

# A Preliminary Outline of the Statistical Inference Process in Genetic Association Studies

Tapati Basak, Nipa Roy

Department of Statistics, Jahangirnagar University, Dhaka, Bangladesh  
Email: [tapati.basak@juniv.edu](mailto:tapati.basak@juniv.edu), [nipaniir@gmail.com](mailto:nipaniir@gmail.com)

**How to cite this paper:** Basak, T. and Roy, N. (2022) A Preliminary Outline of the Statistical Inference Process in Genetic Association Studies. *Open Journal of Statistics*, 12, 200-209.

<https://doi.org/10.4236/ojs.2022.122014>

**Received:** February 20, 2022

**Accepted:** April 11, 2022

**Published:** April 14, 2022

Copyright © 2022 by author(s) and Scientific Research Publishing Inc. This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

## Abstract

The genome-wide association study (GWAS) is a powerful experimental design that is applied to detect disease susceptible genetic variants. The main goal of these studies is to provide a better understanding of the biology of disease, which further facilitates prevention or better treatment. A statistical inferential process is finally carried out in this study, where an association is usually observed between the single-nucleotide polymorphism (SNPs) and the traits in a case-control setting. To detect the disease responsible loci correctly, the investigation of the statistical association should be carefully conducted along with the other necessary steps. This research provides an introductory guideline for conducting such statistical association tests for these studies using SNP genotype data.

## Keywords

GWAS, Statistical Inference, Association, SNP, Case-Control Study

## 1. Introduction

A genome-wide association study (GWAS) is an inclusive genetic analysis to identify associations between specific genetic variations in the form of single-nucleotide polymorphism (SNPs) and phenotypic traits. These studies are very effective in genetic epidemiology as they provide a relatively superficial approach for detecting potential genetic contributors to common and complex diseases using a simple case-control study model [1]-[6].

Correct performance of such genetic association studies requires interdisciplinary knowledge. Specifically, knowledge of genetics, statistics, and bioinformatics are the primary key [6]. In this pathway, in-depth knowledge of the genetic architecture of the human genome was provided by two important re-

search initiatives, the International HapMap Project and the 1000 Genomes project. The International HapMap Project [7] described the patterns of common SNPs within the human DNA (deoxyribonucleic acid) sequence whereas the 1000 Genomes (1KG) project [8] provided a map of both common and rare SNPs [6].

In such studies, very large sample sizes are required to identify and validate findings. The careful attention to data quality has been appreciated as even small sources of systematic or random error can cause spurious results. Hence, a number of strategies for quality control have been developed [6] [9] [10].

Along with these quality control and quality assurance of genotypic data, appropriate statistical association testing will need to be carefully conducted using sophisticated and dedicated genetics software [6].

The method of presenting the GWAS finding regarding the reporting of disease-associated or risk markers are quite different from most clinical or epidemiological studies. Particularly,  $p$ -values of a single SNP test along with its associated odds ratio are emphasized in case of results presentation [5] [11].

Here, the likelihood of the odd-ratios between two different alleles being statistically different than one is reported by the  $p$ -values in a GWAS, where the typical threshold of significance level for the most published GWAS is  $p = 5 \times 10^{-08}$  [5].

This study focuses on the statistical analysis after genotyping calls are made and quality control and assurance measures are taken [11]. The main objective of this paper is to provide an overview of some introductory statistical analysis of GWAS using the SNP genotype DNA sequencing data.

## 2. GWAS Data Preparation

For the independence tests, the SNP genotype data for each gene were generated for 3000 individuals using computer simulation in R-programming language by assigning the equal probability for both the cases (0.5) and the controls (0.5) (Data 1). Another data containing the GWAS results was also generated via computer simulation in R, which is a replica of “PLINK assoc” output (<https://zzz.bwh.harvard.edu/plink/anal.shtml>) containing the following information (Data 2).

---

chr Chromosome number that defaults to PLINK’s “CHR”.

bp Physical position (base-pair). This is the chromosomal position. (defaults to PLINK’s “BP”).

$p$  The  $p$ -value from the association tests that defaults to PLINK’s “P”.

snp SNP identifier or rsID (Reference SNP cluster ID) which defaults to PLINK’s “SNP”.

---

## 3. Demonstration of the Statistical Analysis

### 3.1. Testing Association

GWAS generally tests association for a single SNP which is a contingency table

test of genotype counts and disease phenotype. For example, for a SNP with major allele “A” and minor allele “a”, the genotype counts ( $A/A$ ,  $A/a$  and  $a/a$ ) can be presented in a  $2 \times 3$  contingency table along with a binary disease status or phenotype (case-control) (**Table 1**) [12].

It is expected that the relative allele or genotype frequencies to be the same in case and control groups under the null hypothesis of no association. Usually, the association for a given contingency table is tested by the simple chi-squared ( $\chi^2$ ) [13] test with two degrees of freedom (d.f.), and the  $p$ -value is recorded. Each of these  $p$ -values is compared with the GWAS typical threshold of significance level, which is  $p = 5 \times 10^{-8}$  for the most published GWAS [5].

For the practical application, this contingency table test ( $\chi^2$ -test) was firstly performed for evaluating each of the three randomly selected genes from the genotype data as described in Section 2 (Data 1). The conventional  $\chi^2$ -test was applied for each of SNPs contained in a gene, where the GENE1, GENE2 and GENE3 have 3, 5 and 10 SNPs, respectively. For example, the GENE1 contains 3 SNPs, the three individual  $\chi^2$ -tests were performed for each of the 3 SNPs, and the  $p$ -values were recorded. The tests for the other two genes were also performed in the same manner. The  $p$ -values for these three genes are presented in **Table 2**.

Comparing the  $p$ -values to the GWAS threshold  $p = 5 \times 10^{-8}$ , there is one SNP from the GENE1 is associated with the disease phenotype having  $p = 5.243314 \times 10^{-10}$  (**Table 2**). Hence, GENE1 is associated with the disease phenotype.

**Table 1.** A  $2 \times 3$  table of genotype counts for a single SNP with disease status.

	$A/A$	$A/a$	$a/a$
Case	$n_{11}$	$n_{12}$	$n_{13}$
Control	$n_{21}$	$n_{22}$	$n_{23}$

**Table 2.** The  $p$ -values from the simple  $\chi^2$ -test for the 3 genes.

GENE1	GENE2	GENE3
$1.564368 \times 10^{-07}$	0.02971920	0.01076871
$5.243314 \times 10^{-10}$	0.52543425	0.59685551
$3.288859 \times 10^{-07}$	0.43050584	0.53037024
	0.97167750	0.53108175
	0.08172946	0.80016099
		0.05465315
		0.35180821
		0.20358497
		0.03885251
		0.86894928

### 3.2. Models for the Association Tests

The conventional  $\chi^2$ -test does not include the sense of the genotype ordering (trend). Here, each of the genotypes is assumed to have an independent association with disease phenotype. But, these ordering could be included in the association tests of contingency tables by considering the disease penetrance. The penetrance function is an approach for modeling the relation between SNPs and risk of a given disease with the consideration of genotype ordering [14] [15] [16].

For a single diallelic SNP with alleles “*A*” (major, disease responsible) and “*a*” (minor, normal), the unordered genotype counts are presented in **Table 1**. For a disease status the risk factor is defined by the genotype or allele at a specific marker. Thus, the disease penetrance associated with a given genotype is the risk of disease of the individuals carrying that genotype or allele. This risk of carrying disease responsible genotype could be measured by defining the probabilistic functions, which intern define the conditional probabilities of being affected with a given disease conditional on carrying a specific genotype [12] [17] [18].

For the genotype counts as shown in **Table 1**, the three models can be defined in terms of a genetic penetrance parameter  $\gamma$  ( $\gamma > 1$ ). An additive model implies that risk of developing disease is increased  $\gamma$ -fold for the genotype *A/a* and by  $2\gamma$ -fold for the genotype *A/A*. A recessive model indicates that two copies of allele “*A*” are required for an  $\gamma$ -fold increase in disease risk, and a dominant model specifies that either one or two copies of allele “*A*” are required for an  $\gamma$ -fold increase in disease risk. An intuitive measure of the strength of an association is the relative risk (RR). In this genetic association analysis, each genetic model can be represented with the relation to this genotypic relative risks (GRR) under the assumption of phenocopies, where the GRR presents the increased risk of an individual having a disease responsible genotype over a person without it [9] [19].

Generally, the models should be chosen based on the mode of inheritance (dominant, additive and recessive). But, a common problem is the lack of knowledge concerning the mode of inheritance. Assumption of the incorrect mode of inheritance may lead to significant loss of power. Also, testing for all the possible models may increase type I error rate. Some studies have proposed ways to determine the robust procedures which will correctly specify the underlying model of inheritance. Also, these methods perform the analysis by maximizing the power and preserving the nominal type I error rate. The proposed methods are based on the theory of efficiency robust procedures, the deviations from Hardy-Weinberg equilibrium (HWE) and a combination of test statistics for the selection of the underlying genetic model [20] [21].

Now, the practical application of these concepts of penetrance in an association study using the contingency table can be demonstrated in one of the two ways. In one approach, these models of penetrance could be included in the contingency table analysis by rearranging the genotype counts according to the mode of inheritance (additive, dominant and recessive) [9] [19] [22]. The other

way is to define the penetrance models that will specify the trends of risk with increasing numbers of disease responsible allele. Hence, the association could be tested by the Cochran-Armitage trend test for the additive, dominant and recessive models [9] [23] [24]. Tests with an additive model are common in GWAS when the underlying genetic model is unknown. This is because this model has reasonable power to detect both additive and dominant effects [9] [19].

For the genotype data (Data 1), the association test was further performed by the Cochran-Armitage trend test for additive model. Single SNP test was applied for each of the three genes (GENE1, GENE2 and GENE3), and the  $p$ -values were recorded (Table 3).

From the results of the Cochran-Armitage trend test, it was observed that the two SNPs from the GENE1 having the  $p = 6.284831 \times 10^{-11}$  and  $8.586496 \times 10^{-08}$ , respectively, are significant according to the GWAS threshold  $p = 5 \times 10^{-08}$ , and hence associated with the disease phenotype (Table 3). So, the GENE1 is to be reported as positive among the three.

On the other hand, the  $p$ -values from the two tests (the conventional  $\chi^2$ -test and the Cochran-Armitage trend test) are completely different from each other (Table 2 and Table 3). That is, considering the order of the genotypes is producing the different outputs as compared to the unordered case. More specifically, overall the trend test is producing smaller  $p$ -values except for some cases. Though, the GENE1 is resulting in significant association in both of the two tests. But, the conventional  $\chi^2$ -test is showing only one significant SNP whereas the Cochran-Armitage trend test is resulting in two significant SNPs. The SNP2 is significant for both the tests but the Cochran-Armitage trend test is producing smaller  $p$ -value (Table 2 and Table 3).

### 3.3. Multiple Testing Corrections

The GWAS evaluates several thousand of genes simultaneously under different conditions over the genome, where each gene consists of a different number of SNPs. Hence, such association studies consider huge number of simultaneous testing of the null hypothesis, which constitutes multiple testing.

In order to control the type I error rate accurately, an adjustment is required for the  $p$ -values obtained from such simultaneous testing process, because detection of false positives may occur in such microarray data analysis. Here, the false positives are genes that are found to be statistically different between conditions, but are not in reality.

Different types of multiple testing corrections include Bonferroni, Holm, Benjamini and Hochberg False Discovery Rate, etc. [9] [25] [26]. All of these methods have some underlying principles to be applied in practice.

As a practical application, the Bonferroni correction is applied for the marginal  $p$ -values obtained from both the tests (the conventional  $\chi^2$ -test and the Cochran-Armitage trend test) for the significant gene (GENE1). In this approach, the  $p$ -values are multiplied by the number of comparisons. The corrected  $p$ -values are presented in Table 4.

**Table 3.** The  $p$ -values from the Cochran-Armitage trend test for the 3 genes.

GENE1	GENE2	GENE3
$1.381782 \times 10^{-07}$	0.008027006	0.003282794
$6.284831 \times 10^{-11}$	0.379414203	0.911031670
$8.586496 \times 10^{-08}$	0.198701794	0.443972876
	0.836489351	0.324895632
	0.371924565	0.860287464
		0.326049890
		0.166654955
		0.297718380
		0.021493524
		0.991691621

**Table 4.** The Bonferroni corrected  $p$ -values for both the conventional  $\chi^2$ -test and the Cochran-Armitage trend test for the significant gene, GENE1.

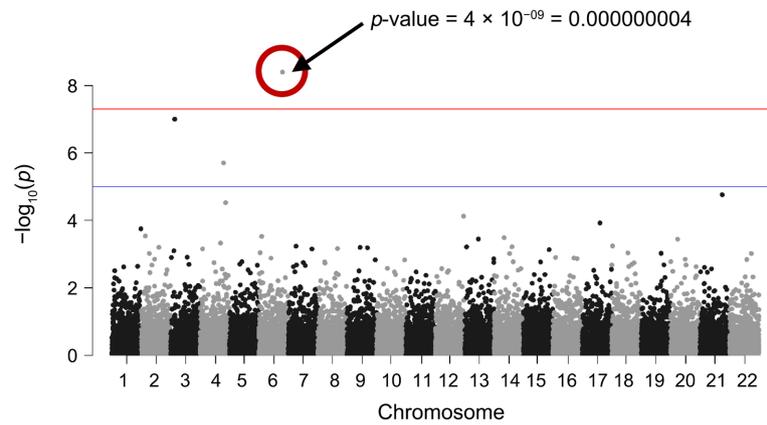
The conventional $\chi^2$ -test		The Cochran-Armitage trend test	
Marginal $p$ -values	Corrected $p$ -values	Marginal $p$ -values	Corrected $p$ -values
$1.564368 \times 10^{-07}$	$4.693104 \times 10^{-07}$	$1.381782 \times 10^{-07}$	$4.145346 \times 10^{-07}$
$5.243314 \times 10^{-10}$	$1.572994 \times 10^{-09}$	$6.284831 \times 10^{-11}$	$1.885449 \times 10^{-10}$
$3.288859 \times 10^{-07}$	$9.866577 \times 10^{-07}$	$8.586496 \times 10^{-08}$	$2.575949 \times 10^{-07}$

Overall, the  $p$ -values are changed after the correction. Specifically, this correction method provides quite larger  $p$ -values as compared to the marginal values. For the same gene (GENE1), the number of the significant SNPs remains the same for the conventional  $\chi^2$ -test. But, the scenario is changed for the Cochran-Armitage trend test. For this case, there were two significant SNPs before the correction whereas only one SNP is showing association after the correction is performed (Table 4).

#### 4. Manhattan Plot

A commonly used plot in most GWAS to display the significant SNPs in terms of  $p$ -values summarizes the results of the millions of tests, which have been performed. It is also a presentation of the  $p$ -values of the entire GWAS on a genomic scale. The horizontal  $x$ -axis is a map of the genome (genomic coordinates) organized from left to right by chromosome, and within chromosome by location. The different colors of each block usually show the extent of each chromosome. Each dot on the Manhattan plot signifies a SNP. The vertical location of the SNP along the  $y$ -axis is its  $p$ -value for the correlation of itself with the phenotype. The  $p$ -values are negative-log-transformed (that is the  $-\log_{10}(p)$  label on the  $y$ -axis) so that the smaller values are higher on the plot [27].

Figure 1 is presenting the Manhattan plot of all the  $p$ -values from the PLINK output (Data 2). The red line indicates the threshold for genome-wide significance ( $p = 5 \times 10^{-08}$ ), and the blue line for suggestive associations ( $p = 1 \times 10^{-05}$ ).



**Figure 1.** Manhattan plot of all the  $p$ -values for the GWAS results (Data 2).

There is one SNP that crossed the red line (the solid grey dot circled by a red border, **Figure 1**). That is, this SNP is significant according to the GWAS threshold. Based on the information of GWAS results (Data 2), this significant SNP is on chromosome 6 with the  $p$ -value  $4 \times 10^{-09}$ . As the  $y$ -axis presents the  $-\log_{10}(p)$ , which is equivalent to the number of zeros after the decimal point plus one. For example, see the presentation of the  $p$ -value  $= 4 \times 10^{-09}$  for the significant SNP indicated in a red circle in **Figure 1**. On the other hand, two SNPs on chromosomes 3 and 4 (the black and the grey dots between the blue and the red lines) have crossed the suggestive significance level (the blue line) having the  $p$ -values  $1 \times 10^{-07}$  and  $2 \times 10^{-06}$ , respectively.

## 5. Dedicated Software for GWAS

The open-source statistical software like R [28] can be used for performing and visualizing all of these analyses as mentioned in this paper. But, there are some customized and dedicated GWAS software. For example, PLINK [29] is the most popular and computationally efficient software program that offers an inclusive and properly documented set of an automated GWAS analysis including the quality control, association testing, etc. The open source software PLINK is written in C++ and can be installed on Windows, Mac and UNIX machines.

## 6. Conclusion

A practical guideline for the GWAS association testing is provided in this paper. Although this application considers simulated data of SNP genotype and the GWAS results, the real data sets can also be handled in similar ways as outlined here. All of these theoretical contexts of statistical association testing along with the practical application would be made GWAS more accessible to statistical researchers without having any formal training in this field.

## Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

## References

- [1] Tängdén, T., Gustafsson, S., Rao, A.S. and Ingelsson, E. (2022) A Genome-Wide Association Study in a Large Community-Based Cohort Identifies Multiple Loci Associated with Susceptibility to Bacterial and Viral Infections. *Scientific Reports*, **12**, Article No. 2582. <https://doi.org/10.1038/s41598-022-05838-z>
- [2] Uffelmann, E., Huang, Q.Q., Munung, N.S., Vries, J., Okada, Y., Martin, A.R., Martin, H.C., Lappalainen, T. and Posthuma, D. (2021) Genome-Wide Association Studies. *Nature Reviews Methods Primers*, **1**, Article No. 59. <https://doi.org/10.1038/s43586-021-00056-9>
- [3] Loos, R.J.F. (2020) 15 Years of Genome-Wide Association Studies and No Signs of Slowing Down. *Nature Communications*, **11**, Article No. 5900. <https://doi.org/10.1038/s41467-020-19653-5>
- [4] Beck, T., Shorter, T. and Brookes, A.J. (2020) GWAS Central: A Comprehensive Resource for the Discovery and Comparison of Genotype and Phenotype Data from Genome-Wide Association Studies. *Nucleic Acids Research*, **48**, D933-D940. <https://doi.org/10.1093/nar/gkz895>
- [5] Patron, J., Serra-Cayuela, A., Han, B., Li, C. and Wishart, D.S. (2019) Assessing the Performance of Genome-Wide Association Studies for Predicting Disease Risk. *PLoS ONE*, **14**, e0220215. <https://doi.org/10.1371/journal.pone.0220215>
- [6] Marees, A.T., Kluiiver, H.D., Stringer, S., Vorspan, F., Curis, E., Marie-Claire, C. and Derks, E.M. (2017) A Tutorial on Conducting Genome-Wide Association Studies: Quality Control and Statistical Analysis. *International Journal of Methods in Psychiatric Research*, **27**, e1608. <https://doi.org/10.1002/mpr.1608>
- [7] The International HapMap Consortium (2003) The International HapMap Project. *Nature*, **426**, 789-796. <https://doi.org/10.1038/nature02168>
- [8] The 1000 Genomes Project Consortium (2010) A Map of Human Genome Variation from Population Scale Sequencing. *Nature*, **467**, 1061-1073. <https://doi.org/10.1038/nature09534>
- [9] Clarke, G.M., Anderson, C.A., Pettersson, F.H., Cardon, L.R., Morris, A.P. and Zondervan, K.T. (2011) Basic Statistical Analysis in Genetic Case-Control Studies. *Nature Protocols*, **6**, 121-133. <https://doi.org/10.1038/nprot.2010.182>
- [10] Laurie, C.C., Doheny, K.F., Mirel, D.B., Pugh, E.W., Bierut, L.J., Bhangale, T., Boehm, F., Caporaso, N.E., Cornelis, M.C., Edenberg, H.J., Gabriel, S.B., Harris, E.L., Hu, F.B., Jacobs, K., Kraft, P., Landi, M.T., Lumley, T., Manolio, T.A., McHugh, C., Painter, I., Paschall, J., Rice, J.P., Rice, K.M., Zheng, X. and Weir, B.S. (2010) Quality Control and Quality Assurance in Genotypic Data for Genome-Wide Association Studies. *Genetic Epidemiology*, **34**, 591-602. <https://doi.org/10.1002/gepi.20516>
- [11] Reed, E., Nunez, S., Kulp, D., Qian, J., Reilly, M.P. and Foulkes, A.S. (2015) A Guide to Genome-Wide Association Analysis and Post-Analytic Interrogation. *Statistics in Medicine*, **34**, 3769-3792. <https://doi.org/10.1002/sim.6605>
- [12] Setu, T.J. and Basak, T. (2021) An Introduction to Basic Statistical Models in Genetics. *Open Journal of Statistics*, **11**, 1017-1025. <https://doi.org/10.4236/ojs.2021.116060>
- [13] Plackett, R.L. (1983) Karl Pearson and the Chi-Squared Test. *International Statistical Review*, **51**, 59-72. <https://www.jstor.org/stable/1402731>  
<https://doi.org/10.2307/1402731>
- [14] Moore, J.H., Hahn, L.W., Ritchie, M.D., Thornton, T.A. and White, B.C. (2004)

- Routine Discovery of Complex Genetic Models Using Genetic Algorithms. *Applied Soft Computing*, **4**, 79-86. <https://doi.org/10.1016/j.asoc.2003.08.003>
- [15] Cooper, D.N., Krawczak, M., Polychronakos, C., Tyler-Smith, C. and Kehrer-Sawatzk, H. (2013) Where Genotype Is Not Predictive of Phenotype: Towards an Understanding of the Molecular Basis of Reduced Penetrance in Human Inherited Disease. *Human Genetics*, **132**, 1077-1130. <https://doi.org/10.1007/s00439-013-1331-2>
- [16] Ford, D., Easton, D.F., Stratton, M., Narod, S., Goldgar, D., Devilee, P., Bishop, D.T., Weber, B., Lenoir, G., Chang-Claude, J., Sobol, H., Teare, M.D., Struewing, J., Arason, A., Scherneck, S., Peto, J., Rebbeck, T.R., Tonin, P., Neuhausen, S., Barkardottir, R., Eyfjord, J., Lynch, H., Ponder, B.A.J., Gayther, S.A., Birch, J.M., Lindblom, A., Stoppa-Lyonnet, D., Bignon, Y., Borg, A., Hamann, U., Haites, N., Scott, R.J., Maugard, C.M., Vasen, H., Seitz, S., Cannon-Albright, L.A., Schofield, A., Zelada-Hedman, M. and The Breast Cancer Linkage Consortium (1998) Genetic Heterogeneity and Penetrance Analysis of the BRCA1 and BRCA2 Genes in Breast Cancer Families. *American Journal of Human Genetics*, **62**, 676-689. <https://doi.org/10.1086/301749>
- [17] Ziegler, A. and König, I.R. (2010) A Statistical Approach to Genetic Epidemiology: Concepts and Applications. Wiley-VCH, Weinheim. <https://www.10.1002/9783527633654>
- [18] Gong, G., Hannon, N. and Whittemore, A.S. (2010) Estimating Gene Penetrance from Family Data. *Genetic Epidemiology*, **34**, 373-381. <https://doi.org/10.1002/gepi.20493>
- [19] Bush, W.S. and Moore, J.H. (2012) Chapter 11: Genome-Wide Association Studies. *PLOS Computational Biology*, **8**, e1002822. <https://doi.org/10.1371/journal.pcbi.1002822>
- [20] Bagos, P.G. (2013) Genetic Model Selection in Genome-Wide Association Studies: Robust Methods and the Use of Meta-Analysis. *Statistical Applications in Genetics and Molecular Biology*, **12**, 285-308. <https://doi.org/10.1515/sagmb-2012-0016>
- [21] Joo, J., Kwak, M. and Zheng, G. (2010) Improving Power for Testing Genetic Association in Case-Control Studies by Reducing the Alternative Space. *Biometrics*, **66**, 266-276. <https://doi.org/10.1111/j.1541-0420.2009.01241.x>
- [22] Horita, N. and Kaneko, T. (2015) Genetic Model Selection for a Case-Control Study and a Meta-Analysis. *Meta Gene*, **5**, 1-8. <https://doi.org/10.1016/j.mgene.2015.04.003>
- [23] Armitage, P. (1955) Tests for Linear Trends in Proportions and Frequencies. *Biometrics*, **11**, 375-386. <https://www.jstor.org/stable/3001775>  
<https://doi.org/10.2307/3001775>
- [24] Cochran, W.G. (1954) Some Methods for Strengthening the Common Chi-Squared Test. *Biometrics*, **10**, 417-451. <https://www.jstor.org/stable/3001616>  
<https://doi.org/10.2307/3001616>
- [25] Pascovici, D., Handler, D.C.L., Wu, J.X. and Haynes, P.A. (2016) Multiple Testing Corrections in Quantitative Proteomics: A Useful but Blunt Tool. *Proteomics*, **16**, 2448-2453. <https://doi.org/10.1002/pmic.201600044>
- [26] Noble, W.S. (2009) How Does Multiple Testing Correction Work? *Nature Biotechnology*, **27**, 1135-1137. <https://doi.org/10.1038/nbt1209-1135>
- [27] Gibson, J., Russ, T.C., Clarke, T.K., Howard, D.M., Hillary, R.F., Evans, K.L., Walker, R.M., Bermingham, M.L., Morris, S.W., Campbell, A., Hayward, C., Murray, A.D., Porteous, D.J., Horvath, S., Lu, A.T., McIntosh, A.M., Whalley, H.C. and Marioni, R.E. (2019) A Meta-Analysis of Genome-Wide Association Studies of Epige-

netic Age Acceleration. *PLOS Genetics*, **15**, e1008104.

<https://doi.org/10.1371/journal.pgen.1008104>

- [28] R Development Core Team (2008) R: A Language and Environment for Statistical Computing. Reference Index: R Foundation for Statistical Computing.

<http://softlibre.unizar.es/manuales/aplicaciones/r/fullrefman.pdf>

- [29] Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., Bakker, P.I.W.de., Daly, M.J. and Sham, P.C. (2007) PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics*, **81**, 559-575.

<https://doi.org/10.1086/519795>