Simultaneous Selection of Multiple Important Single Nucleotide Polymorphisms in Familial Genome Wide Association Studies Data

Subhabrata Majumdar [1] **Saonli Basu** [1] **Matt McGue** [1] **Snigdhansu Chatterjee** [1]

1. Introduction

Genome Wide Association Studies (GWAS) on families instead of unrelated individuals are used to counter environmental heterogeneity issues while detecting small effects of individual SNPs on quantitative behavioral traits such as alcohol consumption, drug abuse, anorexia and depression. However, association analysis using familial data can be computationally challenging. In this context, single-SNP association analysis is the standard tool, where p-values obtained from association tests between the quantitative trait and single SNPs that are lower than a particular threshold are considered to be associated with the trait. Such methods may either ignore the familial dependency structure (Aulchenko et al., 2007; Chen & Abecasis, 2007), or use a mixed effects model to account for it (Li et al., 2011).

Single-SNP methods are prone to be less effective in detecting SNPs with weak signals (Manolio et al., 2009), including situations where multiple SNPs are jointly associated with the phenotype (Yang et al., 2012; Ke, 2012; Schifano et al., 2012). While several alternative methods of multi-SNP analysis exist, such as Kernel based association tests (Schifano et al., 2012; Chen et al., 2013), all of them test for whether a *group* of SNPs is associated with the phenotype being analyzed. Consequently, they may not be detect the individual SNPs primarily associated with the trait outcome.

One way to solve this problem is to perform model selection. However, training multiple models corresponding to multiple predictor sets are computationally intensive to implement in a mixed-model framework necessary for familial data. In this work, we propose a fast and scalable model selection technique to alleviate this problem. Our method fits a single model on the family data being analyzed, and aims to identify important genetic variants with weak signals through joint modelling of multiple variants. We achieve this by extending our proposed framework of e-values (Majumdar & Chatterjee, 2017). A variable selection algorithm using e-values has the following simple and generic steps:

1. Fit the full model, i.e. where all predictor effects are being estimated from the data, and use resampling to estimate its e-value;
2. Set an element of the full model coefficient estimate to 0 and get an e-value for that predictor, quantifying its importance;
3. Select predictors that have e-values below a predetermined threshold as important.

Using a fast Monte-Carlo simulation-based bootstrap (Majumdar & Chatterjee, 2017) allows us to circumvent computational issues related to resampling. We refer the reader to our main paper[1] for details on the methodology, model formulation, and experiments.

2. Analysis of Twin Studies data

We apply our method on the Minnesota Twin Studies dataset (Miller et al., 2012). We assume a nuclear pedigree structure (i.e. parents and two children in each family, families are assumed unrelated). For simplicity we only analyze pedigrees with identical or non-identical twins. We look at the effect of genetic factors behind the response variable, pertaining to the amount of alcohol consumption, which is highly heritable in this specific dataset according to previous studies (McGue et al., 2013). We analyze SNPs inside some of the most-studied genes with respect to alcohol abuse through separate gene-level models, with sex, birth year, age and generation of individuals (parent/offspring) as covariates.

As seen in Table 1, we detect a much higher number of SNPs than the competing methods. A number of SNPs we detect (or SNPs situated close to them) have known associations with alcohol-related behavioral disorders. We summarize this in Table 2. Plots of SNP-specific e-values for the gene GABRA2 identifies two SNPs (longest 2 vertical lines in Figure 1) close to rs278958, the well-known SNP in GABRA2 associated with alcohol abuse (Cui et al., 2012), which was not genotyped in this study. Detailed results, including extensive synthetic experiments, can be found in the main paper.

References

Aulchenko, Y. S. et al. Genome-wide rapid association using mixed model and regression: a fast and simple method

[1] https://arxiv.org/abs/1802.01141

Gene	Total no. of SNPs	No. of SNPs detected by		
		e-value	RFGLS+BH	mBIC2
GABRA2	11	5	0	0
ADH	44	3	1	0
OPRM1	47	25	1	0
CYP2E1	9	5	0	0
ALDH2	6	5	0	1
COMT	15	14	0	0
SLC6A3	18	4	0	0
SLC6A4	5	0	0	0
DRD2	17	0	0	1

Table 1. Table of analyzed genes and number of detected SNPs in them by the three methods. Competing methods are (a) a single-SNP fast mixed effect method (RFGLS, (Li et al., 2011)) followed by Bonferroni correction, and (b) mBIC2 (Frommelet et al., 2012), a variant of the Bayesian Information Criterion (BIC) that selects from multiple SNPs but overlooks familial structure.

Gene	Detected SNPs with known associations	Reference for associated SNP
GABRA2	rs1808851, rs279856: close to rs279858	(Cui et al., 2012)
ADH genes	rs17027523: 20kb upstream of rs1229984	Multiple studies (https://www.snpedia.com/index.php/Rs1229984)
OPRM1	rs12662873: 1 kb upstream of rs1799971	Multiple studies (https://www.snpedia.com/index.php/Rs1799971)
CYP2E1	rs9419624: 600b downstream of rs4646976; rs9419702: 10kb upstream of rs4838767	(Lind et al., 2012)
ALDH2	rs16941437: 10kb upstream of rs671	Multiple studies (https://www.snpedia.com/index.php/Rs671)
COMT	rs4680, rs165774	(Voisey et al., 2011)
SLC6A3	rs464049	(Huang et al., 2017)

Table 2. Table of detected SNPs with known references

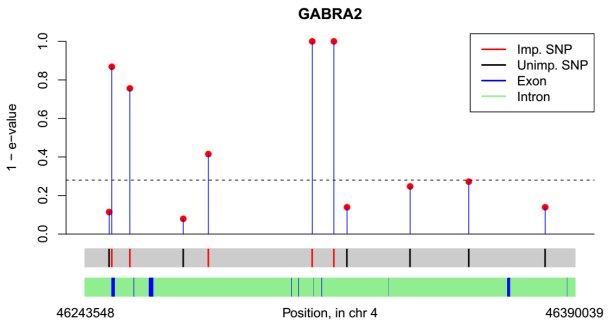

Figure 1. Plot of e-values for GABRA2. For ease of visualization, $1 - e$-values are plotted in the y-axis.

for genome-wide pedigree-based quantitative trait loci association analysis. *Nat. Genet.*, 177:577–585, 2007.

Chen, H., Meigs, J. B., and Dupuis, J. Sequence Kernel Association Test for Quantitative Traits in Family Samples. *Genet. Epidemiol.*, 37:196–204, 2013.

Chen, W. M. and Abecasis, G. Family-based association tests for genome-wide association scans. *Am. J. Hum. Genet.*, 81:913–926, 2007.

Cui, W. Y. et al. Genetics of GABAergic signaling in nicotine and alcohol dependence. *Hum. Genet.*, 131:843–855, 2012.

Frommelet, F. et al. Modified versions of Bayesian Information Criterion for genome-wide association studies. *Comput. Stat. Data Anal.*, 56:1038–1051, 2012.

Huang, C.-C., Kuo, S.-C., Weh, Y.-W., et al. The SLC6A3 gene possibly affects susceptibility to late-onset alcohol dependence but not specific personality traits in a Han Chinese population. *PLoS ONE*, 12(2):e0171170, 2017.

Ke, X. Presence of multiple independent effects in risk loci of common complex human diseases. *Am. J. Hum. Genet.*, 91:185–192, 2012.

Li, X. et al. A Rapid Generalized Least Squares Model for a Genome-Wide Quantitative Trait Association Analysis in Families. *Hum. Hered.*, 71:67–82, 2011.

Lind, P. A. et al. Association between *in vivo* alcohol metabolism and genetic variation in pathways that metabolize the carbon skeleton of ethanol and NADH reoxidation in the Alcohol Challenge Twin Study. *Alcohol Clin. Exp. Res.*, 36:2074–2085, 2012.

Majumdar, S. and Chatterjee, S. Fast and General Model Selection using Data Depth and Resampling. https://arxiv.org/abs/1706.02429, 2017.

Manolio, T. A., Collins, F. S., Cox, N. J., et al. Finding the missing heritability of complex diseases. *Nature*, 461: 747–753, 2009.

McGue, M. et al. A Genome-Wide Association Study of Behavioral Disinhibition. *Behav Genet.*, 43, 2013.

Miller, M. B., Basu, S., Cunningham, J., et al. The Minnesota Center for Twin and Family Research Genome-Wide Association Study. *Twin Res Hum Genet.*, 15:767–774, 2012.

Schifano, E. D. et al. SNP set association analysis for familial data. *Genet. Epidemiol.*, 36(8):797–810, 2012.

Voisey, J., Swagell, C. D., Hughes, I. P., et al. A novel SNP in COMT is associated with alcohol dependence but not opiate or nicotine dependence: a case control study. *Behav. Brain Funct.*, 7(51), 2011.

Yang, J., Ferreira, T., Morris, A. P., et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat. Genet.*, 44:369–375 S361–S363, 2012.