



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 Sugolov¹ · Eric Emmenegger² · Andrew D. Paterson^{3,4} · Lei Sun^{4,5}

Received: 7 September 2022 / Revised: 7 April 2023 / Accepted: 22 May 2023 /
Published online: 1 July 2023
© The Author(s) 2023

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 study (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 ~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 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 Github and 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

Anton Sugolov and Eric Emmenegger have contributed equally to this work.

Extended author information available on the last page of the article

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 [39]. 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 (1 KG) data [2], and (c) deepen their statistical understanding in areas including confounding [15, 23], heterogeneity [21, 22], using principle component analysis to capture population structure [1, 34, 38], multiple hypothesis testing [20, 41], results interpretation and data visualization [9, 24], and reproducible research [19, 32].

Although in this application we focused on genetic data, computational software is important in many areas of large-scale data science, including for example astrostatistics, engineering and manufacturing data management, health data analytics, quantitative finance, and social network modeling and analysis. We highlight that, while application-specific considerations (e.g., domain-specific data quality control procedures) are important, the key statistical concepts introduced in this workshop are useful to analyze data from many domains other than genetic data. For example, multivariate linear regression is the building block for many applications that involve model fitting and statistical inference. Multiple hypothesis testing adjustment is necessary for any large-scale data analysis to prevent overfitting and p-hacking. Principal component analysis is a dimension-reduction technique popular in many data science fields. Finally, data visualization is increasingly recognized as an integral part of good data science practice.

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 [26, 29, 44, 46]. 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 [13, 51]. 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 [5, 16], 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 [30].

Most individual-level genome-wide SNP data is not publicly available due to privacy [27]. 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 [12]. 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. 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*) [11]. 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 [40]. 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 [2].

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 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 h of the workshop is recommended for a general discussion and obtaining feedback from the students, and ideally including short student presentations; see Sect. 2.5.

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 [28], running PLINK v1.9 [36] which is specific to the GWAS domain, and working with large-scale data. A detailed manual is attached as an Appendix. The most updated version, including an analogue of commands for PLINK v2 [8], is openly accessible.²

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

¹ <https://www.internationalgenome.org/about>

² <https://github.com/sugolov/GWAS-Workshop>

³ <https://github.com/LeiSunUofT/How-to-Run-a-GWAS>.

2.1 Datasets

In total, 190 individuals and 1,344,840 bi-allelic SNPs from the 1000 Genomes Project [2, 40] passing quality control from The Centre for Applied Genomic (TCAG)⁴

Quality control is a significant component of conducting a proper GWAS [30]. 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 ~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.

[11] identified that the expression of the gene *ERAP2* had strong genetic association in HapMap 3 individuals [25], many of which overlapped with the 1 KG individuals. Gene expressions of *ERAP2* measured in peripheral blood B cell lines were first extracted from Array Express [31, 43], then matched to the IDs of 1 KG individuals, and finally formatted for PLINK; see Appendix 1. The two largest 1 KG 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 an independent subset of the 1 KG with no family relations and HapMap 3 datasets, and these genetic unrelated individuals were used for the workshop purpose.

Using principal component analysis (PCA) of PLINK v1.9 [36], 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) [35] is necessary for the purpose of this GWAS workshop. 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. The analysis pipeline was originally implemented with PLINK v1.9 [36] but equivalent commands for PLINK v2 [8] are also provided in a separate manual.

⁴ <https://tcag.ca/tools/1000genomes.html>; <https://www.internationalgenome.org/> were used for the genome-wide association study. We note that the 1000 Genomes Project data contains some family data, but the individuals used for this workshop are genetically unrelated to each other. This is to facilitate the application of a simple multivariate linear regression, as accounting for related individuals requires more advanced statistical knowledge; directly applying the simple regression to a dependent sample leads to increased false positive rates.

Depending on the readiness of the student cohort (and length of the workshop), a brief introduction to using R (v4.1.0) [37] could be also part of the workshop; open-resource R introduction materials abound.⁵ The installation and use of R packages such as "qqman" [45], "ggplot2" [48] and "hexbin" [7] 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); GWAS is often performed separately for each population [34], as trait distribution and SNP frequency may differ between populations.

We refer the readers to the open source manual and scripts for additional details, which include further analyses (i.e., the CEU GWAS of the 102 CEU individuals and their SNP data) that could be reproduced using the step-by-step instructions. In the analyzed sample, additional PCA may be conducted to capture fine-scale population structure [38]; see Sect. 3 of the appended manual on population stratification.

1. **Prepare the datasets.** Extract the cleaned 1 KG 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 v1.9⁶ or PLINK v2.0⁷ documentation.

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.⁸ (The 1000 Genomes data quality control performed by [40] does not include a MAF-based QC step.)

Third, we note that Hardy-Weinberg equilibrium is typically part of the QC procedure (using `-hardy`), as severe departure from HWE is usually an indication of genotyping error [14, 30]. However, HWE is a complex phenomenon and HWE QC criterion is unclear [50]. Thus, the workshop analysis did not include a test of HWE, but we note that HWE should be evaluated for any significant SNPs. Additionally, students should only analyze the autosomal common SNPs, as identifying associations on the sex chromosomes [10, 47] and analyzing rare SNPs [17] requires more intricate methods beyond the scope of the workshop.

⁵ <https://cran.r-project.org/doc/manuals/r-release/R-intro.pdf>.

⁶ <https://www.cog-genomics.org/plink/>.

⁷ <https://www.cog-genomics.org/plink/2.0/>.

⁸ <https://www.cog-genomics.org/plink/1.9/filter>.

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 v1.9 command `-linear` or PLINK v2.0 command `-glm`. 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 using `-linear sex` in PLINK v1.9 or `-glm sex` in PLINK v2.

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. 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. Hardy-Weinberg equilibrium may be checked for the top SNP using `-hardy`.

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,

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 Fig. 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 [18, 49].

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 [20]. Further, two most commonly used data visualization plots in GWAS are introduced: the Manhattan plot and the Q-Q plot as shown in

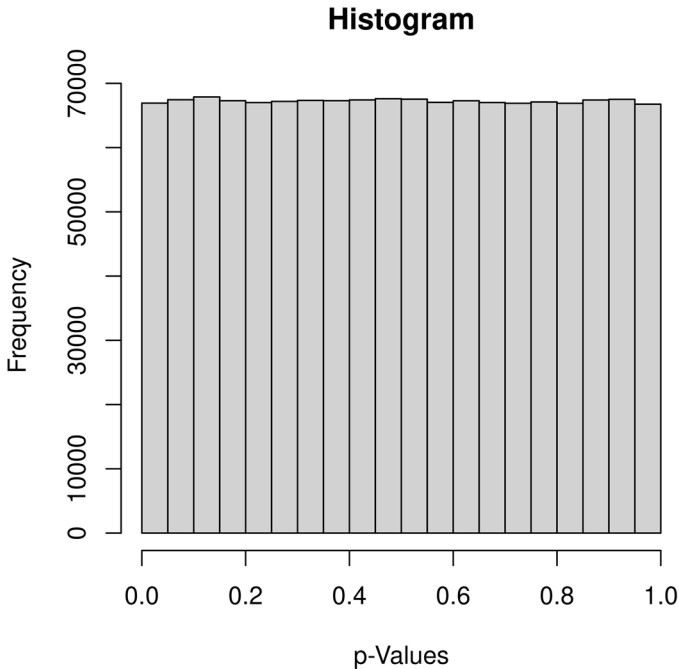


Fig. 1 Histogram of the 1,344,840 p-values from the YRI GWAS of the gene expression of *ERAP2*, obtained using the workshop materials. The histogram is close to $\text{Unif}(0, 1)$, the expected distribution of p-values under the null of no association

Fig. 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 Fig. 2 is a standard statistical plot, showing the quantiles of the observed p-values against those of $\text{Unif}(0,1)$, on 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 Fig. 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 Fig. 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.

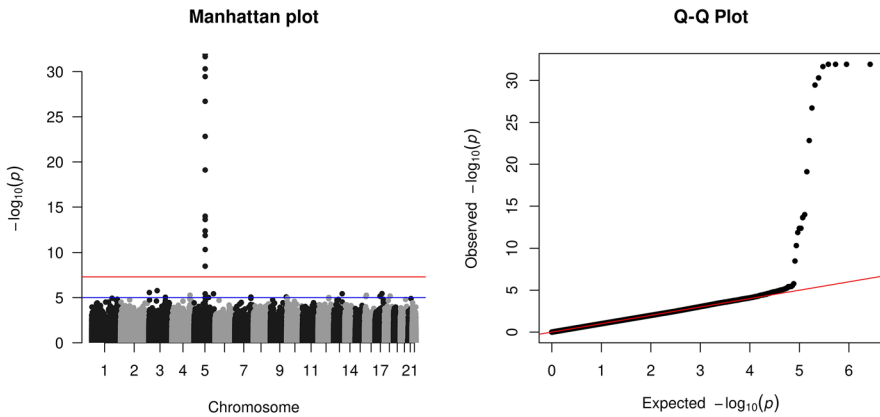


Fig. 2 The Manhattan plot and Q-Q plot of the 1,344,840 p-values from the YRI GWAS of the gene expression of *ERAP2*, obtained using the workshop materials

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 [42]. 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 [3]. The implementation steps associated with the *ERAP2* example are included in the appended workshop manual. Although LD is akin to the statistical concept of correlation, it is an 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 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 1 for the application form. Post-workshop, a survey was conducted to collect participant feedback; see Appendix 1 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 overall 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 discussion; 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 catalog⁹ [4] and to present a 3–5 min 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. Subsequently, notes were added to explain the difficult levels of the five lecture slide decks. 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.

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. Discussion around the cleaning of the 1000 Genomes data could be included in the morning lecture sessions, and cleaning steps for the 1000 Genomes individuals [40] 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 [11] matched with other 1 KG 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 UNIX-based.

⁹ <https://www.ebi.ac.uk/gwas/>.

To adhere to the current standard of reproducible research [33], 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. [11]. R, PLINK, and dataset versions were synchronized, and all scripts were version-controlled and hosted on the workshop GitHub. 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 also made publicly available on Zenodo.¹⁰

Appendix

A Phenotype Extraction and Dataset Generation

Phenotype files containing gene expression data must be matched to cleaned, independent 1000 Genomes individuals to create the datasets. Please refer to [Github.com/sugolov/GWAS-Workshop/Notebooks/DatasetPreparation.Rmd](https://github.com/sugolov/GWAS-Workshop/Notebooks/DatasetPreparation.Rmd) to create a phenotype file using the expression data from the University of Geneva Medical School [31, 43]. Phenotype files for *ERAP2* are provided for the CEU and YRI populations for single population and mixed population analysis on [Github.com/sugolov/GWAS-Workshop/Datasets](https://github.com/sugolov/GWAS-Workshop/Datasets). The PLINK binary format genotype files of independent samples were downloaded from <http://tcag.ca/tools/1000genomes.html> [40]. Refer to [Github.com/sugolov/GWAS-Workshop/Notebooks/YRI_Analysis.Rmd](https://github.com/sugolov/GWAS-Workshop/Notebooks/YRI_Analysis.Rmd) to match the phenotype files with the 1 KG dataset from The Centre for Applied Genomics [2, 40]. For more information on required files and phenotype/genotype combination, please visit Sections 1.4 and 4.2 respectively of the manual, which can be found on the Github.

B High Coverage Dataset

The High Coverage dataset was generated using the 30x High Coverage samples from the New York Genome Center (NYGC) [6]. Please refer to <https://github.com/sugolov/GWAS-Workshop/tree/master/Notebooks> 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.

¹⁰ <https://zenodo.org/record/7809150>.

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 see 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 see 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 see 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 favorite 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 (i.e., 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?

Acknowledgements 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.

Declarations

Conflict of interest The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.




References

1. Abdi H, Williams LJ (2010) Principal component analysis. *Wiley interdisciplinary reviews: computational statistics* 2(4):433–459
2. Auton A, Abecasis GR, Altshuler DM, Durbin RM, Bentley DR, Chakravarti A, Clark AG, Donnelly P, Eichler EE, Flicek P, Gabriel SB, Gibbs RA, Green ED, Hurles ME, Knoppers BM, Korbel JO, Lander ES, Lee C, Lehrach H, Gravel S, (2015) A global reference for human genetic variation. *Nature* 526(7571): 68–74. 10.1038/nature15393

3. Boughton AP, Welch RP, Flickinger M, VandeHaar P, Taliun D, Abecasis GR, Boehnke M (2021) ‘LocusZoom.js: interactive and embeddable visualization of genetic association study results’, *Bioinformatics* . <https://doi.org/10.1093/bioinformatics/btab186>
4. Buniello A, MacArthur JAL, Cerezo M, Harris LW, 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
5. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, Motyer A, Vukcevic D, Delaneau O, O’Connell J, Cortes A (2018) The UK Biobank resource with deep phenotyping and genomic data. *Nature* 562(7726):203–209
6. Byrska-Bishop M, Evani US, Zhao X, Basile AO, Abel HJ, Regier AA, Corvelo A, Clarke WE, Musunuri R, Nagulapalli K, Fairley S, Runnels A, Winterkorn L, Lowy-Gallego E, Consortium THGSV, Flicek P, Germer S, Brand H, Hall IM, Talkowski ME, Narzisi G, Zody MC (2021) ‘High coverage whole genome sequencing of the expanded 1000 genomes project cohort including 602 trios’, *bioRxiv* . <https://www.biorxiv.org/content/early/2021/02/07/2021.02.06.430068>
7. Carr D, Lewin-Koh N, Maechler M, Sarkar D (2021) hexbin: Hexagonal Binning Routines. R package version 1.28.2. <https://CRAN.R-project.org/package=hexbin>
8. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ (2015) Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* 4(1):s13742-015. <https://doi.org/10.1186/s13742-015-0047-8>
9. Chanock SJ, Manolio T, Boehnke M, Boerwinkle E, Hunter DJ, Thomas G, Hirschhorn JN, Abecasis GR, Altshuler D, Bailey-Wilson JE et al (2007) Replicating genotype-phenotype associations. *Nature* 447(7145):655–660
10. Chen B, Craiu RV, Strug LJ, Sun L (2021) The x factor: a robust and powerful approach to x-chromosome-inclusive whole-genome association studies. *Genetic Epidemiol* 45(7):694–709
11. Cheung VG, Spielman RS, Ewens KG, Weber TM, Morley M, Burdick JT (2005) Mapping determinants of human gene expression by regional and genome-wide association. *Nature* 437(7063):1365–1369
12. 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
13. Cordell HJ, Clayton DG (2005) Genetic association studies. *The Lancet* 366(9491):1121–1131
14. Crow JFJF, Kimura M (1970) An introduction to population genetics theory. Harper and Row, New York
15. Cummiskey K, Adams B, Pleuss J, Turner D, Clark N, Watts K (2020) Causal inference in introductory statistics courses. *J Stat Educ* 28(1):2–8
16. Das S, Forer L, Schönherr S, Sidore C, Locke AE, Kwong A, Vrieze SI, Chew EY, Levy S, McGue M et al (2016) Next-generation genotype imputation service and methods. *Nat Genet* 48(10):1284–1287
17. Derkach A, Lawless JF, Sun L (2014) Pooled association tests for rare genetic variants: a review and some new results. *Stat Sci* 29(2):302–321
18. Devlin B, Roeder K, Wasserman L (2001) Genomic control, a new approach to genetic-based association studies. *Theor Popul Biol* 60(3):155–166
19. 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
20. Dudbridge F, Gusnanto A (2008) Estimation of significance thresholds for genomewide association scans. *Genet Epidemiol* 32(3):227–234
21. Gordon D, Finch SJ, Kim W (2020) ‘Heterogeneity in statistical genetics’
22. Higgins JP, Thompson SG (2002) Quantifying heterogeneity in a meta-analysis. *Stat Med* 21(11):1539–1558
23. Hu D, Ziv E (2008) Confounding in genetic association studies and its solutions. *Pharmacogenom Drug Discovery Develop* 448:31–39
24. Hudiburgh LM, Garbinsky D (2020) Data visualization: bringing data to life in an introductory statistics course. *J Stat Educ* 28(3):262–279
25. International HapMap Consortium (2007) A second generation human haplotype map of over 3.1 million snps. *Nature* 449(7164):851–861. <https://doi.org/10.1038/nature06258>
26. Lappalainen T, MacArthur DG (2021) From variant to function in human disease genetics. *Science* 373(6562):1464–1468

27. Lunshof JE, Chadwick R, Vorhaus DB, Church GM (2008) From genetic privacy to open consent. *Nat Rev Genet* 9(5):406–411
28. Maindonald JH (2008) Using R for Data Analysis and Graphics: Introduction. Code and Commentary, Maindonald, J H
29. Manolio TA (2010) Genomewide association studies and assessment of the risk of disease. *New England J Med* 363(2):166–176
30. Mares AT, de Kluiver H, Stringer S, Vorspan F, Curis E, Marie-Claire C, Derks EM (2018) A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. *Int J Methods Psychiatr Res* 27(2):e1608
31. Montgomery SB, Sammeth M, Gutierrez-Arcelus M, Lach RP, Ingle C, Nisbett J, Guigo R, Dermitzakis ET (2010) Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* 464(7289):773–777
32. Ostblom J, Timbers T (2022) Opinionated practices for teaching reproducibility: motivation, guided instruction and practice. *J Stat Data Sci Educ* 30(3):241–250
33. Peng RD (2011) Reproducible research in computational science. *Science* 334(6060):1226–1227
34. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 38(8):904–909
35. Purcell S, Chang C (2021) ‘Plink 1.90b6.24’. www.cog-genomics.org/plink/1.9/
36. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, De Bakker PI, Daly MJ, Sham PC (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Human Genet* 81(3):559–575
37. Core Team R (2021) R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>
38. Reich D, Price AL, Patterson N (2008) Principal component analysis of genetic data. *Nat Genet* 40(5):491–492
39. Risch N, Merikangas K (1996) The future of genetic studies of complex human diseases. *Science* 273(5281):1516–1517
40. Roslin NM, Weili L, Paterson AD, Strug LJ (2016) ‘.5 genotypes’, bioRxiv . <https://www.biorxiv.org/content/early/2016/09/30/078600>
41. Shaffer JP (1995) Multiple hypothesis testing. *Ann Rev Psychol* 46(1):561–584
42. Slatkin M (2008) Linkage disequilibrium—understanding the evolutionary past and mapping the medical future. *Nat Rev Genet* 9(6):477–485
43. Stranger BE, Montgomery SB, Dimas AS, Parts L, Stegle O, Ingle CE, Sekowska M, Smith GD, Evans D, Gutierrez-Arcelus M, Price A, Raj T, Nisbett J, Nica AC, Beazley C, Durbin R, Deloukas P, Dermitzakis ET (2012) Patterns of cis regulatory variation in diverse human populations. *PLoS Genet* 8(4):1–13. <https://doi.org/10.1371/journal.pgen.1002639>
44. Tan VY, Timpson NJ (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. *Ann Rev Genom Human Genet* 23:569–589
45. Turner SD (2014) qqman: an R package for visualizing GWAS results using QQ and manhattan plots. *J Open Source Softw.* <https://doi.org/10.1101/005165>
46. Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, Yang J (2017) 10 years of gwas discovery: biology, function, and translation. *Am J Human Genet* 101(1):5–22
47. Wang Z, Sun L, Paterson AD (2022) Major sex differences in allele frequencies for x chromosomal variants in both the 1000 genomes project and gnomad. *PLoS Genet* 18(5): 1010231
48. Wickham H (2016) ggplot2: Elegant Graphics for Data Analysis. Springer, New York
49. Yang J, Weedon MN, Purcell S, Lettre G, Estrada K, Willer CJ, Smith AV, Ingelsson E, O’connell JR, Mangino M et al (2011) Genomic inflation factors under polygenic inheritance. *Eur J Human Genet* 19(7):807–812
50. Zhang L, Strug LJ, Sun L (2023) Leveraging Hardy-Weinberg disequilibrium for association testing in case-control studies. *Ann Appl Stat* 17(2):1764–1781
51. Zhou X, Stephens M (2012) Genome-wide efficient mixed-model analysis for association studies. *Nature Genet* 44(7):821–824

Authors and Affiliations

Anton Sugolov¹  · Eric Emmenegger² · Andrew D. Paterson^{3,4}  · Lei Sun^{4,5} 

✉ Andrew D. Paterson
andrew.paterson@sickkids.ca

✉ Lei Sun
lei.sun@utoronto.ca

Anton Sugolov
anton.sugolov@mail.utoronto.ca

Eric Emmenegger
eric.emmenegger@mail.utoronto.ca

- ¹ Department of Mathematics, Faculty of Arts and Sciences, University of Toronto, Toronto, Canada
- ² Department of Mechanical and Industrial Engineering, University of Toronto, Toronto, Canada
- ³ Program in Genetics & Genome Biology The Hospital for Sick Children, University of Toronto, Toronto, ON, Canada
- ⁴ Dalla Lana School of Public Health, University of Toronto, Toronto, Canada
- ⁵ Department of Statistical Sciences, Faculty of Arts and Sciences, Dalla Lana School of Public Health, University of Toronto, Toronto, Canada