Selecting Important Single Nucleotide Polymorphisms in Twin Studies: the *e*-value approach

University of Minnesota
Driven to Discover℠

An Interdisciplinary Doctoral Fellowship collaboration

Subho Majumdar[1], Saonli Basu[2,3], Snigdhansu Chatterjee[1], Matt McGue[3,4]

[1]School of Statistics
[2]Division of Biostatistics
[3]Minnesota Ctr. for Twin and Family Research
[4]Department of Psychology

Twin Studies: What and Why

Behavior = Gene + Environment

- **Objective**: detect genes that influence behavioral disorders, e.g. alcohol dependence, drug abuse, anorexia;
- Twin Studies gather data from families with twin children instead of independent individuals;
- Shared environment in families reduce sample size required to detect genetic signals;
- Genetic effect determine by associating behavioral trait with Single Nucleotide Polymorphisms (SNP).

- **Challenges:**
  - ➢ Huge number of SNPs: ~500k
  - ➢ Non-independent data structure makes it hard to model all SNPs simultaneously.

$$\begin{array}{l}\text{Father}\\ \text{Mother}\\ \text{Twin 1}\\ \text{Twin 2}\end{array}\begin{pmatrix}1 & 0 & 0.5 & 0.5\\ 0 & 1 & 0.5 & 0.5\\ 0.5 & 0.5 & 1 & 0.5\\ 0.5 & 0.5 & 0.5 & 1\end{pmatrix} = \mathbf{K}, \text{ the kinship matrix}$$

- **State-of-the-art:**
  - ➢ Ignore dependent data – loses information, needs large samples
  - ➢ Model single-SNPs and choose from ordered *p*-values – ignores dependence among SNPs

Move over *p*-values!

Say we want to detect which of 100 SNPs in a gene are significant.

***p*-values:**

1. Start with model with no SNPs (null model);
2. Add an SNP, get *p*-value with respect to null model;
3. Repeat for all SNPs. Select SNPs with low *p*-values.

Bad:

- **Ignores dependence of SNPs;**
- **Cannot detect weak signals: often the case in SNP studies**

Our solution: *e*-values.

1. Start from model with *all* SNPs: **takes care of correlation of SNPs;**
2. Fix a SNP effect to 0 in the model, get *e*-value with respect to full model by comparing two *model distributions*: **this helps detect weak SNP effects.**

Statistical model

$$\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon} \quad \text{(Linear Mixed Model)}$$

- $\boldsymbol{Y}$ = quantitative trait values for members in a family, $\mathbf{X}$ = matrix of SNP values inside a gene, $\mathbf{Z}$ = random effect design matrix, $\boldsymbol{\gamma} \sim N(\mathbf{0}, \sigma_a^2\mathbf{K})$ is the vector of random effects, and $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma_e^2\mathbf{I})$ the random error term. Dependency inside a family is captured through $\boldsymbol{\gamma}$.
- To detect non-zero entries in $\boldsymbol{\beta}$, first get its maximum likelihood estimate, say $\widehat{\boldsymbol{\beta}}$. Use generalized bootstrap [1] with a large standard deviation to approximate its distribution, say $[\widehat{\boldsymbol{\beta}}]$.
- Replace j^{th} coordinate of $\widehat{\boldsymbol{\beta}}$ and the bootstrap samples with 0. Name them $\widehat{\boldsymbol{\beta}}_{0,j}$ and $[\widehat{\boldsymbol{\beta}}_{0,j}]$, respectively.
- Then *e*-value of SNP = tail probability for q^{th} percentile of $[E(\widehat{\boldsymbol{\beta}}_{0,j})]$ with respect to $[E(\widehat{\boldsymbol{\beta}})]$, where $E(.)$ is an *evaluation function* that takes higher value for a point closer to the center of $[\widehat{\boldsymbol{\beta}}]$, and smaller value for points away from it.
- Select SNPs with *e*-value < 0.5.

Results

Simulation

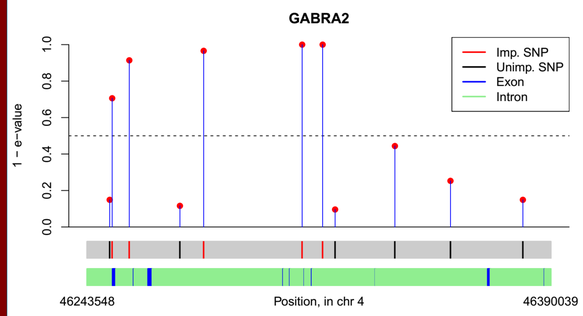
- 50 total SNPs: 4 of them have positive effect, each explaining h/6% pf total variance;
- 100 Families with twins, having kinship matrix **K**.
- **Methods compared:** (a) Linear regression model selection using Bayesian Information Criterion (BIC), (b) Get *p*-values and correct for multiple testing (PVAL).

TP = True Positive proportion, Size = model size:
both averaged from 100 replications

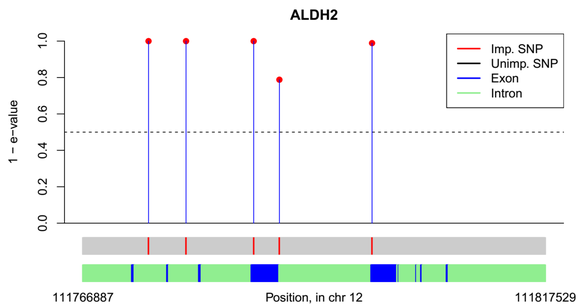
Method	*h* = 5		*h* = 1		*h* = 0	
	TP	**Size**	**TP**	**Size**	**TP**	**Size**
BIC	0.20	3.72	0.08	2.61	0.03	1.94
PVAL	0.03	0.30	0.03	0.30	0.003	0.18
***e*-value (*q* = 0.6)**	0.28	3.88	0.12	2.78	0.03	1.94

Genome-wide Twin Studies application

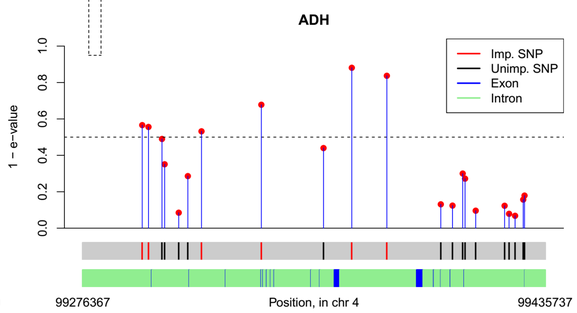
Data from the Minnesota Center for Twin and Family Research (MCTFR) Genome-Wide Association Study sample: 7188 individuals, 527,893 SNP markers [2]. Response variable is **amount of alcohol consumption**.
We consider 9 widely studied genes known to be associated with alcohol consumption and select important SNPs from them using *e*-values (q = 0.9).



(1) GABRA2
5 of 11 SNPs have non-zero effect: 4 very close to exons. The SNPs rs1808851, rs279856 are at perfect linkage disequilibrium with rs279858, a known associated SNP.



(2) ALDH2
All 5 tested SNPs have effect: rs7398343, rs7297186, rs3803167, rs10219736, rs3742004. Importantly all of them overlap with/ very close to coding regions.



(3) ADH1 to ADH7 genes
6 of 21 tested SNPs have *e*-values above threshold. Previously detected rs1229984 is in between two of them. First two are possibly novel: in the uncharacterized gene LOC100507053.

References:

[1] Chatterjee, S. and Bose, A. *Ann. Statist.* **2005**, 33, 414-436.
[2] McGue et al. *Behav. Genet.* **2013**, 43, doi:10.1007/s10519-013-9606-x.

I am grateful to Michael Miller of the Dept. of Psychology for his help regarding use of computational facilities at Minnesota Ctr. For Twin and Family Research (MCTFR), and Kevin Haroian of MCTFR for his support during the IDF application process.