BIOINFORMATION
Discovery at the interface of physical and biological sciences

OPEN ACCESS GOLD

BIOMEDICAL INFORMATICS

www.bioinformation.net
Volume 16(5)

Views

Views on GWAS statistical analysis

Xiaowen Cao[1,3], Li Xing[2], Hua He[1], Xuekui Zhang[3,*]

[1]Department of Mathematics, Hebei University of Technology, Tianjin, China; [2]Department of Mathematics and Statistics, University of Saskatchewan, Saskatoon, SK, Canada; [3]Department of Mathematics and Statistics, University of Victoria, BC, Canada; Xuekui Zhang - xuekui@uvic.ca; *Corresponding author

Contacts: Xiaowen Cao - E-mail: 18920927551@163.com; Li Xing - E-mail: lix491@mail.usask.ca; Hua He - E-mail: hehua@hebut.edu.cn; Xuekui Zhang - E-mail: Xuekui@UVic.ca

Received April 2, 2020; Revised April 15, 2020; Accepted April 17, 2020; Published May 31, 2020

DOI: 10.6026/97320630016393

Declaration on Publication Ethics:
The authors state that they adhere with COPE guidelines on publishing ethics as described elsewhere at https://publicationethics.org/. The authors also undertake that they are not associated with any other third party (governmental or non-governmental agencies) linking with any form of unethical issues connecting to this publication. The authors also declare that they are not withholding any information that is misleading to the publisher in regard to this article.

Declaration on official E-mail:
The corresponding author declares that official e-mail from their institution is not available for all authors

Abstract:
Genome-wide association study (GWAS) is a popular approach to investigate relationships between genetic information and diseases. A number of associations are tested in a study and the results are often corrected using multiple adjustment methods. It is observed that GWAS studies suffer adequate statistical power for reliability. Hence, we document known models for reliability assessment using improved statistical power in GWAS analysis.

Keywords: Genome-Wide Association Studies; Single Nucleotide Polymorphisms; Statistical power, Multiple Testing Adjustment, Linkage Disequilibrium, Supervised Learning, Unsupervised Learning

Background:
Genome-wide association study (GWAS) is a popular approach to investigate associations between genetic information and diseases **[1]**. Known literature on GWAS is large (213,610 records) as of June 2020 in PubMed **[2]**. There were about 5687 GWAS studies (September 2018) submitted to the NHGRI-EBI GWAS Catalog **[3]**. This catalog documented 175,870 associations from 4439 studies (February 2020) **[4]**. The observed data contain the genotype of hundreds of thousands of Single Nucleotide Polymorphisms (SNPs). We test the association between disease outcomes for each SNP one by one. Hence, multiple testing adjustments are critical to control false positive results when many tests are conducted in a single study **[5]**. Typical multiple testing procedures include the Bonferroni correction with the GWAS threshold of 5x10e-8 towards

controlling the False Discovery Rate (FDR). A large proportion of the GWAS studies suffer lack of adequate statistical power due to large data dimensionality. Therefore, it is of interest to review approaches to address 'lack of statistical power' in GWAS analysis with large sample size.

Large GWAS dataset:
Data size is critical in GWAS analysis. The per-sample cost of genomic studies reduced substantially at a speed much faster than Moore's law **[6]** due to the advancement in high-throughput technologies such as the next generational sequencing **[32]**. This makes it possible to conduct genomic studies using large sample size. A number of such studies have been completed. For example, the UK Biobank Data **[7]** gleaned genomic data with related information from over 500,000 volunteers. Large proportion of meta-data (derived data) using the UK Biobank Data is available. The Electronic Medical Records and Genomics (eMERGE) network **[8]** is a National Human Genome Research Institute – funded consortium engaged in the development of methods and best practices to connect genomic data to electronic medical records. This study collected data for 39 million SNPs from 100,000 participants.

A number of software tools are available to analyze GWAS data. Plink **[9]** and snpStats **[10]** are the most popular tools. These tools implement SNP-wise (one-by-one) testing followed by multiple testing adjustments. Plink is a widely used toolset for GWAS. The basic association test is for a disease trait and is based on comparing allele frequencies of SNPs between cases and controls. Alternative tests are also implemented in Plink. These include the Cochran-Armitage trend test, Fisher's exact test, different genetic models (dominant, recessive and general), tests for stratified samples and a test for a quantitative trait. Multiple testing adjustments to control false positive probability are conducted in these tests for every SNP. Popular adjustment options (such as Bonferroni, Sidak, and FDR) are also implemented in Plink.

The snpStats is an R package in Bioconductor for GWAS. The snpStats can handle both quantitative and qualitative phenotypes. It can carry out single SNP tests adjusted for potential perplexing by quantitative and qualitative covariates. Tests having several SNPs taken together as 'tags' are also supported in these analyses. The snpStats package offers options for quality control using Hardy Weinberg equilibrium tests and filtering SNPs using minor allele frequencies. Similar to Plink, snpStats also offer popular multiple testing adjustments options. Plink and snpStats are freely downloaded and the detailed instructions of the various functions in the programs can be found in the respective user manuals.

Statistical models in GWAS:
The linkage disequilibrium:

Improving the statistical power using large sample size is not suitable for all genomic studies. This is because of majority of studies have enough samples to generate huge data as required. Therefore, it is desired to improve the study power using novel statistical analysis methods. The development of novel methods is gaining momentum over the last decades. Standard analysis method tests each SNP separately, but the SNPs are correlated with each other. The relationship between SNPs is called Linkage Disequilibrium (LD), which provides information for other SNPs that are in linkage with each other **[11]**. Most novel statistical models are developed by properly incorporating the LD relationship among SNPs to allow the tests use information from each other.

The number of parameters in the LD matrix is n(n-1)/2, where n is the number of SNPs being investigated. Hence, it is not realistic to obtain a precise estimation of LD matrix using moderate amount of samples in a genomic study. Hence, all reliable models incorporate LD information without clarity clearly. These do not use estimated LD matrix as model parameters as described below.

Supervised learning approaches:
The genomic information is the input data and the disease is the outcome in the association study using a supervised machine-learning model. There are various complex statistical models developed to improve the statistical power for SNP detection. We consider all of these methods as supervised learning methods, which comprises of SNP-set analysis **[12, 13, 14, 15]**, Penalized regression approach **[16, 17, 18]** and Bayesian hierarchical regression models **[19, 20, 21, 22]**.

Unsupervised learning approaches:
Model-based clustering is an unsupervised machine learning method, which can be used to group SNPs. The SNPs in the same group have similar relationship to the outcome, and could borrow information from each other in the GWAS analysis. A recent method proposed a one-step model. This simultaneously clusters SNP and detects significant SNP with FDR control **[23]**.

The patterns of clusters are specified by the difference in minor allele frequencies of SNPs between cases and controls. Thus, the pattern is enforced with a special prior distribution. This model-based clustering have shown more precise controls of FDR and higher statistical power in both simulation studies and real data analysis **[23]**. The limitation is that it can only handle case-control association studies.

Data splitting approach:
The other approach is based on data splitting strategy. The data can be randomly split into a screening set and a testing set. We use the screening set to remove the majority of SNPs with weak signals; and then investigate the retained SNPs in the testing set. The test

sets only consider a very small subset of SNPs. This leads to fewer penalties in the multiple test adjustment on testing set. So, this approach is much more powerful than analyzing the original data with all SNPs. The results of this type of analysis can be heavily affected by which samples are split into the testing set. We use resampling approaches to analyze multiple copies of the data with different random splits to remove unwanted 'split' effect **[24, 25]**. These methods are not popular since they have multiple critical disadvantages. First, multiple testing adjustment method is not available for controlling false positives in these methods. Second, such methods involve multiple tuning parameters whose values have to be selected in an ad-hoc manner.

Conclusion:

GWAS is a popular method to study genome relationship with diseases and their linked phenotypes. The adjustment of multiple testing is critical to reduce false positives in these studies. It is well realized that many GWAS studies suffer statistical power in SNP discovery due to high dimensional description of the problem at hand. Hence, it is of interest to document known information on large studies and known statistical analysis models that are pertinent to GWAS. Thus, application of novel statistical analysis models on large datasets using high performance computing (HPC) infrastructure is highly recommended.

Acknowledgement:

The authors acknowledge the Natural Sciences and Engineering Research Council Discovery Grants RGPIN-2017-04722 and the Canada Research Chair Grant 950-231363 (X.Z.)

Conflicts of Interest:

The authors declare no conflict of interest.

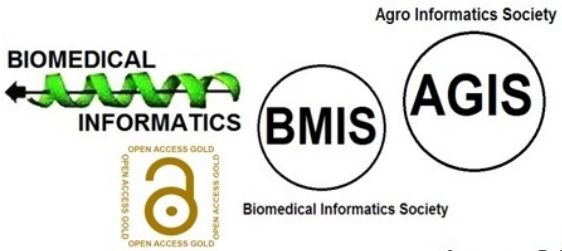
References:

[1] Manolio TA *N Engl J Med* 2010 **363:**166 [PMID: 20647212].
[2] https://pubmed.ncbi.nlm.nih.gov/?term=GWAS
[3] Buniello A *et al. Nucleic Acids Res* 2019 **47:** D1005-D1012 [PMID: 30445434].
[4] https://www.ebi.ac.uk/gwas.
[5] Chen SY *et al. J Thorac Dis* 2017 **9**:1725 [PMID: 28740688].
[6] https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost.
[7] Elliot LT *et al. Nature* 2018 **562:**210 [PMID: 30305740].
[8] Gottesman O *et al. Genet Med* 2013 **15:**761[PMID: 23743551].
[9] Chang CC *et al. GigaScience* 2015 **4:**7 [PMID: 25722852].
[10] Clayton D & Leung HT, *Hum Hered 2007* **64:**45 [PMID: 17483596].
[11] Slatkin M. *Nat Rev Genet* 2008 **9**:477 [PMID: 18427557].
[12] Cologne J *et al. BMC Genomics* 2018 **19:**524 [PMID: 29986644].
[13] Dai H *et al. PLoS One* 2013 **8**:e75897 [PMID: 24098741].
[14] Lu ZH *et al. Genet Epidemiol* 2015 **39:**664 [PMID: 26515609].
[15] Wu MC *et al. Am J Hum Genet* 2010 **86:**929 [PMID*:* 20560208].
[16] Kang HM *et al. Nat Genet* 2010 **42:**348 [PMID: 20208533].
[17] Yu J *et al. Nat Genet* 2006 **38**:203 [PMID: 16380716].
[18] Zhou X & Stephens M, *Nat Genet* 2012 **44**:821 [PMID: 22706312].
[19] Chen C *et al. Genetics* 2017 **206**:1791 [PMID: 28637709].
[20] Fernando RL & Garrick D, *Methods Mol Biol* 2013 **1019**:237 [PMID: 23756894].
[21] Sanyal N *et al. Bioinformatics 2019* **35:**1 [PMID: 29931045].
[22] Wang Q *et al. J Anim Breed Genet* 2016 **133:**253 [PMID: 26582716].
[23] Xu Y *et al. Sci Rep* 2019 **9:**13686 [PMID: 31548641].
[24] Kang G *et al. J Hum Genet 2015* **60:**729 [PMID: 26377241].
[25] Yang JJ *et al. Blood* 2012 **120:**4197 [PMID: 23007406].
[26] Bansal V *et al. Nat Rev Genet* 2010 **11:**773 [PMID: 20940738].
[27] Magrangeas F *et al. Clin Cancer Res* 2016 **22:**4350 [PMID: 27060151].
[28] Xu Z & Taylor JA, *Nucleic Acids Res* 2009 **37:** W600 [PMID: 19417063].
[29] Eskin E *Genome Res* 2008 **18:**653 [PMID: 18353808].
[30] Darnell G *et al. Bioinformatics* 2012 **28:** i147 [PMID: 22689754].
[31] Goode EL *Linkage Disequilibrium. In: Schwab M. (eds) Encyclopedia of Cancer.* 2011 [DOI: 10.1007/978-3-642-16483-5_3368].
[32] Xu et al, *Nat Commun* 2012 **3**, 1258 [DOI: 10.1038/ncomms2256].

Edited by P Kangueane

Citation: Cao ***et al.*** Bioinformation 16(5): 393-397 (2020)

Articles published in BIOINFORMATION are open for relevant post publication comments and criticisms, which will be published immediately linking to the original article for FREE of cost without open access charges. Comments should be concise, coherent and critical in less than 1000 words.



since 2005

indexed in