Detecting Genetic Interactions in Genome-Wide Association Study

Wei Wang
University of North Carolina at Chapel Hill
Systems Genetics

- Complex Models
- Heterogeneous Data
- Powerful / Robust Methods and Systems

- Tissues, cultures
- Cells
- Pathways
- Biochemical reactions
- Proteins, genes...

- Organism specificity
- Functional modules
- Large-scale organization
- Regulatory motifs
- Metabolic pathways
- mRNA
- Proteins
- Metabolites

- Information storage
- Processing
- Execution
Why is GWA Important?

• Once new genetic associations are identified
  ➢ Researchers can develop better strategies to detect, treat and prevent the disease
    ❖ Asthma, Cancer, Diabetes, Heart Disease
  ➢ Health professionals can provide personalized medicine.
Why is GWA Important?

A video clip from Gattaca (1997), (from youtube)
Detecting Genetic Interactions: Motivation

- Example: Mouse Colon Cancer
  
  Two important genes
  \[
  \begin{align*}
  Ptp\text{r}j & \quad (\text{chrom} \ 2) \\
  L\text{r}g1 & \quad (\text{chrom} \ 6)
  \end{align*}
  \]
  
  Detectible only when studying interactions

- True for many common diseases
Detecting Genetic Interactions: Challenges

Statistical – Statistics to capture the interactions

Computational – Hundreds of billions of potential interactions
Detecting Genetic Interactions: Previous Approaches

• Exhaustive [Moore et al. ’06, Purcell et al. ’07]
  ➢ Not scalable
• Heuristic [Carlborg et al.’00, Nakamichi et al.’01]
  ➢ Not optimal
• Two-step [Evans et al. ’06, Yang et al. ’09]
  ➢ Filter, then search (Not optimal)
• Algorithm development is in early stage
Detecting Genetic Interactions: Our Contributions

- **Efficiency and Optimality** [SIGKDD’08, PSB’09]
  - Dramatically reduced the computational burden
  - Guaranteed optimal solution

- **Applicability** [RECOMB’09, ISMB’10]
  - A wide range of study types and statistics

- **The first** to address these issues systematically
Outline of the Talk

• Background
  ➢ SNPs and their interactions
  ➢ Computational problems

• Algorithms for Detecting Genetic Interactions
Single Nucleotide Polymorphism: SNP

SNP – mutation of a single nucleotide in the DNA sequence

The most common form of genetic variation

Valuable for diagnostics and drug development
### SNPs as Binary Variables

**SNPs 1**

<table>
<thead>
<tr>
<th>Sample 1</th>
<th>Sample 2</th>
<th>Sample 3</th>
<th>Sample 4</th>
<th>Sample 5</th>
<th>Sample 6</th>
<th>Sample 7</th>
<th>Sample 8</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 0</td>
<td>1 0</td>
<td>1 1</td>
<td>1 1 1 1</td>
<td>1 1</td>
<td>1 0</td>
<td>1 1 1 1</td>
<td>0 1 1 1</td>
</tr>
</tbody>
</table>

**SNPs 2**

<table>
<thead>
<tr>
<th>Sample 1</th>
<th>Sample 2</th>
<th>Sample 3</th>
<th>Sample 4</th>
<th>Sample 5</th>
<th>Sample 6</th>
<th>Sample 7</th>
<th>Sample 8</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 1 0</td>
<td>1 0 0 0 1</td>
<td>1 1 1 1 1</td>
<td>1 0 0 1 1</td>
<td>1 0 0 0 0</td>
<td>1 0 1 1 0</td>
<td>0 1 1 1 1</td>
<td>0 1 1 1 1</td>
</tr>
</tbody>
</table>

---

**Millions of SNPs in the whole genome**
Phenotype Variation

Phenotype – an observable characteristic or trait

http://www.jax.org/

Coat color

http://derc.ucsd.edu

Body weight
SNP-Phenotype Association Study

- Which SNPs cause the phenotype variation?

<table>
<thead>
<tr>
<th>SNPs</th>
<th>Pheno.</th>
</tr>
</thead>
<tbody>
<tr>
<td>.....0 1 0 1 0 1 .....</td>
<td>8</td>
</tr>
<tr>
<td>.....0 0 0 0 0 0 1 .....</td>
<td>7</td>
</tr>
<tr>
<td>.....0 1 1 0 0 0 1 .....</td>
<td>12</td>
</tr>
<tr>
<td>.....0 0 0 0 0 1 0 .....</td>
<td>11</td>
</tr>
<tr>
<td>.....1 1 1 1 1 1 1 .....</td>
<td>2</td>
</tr>
<tr>
<td>.....1 0 0 1 0 1 .....</td>
<td>5</td>
</tr>
<tr>
<td>.....1 1 0 1 0 1 .....</td>
<td>0</td>
</tr>
<tr>
<td>.....1 0 1 1 0 0 .....</td>
<td>3</td>
</tr>
</tbody>
</table>

Longstanding goal of genetic studies
Traditional Single-SNP Approach

- For every SNP
- Do a statistical test

Large test value

Strong association

\[ T(\text{SNP1}, \text{pheno}) = 28.2 \]
\[ T(\text{SNP2}, \text{pheno}) = 0.6 \]
Detecting SNP-SNP Interactions

• Complex phenotypes
  ➢ Diabetes, heart disease, etc …
  ➢ Joint effect of genetic factors

• SNP-SNP interactions
  ➢ Test for every SNP-pair

• A hot research area in Bioinformatics community
  [Hoh et al.’03, Hirschhorn et al.’05, Musani et al. ’07]
The Computational Problem

- Problem: Find all SNP-pairs that are significantly associated with phenotype

How to define it?
Multiple Testing Problem

Multiple tests increase the probability of false findings

- prob. of at least one false finding (error with each test = 0.05)
Permutation Test for Error Controlling

Goal: Find a threshold $\theta$

- Permute phenotype $K (=1000)$ times
- For each permutation, find max test value
- Threshold $\theta = \alpha \times K$-th largest value $(0.01 \leq \alpha \leq 0.05)$
- SNP-pairs $\geq \theta$ are significant

<table>
<thead>
<tr>
<th>SNPs</th>
<th>Pheno.</th>
</tr>
</thead>
<tbody>
<tr>
<td>...... 0 1 0 1 0 1 .....</td>
<td>0 31 32 31 32 32</td>
</tr>
<tr>
<td>...... 0 0 0 0 0 1 .....</td>
<td>0</td>
</tr>
<tr>
<td>...... 0 1 1 0 0 1 .....</td>
<td>1</td>
</tr>
<tr>
<td>...... 0 0 0 0 1 0 .....</td>
<td>2</td>
</tr>
<tr>
<td>...... 1 1 1 1 1 1 .....</td>
<td>5</td>
</tr>
<tr>
<td>...... 1 0 0 1 0 1 .....</td>
<td>8</td>
</tr>
<tr>
<td>...... 1 1 0 1 0 1 .....</td>
<td>12</td>
</tr>
<tr>
<td>...... 1 0 1 1 0 0 .....</td>
<td>11</td>
</tr>
</tbody>
</table>

$K (=1000)$ values

21.3 10.8 8.7 ...
The Computational Problem (Revisited)

- Problem 1: Find threshold by permutation test
- Problem 2: Find all significant SNP-pairs ($\geq \theta$)
- Brute force: enumerate all SNP-Pairs
- Permutation test is computationally intensive
Challenges

- Statistical – effective tests
  - ANOVA, chi-square, likelihood ratio, etc…

- Computational – huge search space
  - 100K SNPs and 1K permutations
    - Number of tests: 500 Billion
    - Can be easily MUCH LARGER

- Must be handled together
Our Solutions

• Efficiency and Optimality [SIGKDD’08, PSB’09]
  ➢ **Bound** on test statistic
  ➢ **Indexing** search space for bound estimation

• Applicability [RECOMB’09, ISMB’10]
  ➢ Common statistics are **convex**
  ➢ Computing contingency tables
Outline of the Talk

• Background
   SNP-SNP interactions
   Computational problem & Challenges

• Detecting SNP-SNP Interactions
   Algorithms for ANOVA and chi-square tests
   A general approach COE
   A more general approach TEAM
FastANOVA - Key Ideas

SIGKDD 2008

- **Bound** on test statistic
  - Filter out insignificant SNP-pairs

- **Indexing** structure
  - Compute the bound for a group of SNP-pairs

- Removal of redundant computation
The Upper Bound

\[ T(\text{SNP pair, pheno}) \leq \text{constant} + R_1 + R_2 \]

Need to be \( \geq \theta \) to be significant
The Upper Bound

\[
\begin{align*}
R_1 &= f(n_a) \\
R_2 &= f(n_b)
\end{align*}
\]

\[
\begin{align*}
 n_a : \min \# \{0,1\} \text{ in } X_j \text{ (when } X_i = 0) \\
n_b : \min \# \{0,1\} \text{ in } X_j \text{ (when } X_i = 1)
\end{align*}
\]
## Indexing SNP-Pairs

<table>
<thead>
<tr>
<th>$X_1$</th>
<th>$X_2$</th>
<th>$X_3$</th>
<th>$X_4$</th>
<th>$X_5$</th>
<th>$X_6$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Diagram:

- $(X_1X_3)$
- $(X_1X_5)$
- $(X_1X_6)$
- $(X_1X_2)$
- $(X_1X_4)$

Indexing SNP-Pairs...
Properties of the Indexing Structure

- Many pairs **share** an entry
- Pairs in an entry have the **same upper bound**
- **Built only once**, reused in all permutations

![Diagram showing properties of the indexing structure]

- (X_1X_3)
- (X_1X_5)
- (X_1X_6)
- (X_1X_2)
- (X_1X_4)
- (X_1..)
FastANOVA - Overall Process

- For each SNP
  - Index its associated pairs
- For each permutation
  - Find the candidate pairs \((ub \geq \theta)\)
  - Evaluate test values of the candidates
FastANOVA - Complexity

• Time
  ➢ Brute force: $O(KN^2M)$
  ➢ FastANOVA: $O(N^2M + KNM^2 + CM)$

• Space
  ➢ $O((N+K)M)$

\[
\begin{align*}
N &= \# \text{ SNPs} \\
M &= \# \text{ individuals} \\
K &= \# \text{ permutations} \\
C &= \# \text{ candidates}
\end{align*}
\]

$M \ll N$
Brute Force v.s. FastANOVA

#SNPs = 44k, #individuals = 26, phenotype: metabolism (water intake)

Data available at http://www.jax.org
Pruning Power of the Bound

SNPs Pruned

Error Rate
The FastChi Algorithm

ANOVA (for quantitative pheno.)

\[ T(\text{SNP pair, pheno}) \leq \text{constant} + R_1 + R_2 \]

Chi-square (for binary pheno.)

\[ T'(\text{SNP pair, pheno}) \leq \text{constant}' + R_1' + R_2' \]
COE - A General Approach

Available tests
- Chi-square, likelihood ratio, trend, entropy-based, etc …
- Active research, more being proposed …

A unified approach to all above tests?

Convexity is the solution!
Convexity Example: Chi-square Test

- **Theorem**: Convexity is a common property of many tests
- Determine the range of the free variable to get the upper bound
Brute Force vs. FastChi vs. COE

(a) 

Error Rate

Runtime (sec.)

10^2

10^3

10^4

0.05 0.04 0.03 0.02 0.01

(b) 

Number of SNPs x 10^4

Runtime (sec.)

10^1

10^2

10^3

10^4

2 4 6

Brute Force
FastChi
COE Chí
Brute Force vs. COE (on Various Tests)

G-Test

Entropy-based Test

Trend-Test
Summary of the Algorithms

- Key ideas: bound, indexing, convexity

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Supported Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>FastANOA</td>
<td>ANOVA test</td>
</tr>
<tr>
<td>FastChi</td>
<td>Chi-square test</td>
</tr>
<tr>
<td>COE</td>
<td>Convex tests</td>
</tr>
</tbody>
</table>

- Designed for inbred mouse data: small sample size, binary SNPs
<table>
<thead>
<tr>
<th></th>
<th>Previous</th>
<th>TEAM</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>SNPs</strong></td>
<td>{0,1}</td>
<td>{0,1} &amp; {0,1,2}</td>
</tr>
<tr>
<td><strong>Sample size</strong></td>
<td>Small</td>
<td>Large</td>
</tr>
<tr>
<td><strong>Error Control</strong></td>
<td>FWER</td>
<td>FWER &amp; FDR</td>
</tr>
<tr>
<td><strong>Test Statistic</strong></td>
<td>With certain properties</td>
<td>Based on contingency table</td>
</tr>
</tbody>
</table>
The Computational Problem

- SNPs \( \{X_1, X_2, \ldots, X_N\} \)
- Phenotype \( Y \), and permutations \( \{Y_1, Y_2, \ldots, Y_k\} \)

Problem: Computing Test Values for SNP-pairs

Computing Contingency Tables
# Contingency Table

<table>
<thead>
<tr>
<th></th>
<th>$X_j = 0$</th>
<th>$X_j = 1$</th>
<th>$X_j = 2$</th>
<th>$X_j = 0$</th>
<th>$X_j = 1$</th>
<th>$X_j = 2$</th>
<th>$X_j = 0$</th>
<th>$X_j = 1$</th>
<th>$X_j = 2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Y = 0$</td>
<td>$a_1$</td>
<td>$a_2$</td>
<td>$a_3$</td>
<td>$b_1$</td>
<td>$b_2$</td>
<td>$b_3$</td>
<td>$e_1$</td>
<td>$e_2$</td>
<td>$e_3$</td>
</tr>
<tr>
<td>$Y = 1$</td>
<td>$c_1$</td>
<td>$c_2$</td>
<td>$c_3$</td>
<td>$d_1$</td>
<td>$d_2$</td>
<td>$d_3$</td>
<td>$f_1$</td>
<td>$f_2$</td>
<td>$f_3$</td>
</tr>
</tbody>
</table>
## Contingency Table

<table>
<thead>
<tr>
<th></th>
<th>$X_i = 0$</th>
<th>$X_i = 1$</th>
<th>$X_i = 2$</th>
<th></th>
<th>$X_j = 0$</th>
<th>$X_j = 1$</th>
<th>$X_j = 2$</th>
<th>$X_j = 0$</th>
<th>$X_j = 1$</th>
<th>$X_j = 2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Y = 0$</td>
<td>$a_1$</td>
<td>$a_2$</td>
<td>$a_3$</td>
<td>$b_1$</td>
<td>$b_2$</td>
<td>$b_3$</td>
<td>$e_1$</td>
<td>$e_2$</td>
<td>$e_3$</td>
<td></td>
</tr>
<tr>
<td>$Y = 1$</td>
<td>$c_1$</td>
<td>$c_2$</td>
<td>$c_3$</td>
<td>$d_1$</td>
<td>$d_2$</td>
<td>$d_3$</td>
<td>$f_1$</td>
<td>$f_2$</td>
<td>$f_3$</td>
<td></td>
</tr>
</tbody>
</table>

\[ \# \left( X_i, X_j, Y \right) = (1,1,1) \]
### Contingency Table

<table>
<thead>
<tr>
<th></th>
<th>$X_j = 0$</th>
<th>$X_j = 1$</th>
<th>$X_j = 2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_i = 0$</td>
<td>$X_i = 0$</td>
<td>$X_i = 1$</td>
<td>$X_i = 2$</td>
</tr>
<tr>
<td>$Y = 0$</td>
<td>$a_1$</td>
<td>$a_2$</td>
<td>$a_3$</td>
</tr>
<tr>
<td></td>
<td>$b_1$</td>
<td>$b_2$</td>
<td>$b_3$</td>
</tr>
<tr>
<td></td>
<td>$e_1$</td>
<td>$e_2$</td>
<td>$e_3$</td>
</tr>
<tr>
<td>$Y = 1$</td>
<td>$c_1$</td>
<td>$c_2$</td>
<td>$c_3$</td>
</tr>
<tr>
<td></td>
<td>$d_1$</td>
<td>$d_2$</td>
<td>$d_3$</td>
</tr>
<tr>
<td></td>
<td>$f_1$</td>
<td>$f_2$</td>
<td>$f_3$</td>
</tr>
</tbody>
</table>

Only need to compute four variables
Contingency Table

<table>
<thead>
<tr>
<th></th>
<th>$X_j = 0$</th>
<th>$X_j = 1$</th>
<th>$X_j = 2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_i = 0$</td>
<td>$a_1$</td>
<td>$a_2$</td>
<td>$a_3$</td>
</tr>
<tr>
<td>$X_i = 1$</td>
<td>$b_1$</td>
<td>$b_2$</td>
<td>$b_3$</td>
</tr>
<tr>
<td>$X_i = 2$</td>
<td>$e_1$</td>
<td>$e_2$</td>
<td>$e_3$</td>
</tr>
<tr>
<td>$Y = 0$</td>
<td>$c_1$</td>
<td>$c_2$</td>
<td>$c_3$</td>
</tr>
<tr>
<td>$Y = 1$</td>
<td>$d_1$</td>
<td>$\textbf{d}_2$</td>
<td>$d_3$</td>
</tr>
</tbody>
</table>

\[\# \ (X_i, X_j, Y) = (1,1,1)\]
Incremental Update

<table>
<thead>
<tr>
<th>Y</th>
<th>X_1</th>
<th>X_2</th>
<th>X_3</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

\[
\# (X_i, X_j, Y) = (1,1,1) \quad \rightarrow \quad d_2
\]
Incremental Update

\[
\begin{array}{|cccc|}
\hline
Y & X_1 & X_2 & X_3 \\
\hline
0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 \\
0 & 1 & 1 & 0 \\
0 & 1 & 0 & 0 \\
0 & 1 & 0 & 1 \\
1 & 1 & 1 & 0 \\
1 & 0 & 1 & 1 \\
1 & 1 & 0 & 0 \\
1 & 1 & 0 & 1 \\
1 & 1 & 1 & 1 \\
1 & 0 & 1 & 0 \\
\hline
\end{array}
\]

\# (X_i, X_j, Y) = (1,1,1) \quad \leftrightarrow \quad d_2

(\ X_1, \ X_2, \ Y) \quad \leftrightarrow \quad d_2 = 2

(\ X_1, \ X_3, \ Y) \quad \leftrightarrow \quad d_2 = ?
### Incremental Update

<table>
<thead>
<tr>
<th>Y</th>
<th>$X_1$</th>
<th>$X_2$</th>
<th>$X_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

\[
\#(X_i, X_j, Y) = (1, 1, 1) \quad \text{↔} \quad d_2
\]

\[
(X_1, X_2, Y) \quad \text{↔} \quad d_2 = 2
\]

\[
(X_1, X_3, Y) \quad \text{↔} \quad d_2 = 1
\]

- No need to scan all individuals
- Cost proportional to the difference
- Updating order? – Minimal Spanning Tree
TEAM v.s. Brute Force (Human Data)

Data generated by Hapsample

#SNPs = 100K
#Samples = 400
#Permutations = 100
Case/Control = 1
TEAM v.s. COE (Inbred Mouse Data)

Real Mouse Genotype Data (from Jackson Lab)

- #SNPs = 10K
- #Samples = 71
- #Permutations = 100
- Case/Control = 1
TEAM - Summary

- Designed for human study: large sample size, \{0,1,2\} SNPs
- Idea: incrementally update contingency tables
Overall Summary on Detecting Genetic Interactions

- Studying SNP-SNP interactions is important

- Challenges
  - Statistical: effective statistics
  - Computational: enormous search space

- We provide first solutions to
  - Efficiency and Optimality
  - Applicability
Acknowledgements

• Xiang Zhang

• Fernando Pardo-Manuel de Villena
• Wei Sun
• David Threadgill
• Yuying Xie
• Fei Zou
Thank You !