gJLS2: A generalized joint location and scale analysis tool for X-inclusive genome-wide discoveries

Wei Q. Deng[1,2*] and Lei Sun[3,4]

[1]Department of Psychiatry and Behavioural Neurosciences, McMaster University, Address: 100 West 5th Street, Hamilton, L8P 3R2, Canada.

[2]Peter Boris Centre for Addictions Research, St. Joseph's Healthcare Hamilton.

[3]Department of Statistical Sciences, University of Toronto, Address: Ontario Power Building, 700 University Avenue, 9th Floor, Toronto, M5G 1Z5, Canada,

[4]Dalla Lana School of Public Health, University of Toronto, Address:155 College St Room 500, Toronto, M5T 3M7, Canada.

*Correspondence dengwq@mcmaster.ca

Abstract

A joint analysis of location and scale can be a powerful tool in genome-wide association studies to uncover previously overlooked markers that influence a quantitative trait through both mean and variance, as well as to prioritize candidates for gene-environment interactions. This approach has recently been generalized to handle related samples, dosage data, and the analytically challenging X-chromosome. We disseminate the latest advances in methodology through a user-friendly software package. The implemented R package (https://cran.r-project.org/web/packages/gJLS2) can be called via PLINK to enable a streamlined population-based analysis genome-wide, or called directly as a front end R script for intermediate gene-set based analyses, or used in any R GUI for smaller analyses targeting specific variants of interest.

1 Introduction

Genetic association studies examine the relationship between the genotypes of a single-nucleotide polymorphism (SNP; denoted by *G*) and a quantitative phenotype (denoted by *Y*), by testing the mean differences in *Y* according to the genotypes, otherwise known as a location test. More recently, several reports have investigated association with phenotypic variance of complex quantitative traits (Pare, Cook, Ridker, & Chasman, 2010; Shungin et al., 2017; Yang et al., 2012), by testing the variance differences in *Y* across the genotype groups of a SNP, or known as a scale test, in hope to find biologically meaningful markers. One possible explanation for a significant scale test is the presence of gene–gene (GxG) or gene–environment (GxE) interactions; both referred to as GxE hereinafter. Unlike a direct test of GxE interaction, a scale test can be used to indirectly infer GxE without knowledge of the interacting covariates, thus alleviates the multiple hypothesis burden of testing all possible pairwise interactions, and the assumption that all interacting environmental variables could be (accurately) measured.

A more powerful approach to prioritize biologically relevant candidates is to jointly evaluate location and scale effects (Aschard, Hancock, London, & Kraft, 2010; Cao et al., 2014; Soave et al., 2015). As compared to other existing single-marker based joint tests, the joint location-scale (JLS) association testing framework (Soave et al., 2015) is easy to implement on individual-level data or in a meta-analysis, requiring only the location and scale *p*-values for each SNP. Since methods for genome-wide location association are well-established, the main focus was on improving scale tests. Soave et al. (2017) generalized the well-known Levene's test to handle complex data structures often observed in genetic studies, such as correlated samples and dosage data. Following the X-inclusive trend to genome-wide analyses, robust and powerful location (Chen et al, 2021) and scale tests (Deng et al., 2019) tailored for X-chromosome (Xchr) are now available.

This work joins our previous efforts on the JLS and generalized JLS (gJLS) methodology for autosomes, and includes the recently available mean and variance association methods for Xchr, under a unifying software package that enables an X-inclusive, genome-wide, gJLS2 association analysis for common variants.

2 Main Functions

This package supports: 1) location association, 2) scale association, and 3) the joint location and scale analysis, and requires minimal inputs of a quantitative trait and the genotype data, which can be discrete genotype counts, continuous dosage genotype values or imputed genotype probabilities. Additional inputs, such as covariates, are possible and will be adjusted for in both the location and scale association models.

2.1 Location association

For autosomal SNPs, a linear regression fitted using the ordinary least square (OLS) is flexible to accommodate additional covariates and the default option. For related samples, the generalized least square (GLS) method can be used to account for correlated errors in smaller samples. Users can either provide the correlation matrix or specify the 'clust' argument to compute the correlation matrix according to pre-defined subgroups and correlation structure. However, for large population studies (n > 5,000), it is recommended to run the location association analysis using the state-of-art linear mixed models, such as LMM-BOLT (Loh *et al.*, 2015) or SAIGE (Zhou *et al.*, 2018).

The novel contribution of gJLS2 is the addition of our recommended X-chromosome location association strategy (Chen *et al.*, 2021), which has good type I error control, is robust to X-Chromosome inactivation (XCI) uncertainty and skewed XCI. The X-chromosome association requires additionally the sex information and returns results for both the recommended, sex-stratified and other alternative models. Notice that the recommended 3 degrees of freedom (df) test can only be applied to discrete genotypes:

$$\mathrm{E(Y)} = \beta_0 + \beta_\mathrm{S}\,\mathrm{Sex} + \beta_\mathrm{A}\,\mathrm{G_{Additive}} + \beta_\mathrm{D}\,\mathrm{G_{Dominance}} + \beta_\mathrm{GS} G_\mathrm{Additive} \times \mathrm{Sex}, \qquad (1)$$

where *Sex* is the genetic sex, $G_{Additive}$ is the additively coded genotype, and $G_{Dominance}$ is an indicator variable for the heterozygote genotype. Although the resulting 3-df test is robust to the choice of baseline allele and status of XCI, but recent report (Pazokitoroudi *et al.*, 2021) has found "limited contribution of dominance hereditability to complex trait variation". Thus, an alternative 2-df test without the dominance term is recommended and the default option for imputed SNPs.

2.2 Scale association

The scale component builds on two recent work that generalized to dosage genotype, related samples (Soave et al., 2017), and the X-chromosome (Deng et al., 2019).

The imputed data are analyzed either by computing the dosage value and used in place of *G* additively (1 df), or by simply replacing the genotype indicators for each observation, with the corresponding group probabilities (2 df). Similar to the location association, the sample relatedness is dealt with using GLS for autosomal markers at the second stage of linear regression via the correlation matrix. For X-chromosome markers, the recommended scale test uses a 2df test:

$$\mathrm{E(|d|)} = \beta_0 + \beta_\mathrm{S}\,\mathrm{Sex} + \beta_\mathrm{A}\,\mathrm{G_{Additive}} + \beta_\mathrm{GS} G_{Additive} \times \mathrm{Sex}, \qquad (2)$$

where *d* is the residual from the least absolute deviation (LAD) algorithm. A dominance term can be included for a 3df test, however, there is no prior evidence that the skewed XCI would influence the variance or to a degree that is non-negligible (Deng et al., 2019). Additional models, such as a sex-stratified variance test, can also be specified.

2.3 gJLS analysis

The gJLS analyses can be performed automatically as part of the gJLS2 pipeline after running location and scale tests. Alternatively, summary statistics for location and scale test are accepted and then combined via Fisher's method,, giving the corresponding test statistic for the gJLS:

$$W = -2[log(gL) + log(gS)] \sim \chi^2(4). \qquad (3)$$

An important assumption underlying this simple method to combine evidence is the normality of the quantitative trait, which leads to gL and gS being independent under the null of no association. Though empirically the analyses remain valid for approximated normal distributions (Soave et al., 2017), we strongly recommend the user to follow the default option and quantile transform the quantitative trait for scale association, but leaving the choice to the user for the location association, see McCaw et al. (2019) for a thorough investigation on performance in location association. We expect the independence between gL and gS to hold for Xchr SNPs under appropriate location and scale models that accounted for the confounding effect of sex, XCI uncertainty and skewed XCI (Chen *et al.*, 2021). We provide a simulation study to help users gauge the effect of non-normality on the gJLS *p*-values for Xchr markers (Supplementary Materials).

3 Usage scenarios

3.1 Data preparation

The genotype data can be supplied in PLINK format via the R plug-in option using PLINK 1.9 (Chang et al., 2015) or any other format that can be read in R with packages "BGData" and "BEDMatrix" (Grueneberg and Campos, 2019). The phenotype and covariate should ideally be supplied in the same file, and if not, should have a common column to match samples. For smaller analyses or testing purpose, it is possible to use the package in R GUI directly. However, for genome-wide analyses on larger datasets, we recommend the use of a high-performing computing cluster and one of the two

options below. Further details on how to set up each option is included as part of the gJSL2 package, see (https://cloud.r-project.org/web/packages/gJLS2/vignettes/Introduction.html).

3.2 PLINK R-plugin

This option is best suited for directly genotyped data and when the data size is large, for example, when working with UK Biobank sized data. The input can be any PLINK format genotype file that work with PLINK 1.9. The memory requirement for this option is higher and depends heavily on the data size, as the entire genotype matrix is called in the process, thus we recommend running by chromosome.

3.3 Rscript

This option is more flexible but less efficient as the script runs in R. This can be partially resolved by splitting the genome into smaller chunks to process each time, the chunk size can be specified via the '-write' argument. We recommend this option for analyses that require additional arguments, such as correlated samples and dosage data. As an example, we used this option to perform X-chromosome gJLS analyses on UK Biobank data (Bycroft et al., 2018) on four complex traits previously studied (Deng et al, 2019; Supplementary Figures 5-8). It took ~20 hours on a single core with 20G allocated memory (computing node specs 2xIntel E5-2683 v4 Broadwell @ 2.1GHz), to complete the analysis for 276,694 unrelated European samples and 15,179 X-chromosome variants per trait.

4 Conclusions

We hope the availability of gJLS2 software package will encourage more scale/joint analyses in large-scale datasets, and promote the standardized reporting of their *p*-values to be shared with the scientific community.

Acknowledgements

The authors would like to acknowledge Jiafen Gong, Sangook Kim, and Boxi Lin for testing and providing feedbacks on the initial versions of the gJLS2 software. The authors thank David Soave, the original creator of JLS and gJLS, for a critical reading of the paper.

Funding

This research is funded by the Natural Sciences and Engineering Research Council of Canada (NSERC) and the University of Toronto McLaughlin Centre.

Conflict of Interest: none declared.

References

Bycroft, C., Freeman, C., Petkova, D. et al. (2018). The UK Biobank resource with deep phenotyping and genomic data. *Nature*. **562**, 203–209 https://doi.org/10.1038/s41586-018-0579-z

Christopher C Chang, Carson C Chow, Laurent CAM Tellier, Shashaank Vattikuti, Shaun M Purcell, James J Lee, (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets, *GigaScience*, Volume 4, Issue 1, s13742–015–0047–8, https://doi.org/10.1186/s13742-015-0047-8.

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*, **10**.1002/gepi.22422. Advance online publication. https://doi.org/10.1002/gepi.22422

Deng, W. Q., Mao, S., Kalnapenkis, A., Esko, T., Mägi, R., Paré, G., & Sun, L. (2019). Analytical strategies to include the X-chromosome in variance heterogeneity analyses: Evidence for trait-specific polygenic variance structure. *Genetic epidemiology*, ***43***(7), 815–830. https://doi.org/10.1002/gepi.22247.

Grueneberg A, de los Campos G (2019). "BGData - A Suite of R Packages for Genomic Analysis with Big Data." *G3: Genes, Genomes, Genetics*, 9(5), 1377–1383. https://doi.org/10.1534/g3.119.400018.

McCaw, ZR, Lane, JM, Saxena, R, Redline, S, Lin, X. (2020). Operating characteristics of the rank-based inverse normal transformation for quantitative trait analysis in genome-wide association studies. *Biometrics*. **76**: 1262– 1272. https://doi.org/10.1111/biom.13214

Loh, PR., Tucker, G., Bulik-Sullivan, B. *et al*. (2015). Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nature Genetics*. **47,** 284–290. https://doi.org/10.1038/ng.3190

Pazokitoroudi A, Chiu AM, Burch KS, Pasaniuc B, Sankararaman S. (2021) Quantifying the contribution of dominance deviation effects to complex trait variation in biobank-scale data. *American Journal of Human Genetics*. 108(5):799-808. https://doi.org/10.1016/j.ajhg.2021.03.018.

Soave, D., Corvol, H., Panjwani, N., Gong, J., Li, W., Boëlle, P. Y., Durie, P. R., Paterson, A. D., Rommens, J. M., Strug, L. J., & Sun, L. (2015). A Joint Location-Scale Test Improves Power to Detect Associated SNPs, Gene Sets, and Pathways. *American journal of human genetics*. ***97***(1), 125–138. https://doi.org/10.1016/j.ajhg.2015.05.015.

Soave D, Sun L. (2017). A generalized Levene's scale test for variance heterogeneity in the presence of sample correlation and group uncertainty. *Biometrics*. **73**(3):960-971. https://doi.org/10.1111/biom.12651.

Zhou, W., Nielsen, J.B., Fritsche, L.G. *et al*. (2018). Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nature Genetics* **50,** 1335–1341. https://doi.org/10.1038/s41588-018-0184-y.