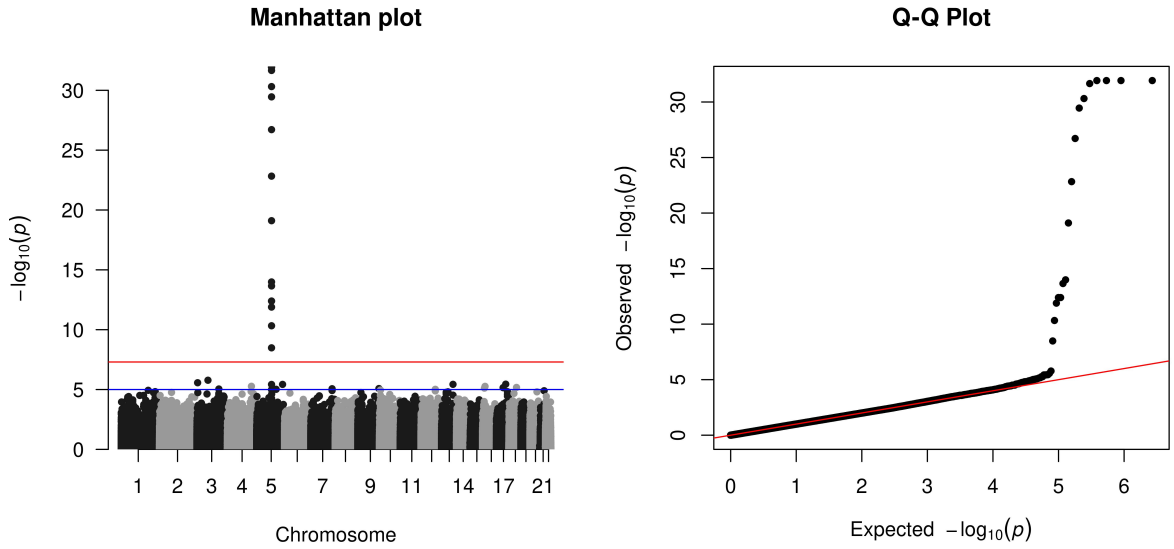

Figure 2: The Manhattan plot and Q-Q plot of the 1,344,840 p-values from the YRI GWAS of the gene expression of *ERAP2*, obtained using the workshop materials.

Without going into the technical details, students are then introduced to the $\alpha = 5.0 \times 10^{-8}$ genome-wide significance threshold used in GWAS to control the family-wise error rate at 0.05 (Dudbridge & Gusnanto 2008). Further, two most commonly used data visualization plots in GWAS are introduced: the Manhattan plot and the Q-Q plot as shown in Figure 2. These two plots complement the histogram which lumps all small p-values in one bin, thus masking the individual significant results.

The Q-Q plot in Figure 2 is a standard statistical plot, showing the quantiles of the observed p-values against those of Unif(0,1), at the $-\log_{10}$ scale. In GWAS, the Q-Q plot serves two purposes. First, it highlights the significant results if there are any at the tail of the distribution.

Second, it also shows the overall distribution of the GWAS p-values (though on the $-\log_{10}$ scale), which is typically expected to follow the main diagonal line.

Based on the Q-Q plot in Figure 2, it is clear that several SNPs are significantly associated with the gene expression of *ERAP2* in the YRI GWAS. However, their genomic locations (e.g. from which chromosome) are unclear. Thus comes the Manhattan plot which contrasts the $-\log_{10}$ p-value of each SNP against its genomic location, with the $\alpha = 5.0 \times 10^{-8}$ genome-wide significance line (7.3 on $-\log_{10}$ scale) marked in red. Other significance thresholds for 'suggestive' association may also be shown, such as the $-\log_{10}(10^{-5})$ blue horizontal line included in Figure 2.

In total, there are 17 genome-wide significant SNPs with p-values less than 5.0×10^{-8}, all from the locus on chromosome 5 (at 96.2 – 96.3 Mb) that is close to the *ERAP2* gene. These are called cis-eQTL SNPs, i.e. SNPs near the gene and whose genotypes associated with differences in the gene expression level.

Another noticeable feature in a typical Manhattan plot is the 'clustering' of significantly associated SNPs. This is due to the phenomenon called linkage disequilibrium (LD) between nearby SNPs (Slatkin 2008). The location, p-values, and the LD between SNPs of a significant locus may be visualized in a Manhattan-like plot using the LocusZoom service (Boughton et al. 2021), with steps and the *ERAP2* example included in the appended workshop manual. Although LD is akin to the basic statistical concept of correlation, it is a more advanced concept in statistical genetics involving population genetics, thus not discussed further in this workshop.

2.5 Summary of the GWAS Workshop Conducted

In the summer of 2021, our team offered this workshop to 15 senior high school students from the University of Toronto Schools (UTS) in Toronto, Ontario, Canada. Due to the pandemic and limited number of TAs available, it was offered online and restricted to 15 participants, which were selected based on their interests and readiness in statistics, genetics and computing; see Appendix C for the application form. Post-workshop, a survey was conducted to collect participant feedback; see Appendix D for the survey questions.

Prior to the workshop, in addition to the survey, an earlier version of the appended manual was distributed to the participating students. Additionally, given the relatively low readiness

of the participating students, the two lead TAs (AS and EE) provided detailed instructions for software installation and configuration, with a troubleshooting guide. Students followed this manual to work in groups, with clarification from the TAs via an online tutorial session as well as Discord discussions; Discord was the preferred social media of this group of students. At the time of the workshop, AS and EE were first year undergraduate students majored, respectively, in mathematics and life sciences, at the University of Toronto; AS and EE were mentored by ADP and LS during the summer of 2020.

Throughout the 4.5-day workshop, the morning lectures providing the necessary background in genetics and statistics were given, respectively, by ADP and LS. The afternoon sessions were guided tutorials, lead by AS and EE with participation of ADP and LS. Notably, on the last morning, students were encouraged to select a trait from the GWAS catalog[8] (Buniello et al. 2019) and to present a 3-5 minute summary of a paper that performed a GWAS for that trait. In addition, students were encouraged to describe their motivation for selecting each particular trait, which provided an emotional connection to the science through personal stories, typically related to family history of diseases. The presented traits ranged from gout, breast cancer, to multiple sclerosis. Finally, to keep the students engaged, music were curated in advance and played during the (frequent) breaks, and the song “Another Brick in the Wall", by Pink Floyd, was much appreciated by the students based on their feedback.

3 Student Feedback

After the workshop, a feedback survey (Appendix D) was distributed and eight responses were collected. Students found the workshop overall interesting, especially working with and interpreting the genetic component of the workshop. The students particularly enjoyed the SNP finding activity, and found the guided afternoon sessions helpful to their understanding.

Due to the high school background of the students, and the workshop’s limited time frame, some found the pace of the lectures to be overwhelming, particularly the statistical section of the lectures. Students unaccustomed to programming found using the terminal-based PLINK to be confusing, and recommended adding a terminal tutorial to the workshop manual, which was later included.

[8] `https://www.ebi.ac.uk/gwas/`

4 Discussion

Depending on the experience of participants, the scope of the workshop may be extended, including covering more advanced lectures, analyses and plots, as well as analyzing additional datasets. Discussions around the cleaning of 1000 Genomes data could be included in the morning lecture sessions, and cleaning steps for 1000 Genomes individuals (Roslin et al. 2016) may be replicated in the afternoons. More thorough descriptions of large-scale multiple testing and fundamentals of regression in the GWAS context may be included. An analysis using individuals with different populations, with PCA adjustment, may be given in the practical hands-on sessions. After conducting a sample GWAS in one population (e.g. the YRI GWAS), gene expressions with various significance (Cheung et al. 2005) matched with other 1KG populations may be provided for students to replicate. Included UNIX commands may be used as an introduction to conducting a remote GWAS on a cloud-based system, which typically are typically UNIX-based.

To adhere to the current standard of reproducible research (Peng 2011), initial GWAS were conducted and documented independently by AS and EE. The two sets of results were then compared with each other, and the analyses and results were successfully reproduced, independently, by the workshop participants. Additionally, the observed *ERAP2* significance replicates the earlier work by Cheung et al. (2005). R, PLINK, and dataset versions were synchronized, and all scripts were version-controlled and hosted on GitHub[9]. The exact analytical steps were recorded in a GWAS documentation, which would later become the appended, open source manual that allows users to reproduce the workshop GWAS materials. Finally, the tested workshop datasets and other materials were made publicly available on Zenodo [10].

5 Acknowledgement

This research is funded by the Natural Sciences and Engineering Research Council of Canada (NSERC, RGPIN-04934), the Canadian Institutes of Health Research (CIHR, PJT-180460), and the University of Toronto Data Sciences Institute (DSI) Catalyst Grant.

[9] `https://github.com/sugolov/GWAS-Workshop/`
[10] `https://zenodo.org/record/6981694`

6 Conflict of Interest Statement

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.

References

Abdi, H. & Williams, L. J. (2010), 'Principal component analysis', *Wiley interdisciplinary reviews: computational statistics* **2**(4), 433–459.

Auton, A., Abecasis, G. R., Altshuler, D. M., Durbin, R. M., Bentley, D. R., Chakravarti, A., Clark, A. G., Donnelly, P., Eichler, E. E., Flicek, P., Gabriel, S. B., Gibbs, R. A., Green, E. D., Hurles, M. E., Knoppers, B. M., Korbel, J. O., Lander, E. S., Lee, C., Lehrach, H., Mardis, E. R., Marth, G. T., McVean, G. A., Nickerson, D. A., Schmidt, J. P., Sherry, S. T., Wang, J., Wilson, R. K., Boerwinkle, E., Doddapaneni, H., Han, Y., Korchina, V., Kovar, C., Lee, S., Muzny, D., Reid, J. G., Zhu, Y., Chang, Y., Feng, Q., Fang, X., Guo, X., Jian, M., Jiang, H., Jin, X., Lan, T., Li, G., Li, J., Li, Y., Liu, S., Liu, X., Lu, Y., Ma, X., Tang, M., Wang, B., Wang, G., Wu, H., Wu, R., Xu, X., Yin, Y., Zhang, D., Zhang, W., Zhao, J., Zhao, M., Zheng, X., Gupta, N., Gharani, N., Toji, L. H., Gerry, N. P., Resch, A. M., Barker, J., Clarke, L., Gil, L., Hunt, S. E., Kelman, G., Kulesha, E., Leinonen, R., McLaren, W. M., Radhakrishnan, R., Roa, A., Smirnov, D., Smith, R. E., Streeter, I., Thormann, A., Toneva, I., Vaughan, B., Zheng-Bradley, X., Grocock, R., Humphray, S., James, T., Kingsbury, Z., Sudbrak, R., Albrecht, M. W., Amstislavskiy, V. S., Borodina, T. A., Lienhard, M., Mertes, F., Sultan, M., Timmermann, B., Yaspo, M.-L., Fulton, L., Fulton, R., Ananiev, V., Belaia, Z., Beloslyudtsev, D., Bouk, N., Chen, C., Church, D., Cohen, R., Cook, C., Garner, J., Hefferon, T., Kimelman, M., Liu, C., Lopez, J., Meric, P., O'Sullivan, C., Ostapchuk, Y., Phan, L., Ponomarov, S., Schneider, V., Shekhtman, E., Sirotkin, K., Slotta, D., Zhang, H., Balasubramaniam, S., Burton, J., Danecek, P., Keane, T. M., Kolb-Kokocinski, A., McCarthy, S., Stalker, J., Quail, M., Davies, C. J., Gollub, J., Webster, T., Wong, B., Zhan, Y., Campbell, C. L., Kong, Y., Marcketta, A., Yu, F., Antunes, L., Bainbridge, M., Sabo, A., Huang, Z., Coin, L. J. M., Fang, L., Li, Q., Li, Z., Lin, H., Liu, B., Luo, R., Shao, H., Xie, Y., Ye, C., Yu, C., Zhang, F., Zheng, H.,

Zhu, H., Alkan, C., Dal, E., Kahveci, F., Garrison, E. P., Kural, D., Lee, W.-P., Fung Leong, W., Stromberg, M., Ward, A. N., Wu, J., Zhang, M., Daly, M. J., DePristo, M. A., Handsaker, R. E., Banks, E., Bhatia, G., del Angel, G., Genovese, G., Li, H., Kashin, S., McCarroll, S. A., Nemesh, J. C., Poplin, R. E., Yoon, S. C., Lihm, J., Makarov, V., Gottipati, S., Keinan, A., Rodriguez-Flores, J. L., Rausch, T., Fritz, M. H., Stütz, A. M., Beal, K., Datta, A., Herrero, J., Ritchie, G. R. S., Zerbino, D., Sabeti, P. C., Shlyakhter, I., Schaffner, S. F., Vitti, J., Cooper, D. N., Ball, E. V., Stenson, P. D., Barnes, B., Bauer, M., Keira Cheetham, R., Cox, A., Eberle, M., Kahn, S., Murray, L., Peden, J., Shaw, R., Kenny, E. E., Batzer, M. A., Konkel, M. K., Walker, J. A., MacArthur, D. G., Lek, M., Herwig, R., Ding, L., Koboldt, D. C., Larson, D., Ye, K., Gravel, S., 1000 Genomes Project Consortium, T., authors, C., committee, S., group, P., of Medicine, B. C., BGI-Shenzhen, of MIT, B. I., Harvard, for Medical Research, C. I., European Molecular Biology Laboratory, E. B. I., Illumina, for Molecular Genetics, M. P. I., at Washington University, M. G. I., of Health, U. N. I., of Oxford, U., Institute, W. T. S., group, A., Affymetrix, of Medicine, A. E. C., University, B., College, B., Laboratory, C. S. H., University, C., Laboratory, E. M. B., University, H., Database, H. G. M., at Mount Sinai, I. S. o. M., University, L. S., Hospital, M. G., University, M. & National Eye Institute, N. I. H. (2015), 'A global reference for human genetic variation', *Nature* **526**(7571), 68–74.
URL: *https://doi.org/10.1038/nature15393*

Boughton, A. P., Welch, R. P., Flickinger, M., VandeHaar, P., Taliun, D., Abecasis, G. R. & Boehnke, M. (2021), 'LocusZoom.js: interactive and embeddable visualization of genetic association study results', *Bioinformatics* .
URL: *https://doi.org/10.1093/bioinformatics/btab186*

Buniello, A., MacArthur, J. A. L., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E. et al. (2019), 'The nhgri-ebi gwas catalog of published genome-wide association studies, targeted arrays and summary statistics 2019', *Nucleic acids research* **47**(D1), D1005–D1012.

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J. et al. (2018), 'The uk biobank resource with deep phenotyping and genomic data', *Nature* **562**(7726), 203–209.

Byrska-Bishop, M., Evani, U. S., Zhao, X., Basile, A. O., Abel, H. J., Regier, A. A., Corvelo, A., Clarke, W. E., Musunuri, R., Nagulapalli, K., Fairley, S., Runnels, A., Winterkorn, L., Lowy-Gallego, E., Consortium, T. H. G. S. V., Flicek, P., Germer, S., Brand, H., Hall, I. M., Talkowski, M. E., Narzisi, G. & Zody, M. C. (2021), 'High coverage whole genome sequencing of the expanded 1000 genomes project cohort including 602 trios', *bioRxiv* .
URL: *https://www.biorxiv.org/content/early/2021/02/07/2021.02.06.430068*

Carr, D., Lewin-Koh, N., Maechler, M. & Sarkar, D. (2021), *hexbin: Hexagonal Binning Routines*. R package version 1.28.2.
URL: *https://CRAN.R-project.org/package=hexbin*

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M. & Lee, J. J. (2015), 'Second-generation PLINK: rising to the challenge of larger and richer datasets', *GigaScience* **4**(1). s13742-015-0047-8.
URL: *https://doi.org/10.1186/s13742-015-0047-8*

Chanock, S. J., Manolio, T., Boehnke, M., Boerwinkle, E., Hunter, D. J., Thomas, G., Hirschhorn, J. N., Abecasis, G. R., Altshuler, D., Bailey-Wilson, J. E. et al. (2007), 'Replicating genotype-phenotype associations'.

Chen, B., Craiu, R. V., Strug, L. J. & Sun, L. (2021), 'The x factor: A robust and powerful approach to x-chromosome-inclusive whole-genome association studies', *Genetic epidemiology* **45**(7), 694–709.

Cheung, V. G., Spielman, R. S., Ewens, K. G., Weber, T. M., Morley, M. & Burdick, J. T. (2005), 'Mapping determinants of human gene expression by regional and genome-wide association', *Nature* **437**(7063), 1365–1369.

Clarke, L., Zheng-Bradley, X., Smith, R., Kulesha, E., Xiao, C., Toneva, I., Vaughan, B., Preuss, D., Leinonen, R., Shumway, M. et al. (2012), 'The 1000 genomes project: data management and community access', *Nature methods* **9**(5), 459–462.

Cordell, H. J. & Clayton, D. G. (2005), 'Genetic association studies', *The Lancet* **366**(9491), 1121–1131.

Cummiskey, K., Adams, B., Pleuss, J., Turner, D., Clark, N. & Watts, K. (2020), 'Causal inference in introductory statistics courses', *Journal of Statistics Education* **28**(1), 2–8.

Das, S., Forer, L., Schönherr, S., Sidore, C., Locke, A. E., Kwong, A., Vrieze, S. I., Chew, E. Y., Levy, S., McGue, M. et al. (2016), 'Next-generation genotype imputation service and methods', *Nature genetics* **48**(10), 1284–1287.

Derkach, A., Lawless, J. F. & Sun, L. (2014), 'Pooled association tests for rare genetic variants: a review and some new results', *Statistical Science* **29**(2), 302–321.

Devlin, B., Roeder, K. & Wasserman, L. (2001), 'Genomic control, a new approach to genetic-based association studies', *Theoretical population biology* **60**(3), 155–166.

Dragicevic, P., Jansen, Y., Sarma, A., Kay, M. & Chevalier, F. (2019), Increasing the transparency of research papers with explorable multiverse analyses, *in* 'Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems', pp. 1–15.

Dudbridge, F. & Gusnanto, A. (2008), 'Estimation of significance thresholds for genomewide association scans', *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society* **32**(3), 227–234.

Gordon, D., Finch, S. J. & Kim, W. (2020), 'Heterogeneity in statistical genetics'.

Higgins, J. P. & Thompson, S. G. (2002), 'Quantifying heterogeneity in a meta-analysis', *Statistics in medicine* **21**(11), 1539–1558.

Hu, D. & Ziv, E. (2008), 'Confounding in genetic association studies and its solutions', *Pharmacogenomics in Drug Discovery and Development* pp. 31–39.

Hudiburgh, L. M. & Garbinsky, D. (2020), 'Data visualization: Bringing data to life in an introductory statistics course', *Journal of Statistics Education* **28**(3), 262–279.

International HapMap Consortium (2007), 'A second generation human haplotype map of over 3.1 million snps', *Nature* **449**(7164), 851–861.
URL: *https://doi.org/10.1038/nature06258*

Lappalainen, T. & MacArthur, D. G. (2021), 'From variant to function in human disease genetics', *Science* **373**(6562), 1464–1468.

Lunshof, J. E., Chadwick, R., Vorhaus, D. B. & Church, G. M. (2008), 'From genetic privacy to open consent', *Nature Reviews Genetics* **9**(5), 406–411.

Maindonald, J. H. (2008), *Using R for Data Analysis and Graphics: Introduction, Code and Commentary*, Maindonald, J H.

Manolio, T. A. (2010), 'Genomewide association studies and assessment of the risk of disease', *New England journal of medicine* **363**(2), 166–176.

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)**(e1608).

Montgomery, S. B., Sammeth, M., Gutierrez-Arcelus, M., Lach, R. P., Ingle, C., Nisbett, J., Guigo, R. & Dermitzakis, E. T. (2010), 'Transcriptome genetics using second generation sequencing in a caucasian population', *Nature* **464**(7289), 773—777.
URL: *https://europepmc.org/articles/PMC3836232*

Ostblom, J. & Timbers, T. (2022), 'Opinionated practices for teaching reproducibility: motivation, guided instruction and practice', *Journal of Statistics and Data Science Education* (just-accepted), 1–22.

Peng, R. D. (2011), 'Reproducible research in computational science', *Science* **334**(6060), 1226–1227.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A. & Reich, D. (2006), 'Principal components analysis corrects for stratification in genome-wide association studies', *Nature genetics* **38**(8), 904–909.

Purcell, S. & Chang, C. (2021), 'Plink 1.90b6.24'.
URL: *www.cog-genomics.org/plink/1.9/*

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., de Bakker, P. I. W., Daly, M. J. & Sham, P. C. (n.d.), 'Plink: a tool set for whole-genome association and population-based linkage analyses.', *American journal of human genetics* **81**(3), 559–575.
URL: *https://doi.org/10.1086/519795*

R Core Team (2021), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
URL: *https://www.R-project.org/*

Reich, D., Price, A. L. & Patterson, N. (2008), 'Principal component analysis of genetic data', *Nature genetics* **40**(5), 491–492.

Risch, N. & Merikangas, K. (1996), 'The future of genetic studies of complex human diseases', *Science* **273**(5281), 1516–1517.

Roslin, N. M., Weili, L., Paterson, A. D. & Strug, L. J. (2016), 'Quality control analysis of the 1000 genomes project omni2.5 genotypes', *bioRxiv* .
URL: *https://www.biorxiv.org/content/early/2016/09/30/078600*

Shaffer, J. P. (1995), 'Multiple hypothesis testing', *Annual review of psychology* **46**(1), 561–584.

Slatkin, M. (2008), 'Linkage disequilibrium—understanding the evolutionary past and mapping the medical future', *Nature Reviews Genetics* **9**(6), 477–485.

Stranger, B. E., Montgomery, S. B., Dimas, A. S., Parts, L., Stegle, O., Ingle, C. E., Sekowska, M., Smith, G. D., Evans, D., Gutierrez-Arcelus, M., Price, A., Raj, T., Nisbett, J., Nica, A. C., Beazley, C., Durbin, R., Deloukas, P. & Dermitzakis, E. T. (2012), 'Patterns of cis regulatory variation in diverse human populations', *PLOS Genetics* **8**(4), 1–13.
URL: *https://doi.org/10.1371/journal.pgen.1002639*

Tan, V. Y. & Timpson, N. J. (2022), 'The uk biobank: A shining example of genome-wide association study science with the power to detect the murky complications of real-world epidemiology', *Annual Review of Genomics and Human Genetics* **23**.

Turner, S. (2018), 'qqman: an r package for visualizing gwas results using q-q and manhattan plots', *The Journal of Open Source Software* .

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.

Võsa, U., Claringbould, A., Westra, H.-J., Bonder, M. J., Deelen, P., Zeng, B., Kirsten, H., Saha, A., Kreuzhuber, R., Brugge, H. et al. (2021), 'Large-scale cis-and trans-eqtl analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression', *Nature genetics* **53**(9), 1300–1310.

Wang, Z., Sun, L. & Paterson, A. D. (2022), 'Major sex differences in allele frequencies for x chromosomal variants in both the 1000 genomes project and gnomad', *PLoS genetics* **18**(5), e1010231.

Wickham, H. (2016), *ggplot2: Elegant Graphics for Data Analysis*, Springer-Verlag New York. **URL:** *https://ggplot2.tidyverse.org*

Yang, J., Weedon, M. N., Purcell, S., Lettre, G., Estrada, K., Willer, C. J., Smith, A. V., Ingelsson, E., O'connell, J. R., Mangino, M. et al. (2011), 'Genomic inflation factors under polygenic inheritance', *European Journal of Human Genetics* **19**(7), 807–812.

Zhou, X. & Stephens, M. (2012), 'Genome-wide efficient mixed-model analysis for association studies', *Nature genetics* **44**(7), 821–824.

APPENDICES

A Phenotype Extraction and Dataset Generation

Please refer to Github.com/sugolov/GWAS-Workshop/Notebooks/DatasetPreparation.Rmd to create a phenotype file using the expression data from the University of Geneva Medical School Montgomery et al. (2010), Stranger et al. (2012). Example phenotype files for ERAP2 are provided for the CEU and YRI populations both separately and together at Github.com/sugolov/GWAS-Workshop/Datasets. Please refer to Github.com/sugolov/GWAS-Workshop/Notebooks/YRI_Analysis.Rmd to combine the phenotype files with the 1KG dataset from The Centre for Applied Genomics Auton et al. (2015), Roslin et al. (2016)

B High Coverage Dataset

The High Coverage dataset was generated using the 30x High Coverage samples from the New York Genome Center (NYGC) Byrska-Bishop et al. (2021). Please refer to Github.com/sugolov/GWAS-Workshop/Notebooks/
High_Coverage.Rmd to generate a set of High Coverage data.

C Application Form

The application form for students consisted of the following questions sent out as a Google Form.

1. Your name (First, Last)
2. Your email address
3. Please list relevant courses (UTS course codes and names) taken in statistics/data science, computer science and biology. (This is to help the workshop organizers to team up participants with complementing skills, if needed depending on the number of applicants.)
4. Check 1-2 boxes that reflect your strengths
    - Statistics/Data Science

- Computing
- Biology

5. Explain why are you particularly interested in this workshop? (200 words)
6. Any preference or suggestion on the platform(s) to be used for the virtual workshop, and for the on-line discussion board?
7. Any other comments?

D End of Workshop Survey

The following questions were sent to the students as a Google Form after the end of the workshop. 8 students out of 17 responded.

1. On a scale of 1 to 10, how difficult did you find the genetic component?
2. If you answered greater than 7 to the question above please specify what you found too difficult. If you answered below 5 to the question please specify what you found too easy. If <=5 your answer <=7, still say something:-)
3. Did you find the pace of the genetic component too quick or too slow? Please specify.
4. What would you have liked to seen more of?
5. On a scale of 1 to 10, how difficult did you find the statistics component?
6. If you answered greater than 7 to the question above please specify what you found too difficult. If you answered below 5 to the question please specify what you found too easy. If <=5 your answer <=7, still say something:-)
7. Did you find the pace of the statistic component too quick or too slow? Please specify.
8. What would you have liked to seen more of?
9. On a scale of 1 to 10, how difficult did you find the computing/hands on component?

10. If you answered greater than 7 to the question above please specify what you found too difficult. If you answered below 5 to the question please specify what you found too easy. If <=5 your answer <=7, still say something:-)

11. Did you find the pace of the computing component too quick or too slow? Please specify.

12. What would you have liked to seen more of?

13. If we were to do this workshop again what would you have liked to see more of? Select all that apply.

    - Statistics
    - Genetics
    - Computing
    - Nothing. The balance was perfect.
    - Other:

14. What was your favourite aspect of the workshop? Select all that apply.

    - Statistics
    - Genetics
    - Computing
    - None. I did not enjoy anything
    - All. I loved everything
    - Other:

15. Would you like to have been presented with more references and resources before the workshop (ie. terminal commands, file directory structure, etc)? If this is the case please specify.

16. From a scale of 1 - 10 how much did you enjoy the music during breaks?

17. Any other final remarks?