Accounting for population structure and relatedness in gene expression genome-wide association testing using a mixed-model approach

Greg Gibson
Department of Genetics, North Carolina State University, Raleigh NC, USA/ School of Biological Sciences, University of Queensland, Queensland, Australia

Youssef Idaghdour
Department of Genetics, North Carolina State University, Raleigh NC, USA

Wendy Czika
SAS Institute Inc., Cary NC, USA

Kelci Miclaus
SAS Institute Inc., Cary NC, USA

Sang H. Lee
Queensland Institute of Medical Research, Brisbane, Queensland, Australia

Peter M. Visscher
Queensland Institute of Medical Research, Brisbane, Queensland, Australia

Russell D. Wolfinger
SAS Institute Inc., Cary NC, USA

Method Article

Keywords: gene expression, GWAS, eSNP, population structure, ethnicity, relatedness, mixed-model, SAS, JMP Genomics

Posted Date: December 15th, 2009

DOI: https://doi.org/10.1038/nprot.2009.216

Abstract

Introduction

Studies of the genetics of gene expression reveal expression SNPs \(eSNP) that explain variation in transcript abundance. Here we report a step-by-step protocol of a whole genome gene expression association mapping procedure using JMP Genomics \(SAS Institute, Cary, NC) as described in ref. 1. The method accounts for population structure, ethnicity, relatedness and other variables and was applied on a sample of 194 individuals from a city and two villages in southern Morocco representing two ethnicities and a wide range of family relationships. Gene expression genome-wide association testing between 516,792 SNPs from Illumina Human 610-Quad beadchips with expression measurements at 22,300 probes from Illumina HumanHT12 beadchips was performed. Filtering via Pearson correlation tests between expression and SNP genotypes was used prior to testing with more complex models that contained several fixed effects and relatedness between individuals as a random effect. These models were run using processes implemented in JMP Genomics 4.1 released by SAS in December 2009.

Equipment

1- A data set of gene expression measurements for a group of individuals 2- A data set of genotypes \(SNPs) from the same group of individuals 3- Annotation files for genes and SNPs 4- Experimental design files containing various relevant data on the individuals 5- Intel® Core™2 Duo CPU, 2.33 GHz, 3.23 GB Ram with Windows XP 6- JMP Genomics - "http://www.jmp.com/software/genomics/":http://www.jmp.com/software/genomics/

Procedure

The procedure of gene expression genome-wide association testing is performed once normalized \(or standardized) gene expression and genotypic data are obtained. Detailed description of quality control measures to obtain gene expression and genotypic data is beyond the scope of this protocol and can be found in ref. 1. 1- Input files for the procedure 1-1- Gene expression data - Obtain raw gene expression data - Do a log2 transformation followed by appropriate data standardization or normalization using JMP Genomics that provides several methods for normalizing gene expression data sets3. - Retain the transcripts with expression above background levels. This procedure is described in ref. 2. 1-2- Genotypic data - Extract genotypic data after stringent quality control. In ref. 1, a total of 579,144 SNPs were generated using BeadStudio \(Illumina). - Generate molecular properties \(minor allele frequency, heterozygosity….etc) for each marker using JMP Genomics. - Numerically code the entire allelic data set using JMP Genomics as 0, 1, or 2 where each number represents the number of copies of the minor allele. - Retain markers with minor allele frequencies MAF > 5% for association testing. In ref. 1, a total of 516,792 have a MAF > 5%. Save both gene expression and genotypic data sets in JMP Genomics wide format with the samples as rows and molecular entities \(numerical genotypes and gene expression measurements) as columns. Each data set must contain a column of individual identifiers which is used

for merging the data sets together. Save the data also in JMP Genomics tall format along with an accompanying experimental design data set for certain analyses as described in JMP Genomics user guide3 where details about the different input file formats are described. Create also a data set comprising all gene expression and SNP genotype data as described below. 2- Variables included in the association model 2-1- Matrix of relatedness measures - Calculate relatedness Âij between each pair of individuals using the method described in ref. 4 for all individual pairs where Âij is averaged over l = 1 to n loci: Âij = \[? \(x il – 2 p). \(\(x jl –2 p) / 2 pq)] / n where x il = 0, 1 or 2 according to whether individual i has genotype aa, Aa or AA at locus l, p \(q) is allele frequency of A \(a), and 2p is the mean of x l. Use only autosomal SNPs that satisfy quality control criteria for Hardy-Weinberg equilibrium \(P value > 0.0001) and minor allele frequency \(MAF > 0.05) to generate the matrix of relatedness measures. 2-2- Genotypic principal component eigenvectors and clustered ethnicity - Perform principal component analyses on the entire genotypic dataset using the Eigenstrat method5 as implemented in JMP Genomics. - Retain scores from the significant genotypic PCs for inclusion in the association tests. - Inclusion of clustered ethnicity based on PC plots is optional. For example, the 194 samples studied in ref. 1 were clustered into four clusters to account for location in an unbiased manner relative to ethnicity. Deciding to include clustered ethnicity or not as a variable is best done on a case-by-case basis depending on the observed population structure patterns. 3- Association testing - First use the JMP Genomics Cross Correlation process to test for Pearson correlation between all pairs of expression measurements and numerically coded SNP genotypes \(>11 billion pairs in ref. 1) using the following basic correlation model, where Baseline is the mean measure of transcript abundance, and the error ? is assumed to be normally distributed with a mean of zero Expression = Baseline + SNP + ? \(Model 1) - To significantly reduce computational time and the need for disk space for data storage save only hits below 0.05. This cut-off yielded 690 million significantly correlated SNP/gene pairs using the dataset in ref.1. - If storage and computing power is not an issue all P values should be saved. - Bring the 10,000 most significant SNP-expression associations from this model \(~P < 10-7) forward for two further analyses in a mixed model framework. - Apply more complex models to each of the 10,000 most significant SNP-expression pairs using the JMP Genomics Q-K Mixed Model process. For example the following model is used in ref. 1 to account for various effects including relatedness and population structure and various interaction effects: Expression = Baseline + Location + Gender + Relatedness + Significant Genotypic Principal Components + Clustered Ethnicity + SNP + SNP** Location + Gender** Clustered Ethnicity + Gender*Location + ? \(Model2) - The JMP Genomics Q-K Mixed Model process fits a model as described in ref 6. to test for association between the gene expression variants and SNP genotypes while adjusting for fixed effects and the random effect relatedness. JMP Genomics calls the MIXED procedure in SAS/STAT to perform these analyses. - First create a data set comprising all expression and SNP genotype data as input for the Q-K Mixed Model process. Create a second data set containing the columns of the relatedness matrix named Col1, Col2,…, Coln as well as a Row column indicating the row of the matrix . - In ref.1, because of the large number of SNP-expression pairs to be run, a JSL script was used to auto-generate the required JMP Genomics setting files. Each settings file contained the parameters needed for a single run of Q-K Mixed Model for one model for a single SNP-expression pair. All settings were run as a group in the JMP Genomics Workflow Builder and results were gathered into a

single data set. Model 2 adjusted for relatedness, by specifying the column of the relatedness data set containing the relatedness measurements as a Random Effect, and using the PROC MIXED options LDATA= and TYPE=LIN\(1) for the Random Statement Options.

Timing

The analyses described in the Association Testing section below took approximately 47 hours using the computer described above.

Troubleshooting

JMP Genomics and SAS provide comprehensive troubleshooting for the analyses involved in this procedure.

Anticipated Results

This procedure yields various statistics for each pair SNP-gene expression measurement association test.

References

1. Idaghdour, Y. et al. \(2009) Geographical Genomics of Human Leukocyte Gene Expression Variation in Southern Morocco. Nat Genet doi: 10.1038/ng.495 2. Idaghdour, Y., Storey, J.D., Jadallah, S.J., & Gibson, G. A Genome-Wide Gene Expression Signature of Environmental Geography in Leukocytes of Moroccan Amazighs. PLoS Genet 4\(4): e1000052. doi:10.1371/journal.pgen.1000052 \(2008). 3. SAS Institute Inc. JMP Genomics User Guide. Cary, NC: SAS Press. \(2008) 4. Hayes, B.J., Visscher, P.M. & Goddard, M.E. Increased accuracy of artificial selection by using the realized relationship matrix. Genet. Res. 91, 47-60 \(2009). 5. Price, A.L., Patterson, N.J., Plenge, R.M., et al. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38: 904-909 \(2006). 6. Yu, J., Pressoir, G., Briggs, W.H., et al. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness, Nature Genetics 38, 203-208 \(2006).

Acknowledgements

YI is supported by a Fulbright scholarship and GG by an ARC Australian Professorial Fellowship.