Bioinformatics, 32(2), 2016, 295–297
doi: 10.1093/bioinformatics/btv567
Advance Access Publication Date: 1 October 2015
Applications Note

OXFORD

Genome analysis

JEPEGMIX: gene-level joint analysis of functional SNPs in cosmopolitan cohorts

Donghyung Lee*, Vernell S. Williamson, T. Bernard Bigdeli, Brien P. Riley, Bradley T. Webb, Ayman H. Fanous, Kenneth S. Kendler, Vladimir I. Vladimirov and Silviu-Alin Bacanu

Department of Psychiatry, Virginia Commonwealth University, Richmond, VA 23298, USA

*To whom correspondence should be addressed.
Associate Editor: John Hancock

Received on July 13, 2015; revised on September 1, 2015; accepted on September 22, 2015

Abstract

Motivation: To increase detection power, gene level analysis methods are used to aggregate weak signals. To greatly increase computational efficiency, most methods use as input summary statistics from genome-wide association studies (GWAS). Subsequently, gene statistics are constructed using linkage disequilibrium (LD) patterns from a relevant reference panel. However, all methods, including our own Joint Effect on Phenotype of eQTL/functional single nucleotide polymorphisms (SNPs) associated with a Gene (JEPEG), assume homogeneous panels, e.g. European. However, this renders these tools unsuitable for the analysis of large cosmopolitan cohorts.

Results: We propose a JEPEG extension, JEPEGMIX, which similar to one of our software tools, Direct Imputation of summary STatistics of unmeasured SNPs from MIXed ethnicity cohorts, is capable of estimating accurate LD patterns for cosmopolitan cohorts. JEPEGMIX uses this accurate LD estimates to (i) impute the summary statistics at unmeasured functional variants and (ii) test for the joint effect of all measured and imputed functional variants which are associated with a gene. We illustrate the performance of our tool by analyzing the GWAS meta-analysis summary statistics from the multi-ethnic Psychiatric Genomics Consortium Schizophrenia stage 2 cohort. *This practical application supports the immune system being one of the main drivers of the process leading to schizophrenia.*

Availability and implementation: Software, annotation database and examples are available at http://dleelab.github.io/jepegmix/.

Contact: donghyung.lee@vcuhealth.org

Supplementary information: Supplementary material is available at *Bioinformatics* online.

1 Introduction

Univariate analysis of genome-wide association studies (GWAS) has emerged as the main tool for identifying trait/disease-associated genetic variants (Burton *et al.*, 2007). However, most variants reported by complex trait GWAS are common single nucleotide polymorphisms (SNPs) with weak or moderate effect sizes, which account for only a small fraction of the overall phenotypic variation (Manolio *et al.*, 2009). This is due to the fact that, due to their small effect sizes, most common causal variants are unlikely to be detected in GWAS (Yang *et al.*, 2010).

A reasonable approach to increase the power to detect true association signals with small effect sizes is to aggregate them by jointly analyzing multiple SNPs. To leverage information from multiple SNPs, multivariate association tests (Ehret *et al.*, 2012; Wood *et al.*, 2011; Yang *et al.*, 2012) have been also proposed. However, these methods typically test all SNPs, regardless of their functionality.

Given that functional SNPs are likely to jointly impact on gene expression, to increase detection power, our group proposed JEPEG (Joint Effect on Phenotype of eQTL/functional SNPs associated with

a Gene; Lee *et al.*, 2015b), which (i) uses only summary association statistics, (ii) imputes summary statistics of unmeasured functional SNPs and (iii) boosts detection power by jointly analyzing measured and imputed functional variants. However, similar to direct imputation methods based on summary statistics, e.g. DIST (Lee *et al.*, 2013) and ImpG (Pasaniuc *et al.*, 2014), it is only applicable to homogeneous cohorts. To overcome this limitation, concurrently with Adapt-Mix (Park *et al.*, 2015) and DISSCO (Xu *et al.*, 2015), our group developed DISTMIX (Direct Imputation of summary STatistics of unmeasured SNPs from MIXed ethnicity; Lee *et al.*, 2015a). It extends DIST capabilities to the analysis of mixed ethnicity cohorts by estimating their linkage disequilibrium (LD) patterns as a mixture of the LD patterns from the constituent ethnicities of large reference panels, e.g. 1000 Genomes data (1KG) (Altshuler *et al.*, 2010). Here, for the gene level analysis of the ever more common (and well powered) mixed ethnicity cohorts, we propose JEPEG for MIXed ethnicity cohorts (JEPEGMIX), which adapts the LD estimation strategy used by DISTMIX, while retaining all JEPEG advantages.

2 Methods

Similar to DISTMIX, to accurately estimate LD patterns for mixed ethnicity cohorts, JEPEGMIX first estimates the ethnic proportions of study cohorts using study allele frequency (AF) information [see Supplementary Text S1 in supplementary data (SD) for details]. (Alternatively, when AF information is not available, user can prespecify the proportions based on the ethnic composition information typically provided by published studies.) Next, using the estimated/user-specified ethnic proportions, the software estimates LD patterns of the study cohort as a weighted mixture of the LD matrices of all ethnic groups in a reference panel (Supplementary Text S2 of SD). Finally, it uses these estimated mixture LD patterns and association summary statistics to (i), when necessary, rapidly and accurately impute summary statistics of unmeasured functional SNPs (Supplementary Text S3 of SD) and (ii) jointly test the effect of measured and imputed functional SNPs associated with each gene (Supplementary Text S4 of SD).

3 Results

To estimate false positive rates, null hypothesis cosmopolitan cohorts were simulated using haplotypic patterns from 1KG (see Supplementary Text S5 of SD). When compared with JEPEG, JEPEGMIX maintains the false positive rates at or below nominal thresholds (Supplementary Fig. S1 in SD). We obtained gene-level statistics by applying the method to association summary statistics from the large-scale cosmopolitan Psychiatric Genomics Consortium Schizophrenia stage 2 (PGC SCZ2) cohort (Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014). A subsequent Ingenuity Pathway Analysis (www.ingenuity.com) of the 61 significant JEPEGMIX genes (Supplementary Table S1 in SD), i.e. those with false discovery rate q-values < 0.05, yields a large number of immune pathways and only one (in italics) which is neurologically related (Table 1). The pattern is maintained even when excluding the 21 genes located in the immune related Major Histocompatibility (MHC) region from chromosome 6p (Supplementary Table S2).

Table 1. Pathways significant at a Type I error of 0.05

Pathway	*P*-value
Antigen presentation pathway	0.0002
Graft-versus-host disease signaling	0.0004
Autoimmune thyroid disease signaling	0.0005
Granzyme A signaling	0.002
Dendritic cell maturation	0.002
Allograft rejection signaling	0.002
OX40 signaling pathway	0.003
Crosstalk between dendritic cells and natural killer cells	0.003
Communication between innate and adaptive immune cells	0.003
Cytotoxic T lymphocyte-mediated apoptosis	0.004
Type I diabetes mellitus signaling	0.005
Role of RIG1-like receptors in antiviral innate immunity	0.008
Neuroprotective Role of THOP1 in Alzheimer's Disease	*0.009*
Nur77 signaling in T lymphocytes	0.01
Cdc42 signaling	0.01
Calcium-induced T Lymphocyte apoptosis	0.02
Caveolar-mediated endocytosis signaling	0.02
CTLA4 signaling in cytotoxic T lymphocytes	0.03
Virus entry via endocytic pathways	0.03
p53 Signaling	0.04
G-protein coupled receptor signaling	0.05

4 Conclusions

For multi-ethnic cohorts, unlike existing methods, JEPEGMIX controls the Type I error rates at or below nominal levels. Due to ridge adjustment being inversely related to the size of 1KG relevant subpopulations (Supplementary Text S2 of SD), at present the method is rather conservative. However, the conservativeness is expected to become negligible with the advent of extremely large reference panels (http://www.haplotype-reference-consortium.org). Thus, to the capabilities of JEPEG, JEPEGMIX adds the much needed applicability to the analysis of large cosmopolitan cohorts, which are the state-of-the-art in detecting genetic signals. For such cohorts, it (i) imputes unmeasured functional SNPs, (ii) pools in a synthetic variable the information of measured and imputed SNPs from the same functional category and (iii) combines these synthetic variables in a gene-level Mahalanobis test. JEPEGMIX application to PGC SCZ2 cohort suggests that, in the etiology of SCZ, the immune system might play a more substantial role than currently accepted.

Funding

This work was supported by National Institute on Drug Abuse [R25DA026119 to D.L.], National Institutes of Mental Health [R21MH100560 to S.A.B. and B.P.R.] and National Institute on Alcohol Abuse and Alcoholism [R21AA022717 to S.A.B. and V.I.V.; P50AA022537 to S.A.B. and K.S.K.].

Conflict of Interest: none declared.

References

Altshuler,D. *et al.* (2010) A map of human genome variation from population-scale sequencing. *Nature*, **467**, 1061–1073.

Burton,P.R. *et al.* (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, **447**, 661–678.

Ehret,G.B. *et al.* (2012) A multi-SNP locus-association method reveals a substantial fraction of the missing heritability. *Am. J. Hum. Genet.*, **91**, 863–871.

Lee,D. *et al.* (2013) DIST: direct imputation of summary statistics for unmeasured SNPs. *Bioinformatics*, **29**, 2925–2927.

Lee,D. *et al.* (2015a) DISTMIX: direct imputation of summary statistics for unmeasured SNPs from mixed ethnicity cohorts. *Bioinformatics*, **31**, 3099–3104.

Lee,D. *et al.* (2015b) JEPEG: a summary statistics based tool for gene-level joint testing of functional variants. *Bioinformatics*, **31**, 1176–1182.

Manolio,T.A. *et al*. (2009) Finding the missing heritability of complex diseases. *Nature*, **461**, 747–753.

Park,D.S. *et al*. (2015) Adapt-Mix: learning local genetic correlation structure improves summary statistics-based analyses. *Bioinformatics*, **31**, i181–i189.

Pasaniuc,B. *et al*. (2014) Fast and accurate imputation of summary statistics enhances evidence of functional enrichment. *Bioinformatics*, **30**, 2906–2914.

Schizophrenia Working Group of the Psychiatric Genomics Consortium. (2014) Biological insights from 108 schizophrenia-associated genetic loci. *Nature*, **511**, 421–427.

Wood,A.R. *et al*. (2011) Allelic heterogeneity and more detailed analyses of known loci explain additional phenotypic variation and reveal complex patterns of association. *Hum. Mol. Genet.*, **20**, 4082–4092.

Xu,Z. *et al*. (2015) DISSCO: direct imputation of summary statistics allowing covariates. *Bioinformatics*, **31**, 2434–2442.

Yang,J. *et al*. (2010) Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.*, **42**, 565–569.

Yang,J. *et al*. (2012) Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat. Genet.*, **44**, 369–375.