Improving the Accuracy of Genome Wide Association Studies

February 17, 2017
Ben Lengerich
Genome-Wide Association Studies

Goal: Find the genetic polymorphisms associated with each phenotype
GWAS have been used to identify drug abuse risk factors

A genome-wide association study of alcohol dependence

Laura J. Bierut\textsuperscript{a,}\textsuperscript{1}, Arpana Agrawal\textsuperscript{a}, Kathleen K. Bucholz\textsuperscript{a}, Kimberly F. Doheny\textsuperscript{b}, Cathy Laurie\textsuperscript{c}, Elizabeth Pugh\textsuperscript{b}, Sherri Fisher\textsuperscript{a}, Louis Fox\textsuperscript{a}, William Howells\textsuperscript{a}, Sarah Bertelsen\textsuperscript{a}, Anthony L. Hinrichs\textsuperscript{a}, Laura Almasy\textsuperscript{d}, Naomi Breslau\textsuperscript{e}, Robert C. Culverhouse\textsuperscript{f}, Danielle M. Dick\textsuperscript{g}, Howard J. Edenberg\textsuperscript{h}, Tatiana Foroud\textsuperscript{i}, Richard A. Grucza\textsuperscript{a}, Dorothy Hatsukami\textsuperscript{j}, Victor Hesselbrock\textsuperscript{k}, Eric O. Johnson\textsuperscript{l}, John Kramer\textsuperscript{m}, Robert F. Krueger\textsuperscript{a}, Samuel Kuperman\textsuperscript{n}, Michael Lynskey\textsuperscript{a}, Karl Mann\textsuperscript{o}, Rosalind J. Neuman\textsuperscript{a}, Markus M. Nöthen\textsuperscript{p}, John I. Nurnberger, Jr.\textsuperscript{q}, Bernice Porjesz\textsuperscript{r}, Monika Ridinger\textsuperscript{s}, Nancy L. Saccone\textsuperscript{t}, Scott F. Saccone\textsuperscript{a}, Marc A. Schuckit\textsuperscript{u}, Jay A. Tischfield\textsuperscript{v}, Jen C. Wang\textsuperscript{a}, Marcella Rietschel\textsuperscript{w}, Alison M. Goate\textsuperscript{a}, and John P. Rice\textsuperscript{a}, as part of the Gene, Environment Association Studies (GENEVA) Consortium
GWAS have been used to identify drug abuse risk factors.

**KAT2B polymorphism identified for drug abuse in African Americans with regulatory links to drug abuse pathways in human prefrontal cortex**

Eric O. Johnson¹, Dana B. Hancock², Joshua L. Levy³, Nathan C. Gaddis³, Grier P. Page⁴, Cristie Glosheen², Nancy L. Saccone⁵, Laura J. Bierut⁶ & Alex H. Kral⁷

Fellow Program and Behavioral Health and Criminal Justice Division, RTI International, Research Triangle Park, NC, USA¹, Behavioral and Urban Health Program, Behavioral Health and Criminal Justice Division, RTI International, Research Triangle Park, NC, USA², Research Computing Division, RTI International, Research Triangle Park, NC, USA³, Fellow Program, Center for Genomics in Public Health and Medicine, and Genomics, Statistical Genetics, and Environmental Research Program, RTI International, Atlanta, GA, USA⁴, Department of Genetics, Washington University School of Medicine, St. Louis, MO, USA⁵, Department of Psychiatry, Washington University School of Medicine, St. Louis, MO, USA⁶, Behavioral and Urban Health Program, Behavioral Health and Criminal Justice Division, RTI International, San Francisco, CA, USA⁷
GWAS have been used to identify drug abuse risk factors

The interplay of genes and adolescent development in substance use disorders: leveraging findings from GWAS meta-analyses to test developmental hypotheses about nicotine consumption

Scott I. Vrieze · Matt McGue · William G. Iacono
GWAS have been used to identify drug abuse risk factors

Genome-Wide Association Studies of Alcohol Dependence and Substance Use Disorders

Jens Treutlein · Marcella Rietschel

The interplay of genes and adolescent development in substance use disorders: leveraging findings from GWAS meta-analyses to test developmental hypotheses about nicotine consumption

Scott I. Vrieze · Matt McGue · William G. Iacono
Traditional GWAS Use Independent Hypothesis Testing

- Chi-square Test
- Wald Test
- Fisher Exact Test

Associated SNP
Hypothesis testing is limited for GWAS

- A trait may not be determined by a single SNP
- A single SNP may be weakly associated with many phenotypes
- Rare SNPs are hard to detect
- Genomic data is structured
Hypothesis testing is limited for GWAS

- A trait may not be determined by a single SNP
- A single SNP may be weakly associated with many phenotypes
- Rare SNPs are hard to detect
- Genomic data is structured
Hypothesis testing is limited for GWAS

- A trait may not be determined by a single SNP
- A single SNP may be weakly associated with many phenotypes
- Rare SNPs are hard to detect
- Genomic data is structured
Hypothesis testing is limited for GWAS

- A trait may not be determined by a single SNP
- A single SNP may be weakly associated with many phenotypes
- Rare SNPs are hard to detect
- Genomic data is structured
Hypothesis testing is limited for GWAS

- A trait may not be determined by a single SNP
- A single SNP may be weakly associated with many phenotypes
- Rare SNPs are hard to detect
- Genomic data is structured
Many types of structure in genomic data

Epistasis

Population Structure

Linkage Disequilibrium

Phenotype Structure
Genome Structure

Phenome Structure

Structured Association Mapping

Transcriptome Structure
Many methods of structured association mapping

- **Phenotype Graph**
  - Graph-guided fused Lasso (Kim & Xing, PLoS Genetics, 2009)

- **Phenotype Tree**
  - Tree-guided Fused Lasso (Kim & Xing, ICML 2010)
Many methods of structured association mapping

**Linkage Disequilibrium**

- **Precision Lasso** (Wang, Lengerich, Aragam, Xing)
- Stochastic block Regression (Kim & Xing UAI 2008)

**Population Structure**

- Multi-population Group Lasso (Puniyani, Kim, Xing, ISMB, 2010)
- Linear Mixed Model
Variable Selection on structured data is common

- Given a high-dimensional dataset, which features should we investigate for causality?
- Similar methods for both SNP and microarray data

“Hybrid Subspace Learning for High-Dimensional Data”. Marchetti-Bowick and Lengerich, et al. 2017
Precision Lasso

- Goal: Select active variables, even when variables are correlated and/or linearly dependent
Lasso

- Traditional Lasso solves the optimization problem:

\[
\arg \min_{\beta} (y - X\beta)^2 + \lambda |\beta|
\]

Traits (n x 1)  SNPs (n x p)  Effect sizes (p x 1)  Sparse effect sizes
Lasso fails on correlated variables

- If two explanatory variables are highly correlated and effect sizes are unconstrained, then the explanatory variables show very similar influence on the response variable.
Lasso fails on correlated variables

- If two explanatory variables are highly correlated and effect sizes are unconstrained, then the explanatory variables show very similar influence on the response variable.
  - But the Lasso will only select one variable.
Lasso fails on correlated variables

- If two explanatory variables are highly correlated and effect sizes are unconstrained, then the explanatory variables show very similar influence on the response variable.
  - But the Lasso will only select one variable.

- Given explanatory variables $X^i, X^j, X^k$ with effect sizes $\beta_i, \beta_j, \beta_k$, if $X^k = aX^i + cX^j$, $a\beta_i \geq 0$ and $c\beta_j \geq 0$, then the Lasso will select the combined variable $X^k$. 
Correlated and Linearly Dependent Variables are Common in Genomic Data

<table>
<thead>
<tr>
<th></th>
<th>Glioblastoma</th>
<th></th>
<th>Breast</th>
<th></th>
<th>Lung</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NIV</td>
<td>CV</td>
<td>NIV</td>
<td>CV</td>
<td>NIV</td>
<td>CV</td>
</tr>
<tr>
<td>Gene Expression</td>
<td>0.083%</td>
<td>0.0</td>
<td>0.095%</td>
<td>0.218%</td>
<td>0.174%</td>
<td>0</td>
</tr>
<tr>
<td>Methylation</td>
<td>0.749%</td>
<td>0.0</td>
<td>5.084%</td>
<td>0.086%</td>
<td>4.261%</td>
<td>4.546%</td>
</tr>
<tr>
<td>miRNA</td>
<td>0.919</td>
<td>0.374%</td>
<td>4.018%</td>
<td>4.510%</td>
<td>3.714%</td>
<td>13.35%</td>
</tr>
</tbody>
</table>

“CV” = Frequency of Highly Correlated Variables
(correlation > 0.99)

“NIV” = Frequency of Non-Irrepresentable Variables
(Linearly Dependent)
Related Work: Group Lasso

\[ \| \mathbf{w} \|_{GL} = \sum_{i=1}^{k} \| \mathbf{w}_{S_i} \|_2. \]

- Given a partitioning of the variables, enforces sparsity at the group level
Related Work: Group Lasso

\[ \| \mathbf{w} \|_{GL} = \sum_{i=1}^{k} \| \mathbf{w} S_i \|_2. \]

- Given a partitioning of the variables, enforces sparsity at the group level
- Requires knowing the partitioning of variables
Related Work: Elastic Net

\[ L(\lambda_1, \lambda_2, \beta) = |y - X\beta|^2 + \lambda_2|\beta|^2 + \lambda_1|\beta|_1 \]

- Strictly convex \(\Rightarrow\) similar estimated effect sizes for correlated variables
Related Work: Elastic Net

$$L(\lambda_1, \lambda_2, \beta) = |y - X\beta|^2 + \lambda_2|\beta|^2 + \lambda_1|\beta|_1$$

- Strictly convex => similar estimated effect sizes for correlated variables
- Blindly applies same penalty to all variables
Related Work: Trace Lasso

\[
\min_{\mathbf{w}} \frac{1}{2} \left\| \mathbf{y} - \mathbf{Xw} \right\|_2^2 + \lambda \left\| \mathbf{X Diag(w)} \right\|_\star.
\]

- Convex approximation of the dimension of the subspace spanned by the selected predictors
- Behaves like Lasso regularization for uncorrelated predictors, Tikhonov regularization for correlated predictors
Related Work: Trace Lasso

\[
\min_w \frac{1}{2} \| y - Xw \|_2^2 + \lambda \| X \text{Diag}(w) \|_*. 
\]

- Convex approximation of the dimension of the subspace spanned by the selected predictors
- Behaves like Lasso regularization for uncorrelated predictors, Tikhonov regularization for correlated predictors
- Does not consider linearly dependent variables
Precision Lasso: Use covariance structure of Precision Matrix to Adapt Regularization

\[
\arg \min_\beta \frac{1}{2} \| Y - X \beta \|_2^2 + \lambda \| \gamma (X^T X) + (1 - \gamma) (X^T X)^{-\frac{1}{2}} \| \text{diag}(\beta) \|_\star
\]

Reconstruction Error
Covariance
Linear Dependence
Experiments

Three types:

- **Synthetic data** to estimate accuracy
- **Semi-Synthetic data** to estimate accuracy on data that follows real distributions
- **Real microarray data** to estimate utility
Results: PL selects Active Variables (Synthetic Data)

F1 Scores from synthetic data with n=100 samples, p=1000 total variables, and k active variables.

Methods are: Wald Hypothesis Testing (Wald), Lasso, Ridge Regression (RR), Adaptive Lasso (AL), SCAD, MCP, Trace Lasso (TL), Inverse Covariance Regularizer (IC) and Precision Lasso (PL)
Results: PL selects Active Variables (Semi-Synthetic Data)

F1 Scores from semi-synthetic data with n=100-500 samples, p=13,000 total variables, and k=130 active variables. Design matrix taken from real gene expression data, effect sizes drawn from standard normal.

Methods are: Wald Hypothesis Testing (Wald), Lasso, Ridge Regression (RR), Adaptive Lasso (AL), SCAD, MCP, Trace Lasso (TL), Inverse Covariance Regularizer (IC) and Precision Lasso (PL)
Precision Lasso Selects Meaningful Genes from Breast Cancer Tumor Samples

<table>
<thead>
<tr>
<th>Variable Selection Method</th>
<th>Selected Genes with known carcinogenic somatic mutations</th>
<th>Known Tumor Tissues</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wald Test</td>
<td><strong>PPARG</strong></td>
<td>Thyroid</td>
</tr>
<tr>
<td></td>
<td><strong>EBF1</strong></td>
<td>Skin</td>
</tr>
<tr>
<td></td>
<td><strong>TPM3</strong></td>
<td>Thyroid, Breast, Lymph, Lung, Skin</td>
</tr>
<tr>
<td>Lasso</td>
<td><strong>HMGA2</strong></td>
<td>Skin, Uterus, Salivary Gland</td>
</tr>
<tr>
<td></td>
<td><strong>COL1A1</strong></td>
<td>Skin, Bone</td>
</tr>
<tr>
<td>Precision Lasso</td>
<td><strong>RAC1</strong></td>
<td>All Epithelial, Skin</td>
</tr>
<tr>
<td></td>
<td><strong>AKT1</strong></td>
<td>Breast, Colon, Ovary, Lung</td>
</tr>
<tr>
<td></td>
<td><strong>CLTC</strong></td>
<td>Breast, Lymph, Kidney</td>
</tr>
<tr>
<td></td>
<td><strong>ZNF384</strong></td>
<td>Blood, Bone Marrow</td>
</tr>
<tr>
<td></td>
<td><strong>DAXX</strong></td>
<td>Pancreas, Brain, Spine</td>
</tr>
</tbody>
</table>

**Bolded genes** are known to be associated with breast cancer. No other method selected a relevant gene.
Visual Machine Learning for Next-Generation GWAS

GENAMAP

TRY NOW

Email Address

KEEP ME INFORMED
● Live:
  ● genamap.org
  ● Click “Get Started” to get started

● Source Code:
  ● github.com/blengerich/GenAMap
  ● Set up your own server
GenAMap

Data Management  Algorithm Automation  Interactive Visualizations
Data Management - Local Data
Data Management - GDC Data
Algorithm Automation

Hypothesis Testing

Run Analysis

Association Study

Hypothesis Analysis

- Fisher Exact Test
- Chi-squared Test
- Wald Test

Marker1

Trait1
Algorithm Automation

Hypothesis Testing

Structured Association

Run Analysis

Association Study

Structure Association

- Linear Regression
- Adaptive Multi-Task Lasso
- GfLasso
- Multi-Population Lasso
- Tree Lasso

CANCEL  RUN ANALYSIS
Algorithm Automation

- Hypothesis Testing
- Structured Association
- Population Stratification

Run Analysis

Association Study

Structure Association

- LMM
- sparse LMM
- ...

Marker1

Trait1
Interactive Visualizations

- **Matrix View**
  - Trait names
  - Detailed Information card
  - Color bar
  - Chromosome browser
  - Significance threshold
  - Marker names
  - Other Visualizations

Matrix View, each square denotes an effect size
Interactive Visualizations

Dendrogram

Manhattan Plot
Future Work

- Precision Lasso Scalability
- Network Inference and Network-guided GWAS
- Transcriptomic Structure
- GenAMap Improvements
  - Scalability
  - Visualizations
  - User studies
Thanks for your attention!