Genetic Effects and Statistical Power of Gene Hunting Using GWAS and Sequence Data

Zheyang Wu, Ph.D.

Department of Mathematical Sciences, and Program of Bioinformatics and Computational Biology
Worcester Polytechnic Institute

September 17, 2013
Content

Whole Genome Association Tests and Missing Heritability

Interactive Genetic Effects: Power of SNV Searching Strategies

Weak and Sparse Genetic Effects: Detection Boundary and Optimal SNV-set Method
Whole Genome Association Tests and Missing Heritability

Interactive Genetic Effects: Power of SNV Searching Strategies

Weak and Sparse Genetic Effect: Detection Boundary and Optimal SNV-set Method
Gene Hunting on Whole Genome (Array or Sequencing)

Candidate gene studies

Genome-wide Association Studies

From Francis S. Collins

From Francis S. Collins
GWAS Achievements

Published Genome-Wide Associations through 12/2012
Published GWA at $p \leq 5 \times 10^{-8}$ for 17 trait categories
Missing Heritability

Cumulative fraction of genetic variance explained by 71 Crohn’s disease risk loci (Franke et al., *Nature Genetics* 2010).
### Data

<table>
<thead>
<tr>
<th>Outcomes</th>
<th>Genetic Markers</th>
<th>Envir. Factors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Height</td>
<td>SNP 3&lt;sup&gt;+&lt;/sup&gt;</td>
<td>nutrition</td>
</tr>
<tr>
<td>Cancer</td>
<td>SNP 4</td>
<td>Gender</td>
</tr>
<tr>
<td></td>
<td>SNP 5</td>
<td></td>
</tr>
</tbody>
</table>

**Genetic DNA**

- **SNP 3<sup>+</sup>**
  - Person 1: G, G, C, ?
  - Person 2: T, T, T, ?
  - Person 3: T, T, C, ?

- **SNP 4**
  - Person 1: C, C, A, A
  - Person 2: C, C, A, A
  - Person 3: C, C, A, A

- **SNP 5**
  - Person 1: T, T, C, ?
  - Person 2: T, T, C, ?
  - Person 3: T, T, C, ?

**Height**

- Person 1: 6.1
- Person 2: 5.4
- Person 3: 4.7

**Cancer**

- Person 1: 1
- Person 2: 0
- Person 3: 0

**Nutrition**

- Person 1: 2.3
- Person 2: 1.6
- Person 3: ?

**Gender**

- Person 1: M
- Person 2: M
- Person 3: F

Features of Data and Genetic Effects

Features of data:
- High dimension: up to several millions of SNVs and other genetic factors (i.e., CNVs).
- Linkage disequilibrium (correlation).

Features of genetic effects:
- Epistasis (gene-gene interaction).
- Relatively weak and sparse signals at population level.

Single-SNV analysis are likely not optimal.
Genetic Effects: Interaction

Example: Rheumatoid Arthritis GWAS data from NARAC. Association tests for single SNVs, and for pair-wise SNVs with interaction terms. **Interaction effects could be significant.**

Top LLR test statistics of the interactions terms.
Genetic Effects: Interaction

Marginal associations of strong interactive SNVs could be weak.

Log of the ranks of marginal association with large interaction estimates.
Genetic Effects: Weak Associations At Population Level

Whole Genome Association Tests and Missing Heritability

Interactive Genetic Effects: Power of SNV Searching Strategies

Weak and Sparse Genetic Effect: Detection Boundary and Optimal SNV-set Method
SNV Search Strategies

**Strategies:** Detect disease-associated SNVs through model selection: SNVs in the top models are selected

▶ Marginal search: single-SNV analysis
▶ Exhaustive search
▶ Forward search
▶ Two-stage screen search

**Key question:** Which model selection methods have higher power, considering gene-gene interaction pattern.
Genetic Models

Genotype of the $j$th marker in the $i$th individual is

$$X_{ij} = \begin{cases} 
2 & \text{Genotype } = A_jA_j, \text{ with probability } p_j^2 \\
1 & \text{Genotype } = A_ja_j, \text{ with probability } 2p_j(1 - p_j) \\
0 & \text{Genotype } = a_ja_j, \text{ with probability } (1 - p_j)^2 
\end{cases}$$

Assume SNVs 1 and 2 are associated.

- **Quantitative trait model:**
  $$Y_i = b_0 + b_1 X_{i1} + b_2 X_{i2} + b_3 X_{i1}X_{i2} + \epsilon_i.$$

- **Binary trait model:** The odds of disease is
  $$O(x_1, x_2) = \frac{p(D|x_1, x_2)}{p(\bar{D}|x_1, x_2)}.$$
Linkage disequilibrium (LD)

- Causative SNVs may not be observed but they have LD with the genotyped markers.
- Assume causative but unobserved SNVs indexed by, say -1 and -2, respectively.
- The observed SNVs 1 and 2 are indirectly associated with the disease through the LD.
- The odds of disease at SNVs 1 and 2 is

\[
\frac{p(D|x_1, x_2)}{p(\bar{D}|x_1, x_2)} = \frac{\sum_{x_{-1}, x_{-2}} p(D|x_{-1}, x_{-2}) p(x_1|x_{-1}) p(x_2|x_{-2}) p(x_{-1}, x_{-2})}{\sum_{x_{-1}, x_{-2}} p(\bar{D}|x_{-1}, x_{-2}) p(x_1|x_{-1}) p(x_2|x_{-2}) p(x_{-1}, x_{-2})}.
\]
1D and 2D Scan

SNV search by regression model fitting:

- 1D scan for individual SNVs.

\[
\text{Link}(Y_i) = \hat{\beta}_0 + \hat{\beta}_1 x_{ij},
\]

- 2D scan with interaction terms:

\[
\text{Link}(Y_i) = \hat{\beta}_{0jk} + \hat{\beta}_{1jk} x_{ij} + \hat{\beta}_{2jk} x_{ik} + \hat{\beta}_{3jk} x_{ij} x_{ik}.
\]

Test statistics:

- For quantitative trait: T and F statistics for regression.
- For binary trait: Score test for logit.
1D and 2D Scan

Correlation structure among the test statistics:

1-D search

\[ T_1 \quad T_2 \quad T_3 \quad T_4 \quad T_5 \quad \ldots \]

2-D search

\[ F_{12} \quad F_{13} \quad F_{14} \quad F_{15} \quad F_{16} \quad \ldots \]

\[ F_{23} \quad F_{24} \quad F_{25} \quad F_{26} \quad \ldots \]

\[ F_{34} \quad F_{35} \quad F_{36} \quad \ldots \]

\[ F_{45} \quad F_{46} \quad \ldots \]

\[ F_{56} \quad \ldots \]
The **statistical power** is defined as the probability of either

- Finding the true model (or both SNVs in marginal selection), or
- Finding a model with either true SNV, under the

**Error control:**

- Discovery number control: Select SNVs in the top $R$ most significant models.
- Bonferroni type I error rate control.
Power Comparisons: Quantitative Trait

Finding both SNVs
Marginal Search vs. Exhaustive Search

\[ b_1 = b_2 \in (-1, 1) \]
\[ b_3 \in (-1, 1) \]
\[ R = 10 \]
\[ n = 1000 \]
\[ q = 0.3 \]
\[ \sigma^2 = 3 \]
\[ p = 300,000 \]
Power comparisons: Quantitative Trait

Finding **either** SNVs
Marginal Search vs. Exhaustive Search

\[ b_1 = b_2 \in (-1,1) \]
\[ b_3 \in (-1,1) \]
\[ R = 10 \]
\[ n = 1000 \]
\[ q = 0.3 \]
\[ \sigma^2 = 3 \]
\[ \rho = 300,000 \]
Power comparisons: Quantitative Trait

Power.Marg – Power.Exha
Power.Marg – Power.Forz
Power.Forz – Power.Exha

Both SNPs
$R = 1$

Both SNPs
$R = 10$

Either SNP
$R = 1$

Either SNP
$R = 10$
Power comparisons: Binary Trait

True model is assumed logit.

Discovery number $R = 10$. 
Power comparisons: Binary Trait

True model is assumed logit.
Bonferroni type I error rate $\alpha = 0.05$. 
When Marginal Effects Fixed

Setup from Marchini et. al *Nat Genet* 2005

Consider three genetic models M1-M3:

**M1:**

<table>
<thead>
<tr>
<th>$x_1=2$</th>
<th>$x_2=2$</th>
<th>$x_2=1$</th>
<th>$x_2=0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_1=2$</td>
<td>$\alpha (1 + \theta_1)^2 (1 + \theta_2)^2$</td>
<td>$\alpha (1 + \theta_1)^2 (1 + \theta_2)$</td>
<td>$\alpha (1 + \theta_1)^2$</td>
</tr>
<tr>
<td>$x_1=1$</td>
<td>$\alpha (1 + \theta_1) (1 + \theta_2)^2$</td>
<td>$\alpha (1 + \theta_1) (1 + \theta_2)$</td>
<td>$\alpha (1 + \theta_1)$</td>
</tr>
<tr>
<td>$x_1=0$</td>
<td>$\alpha (1 + \theta_2)^2$</td>
<td>$\alpha (1 + \theta_2)$</td>
<td>$\alpha$</td>
</tr>
</tbody>
</table>

**M2:**

<table>
<thead>
<tr>
<th>$x_1=2$</th>
<th>$x_2=2$</th>
<th>$x_2=1$</th>
<th>$x_2=0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_1=2$</td>
<td>$\alpha (1 + \theta)^4$</td>
<td>$\alpha (1 + \theta)^2$</td>
<td>$\alpha$</td>
</tr>
<tr>
<td>$x_1=1$</td>
<td>$\alpha (1 + \theta)^2$</td>
<td>$\alpha (1 + \theta)$</td>
<td>$\alpha$</td>
</tr>
<tr>
<td>$x_1=0$</td>
<td>$\alpha$</td>
<td>$\alpha$</td>
<td>$\alpha$</td>
</tr>
</tbody>
</table>

**M3:**

<table>
<thead>
<tr>
<th>$x_1=2$</th>
<th>$x_2=2$</th>
<th>$x_2=1$</th>
<th>$x_2=0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_1=2$</td>
<td>$\alpha (1 + \theta)$</td>
<td>$\alpha (1 + \theta)$</td>
<td>$\alpha$</td>
</tr>
<tr>
<td>$x_1=1$</td>
<td>$\alpha (1 + \theta)$</td>
<td>$\alpha (1 + \theta)$</td>
<td>$\alpha$</td>
</tr>
<tr>
<td>$x_1=0$</td>
<td>$\alpha$</td>
<td>$\alpha$</td>
<td>$\alpha$</td>
</tr>
</tbody>
</table>
When Marginal Effects Fixed

Setup from Marchini et. al *Nat Genet* 2005

- Assume the three genetic models M1-M3 have the same marginal association: heterozygote odds ratio $\lambda = 1.5$ at each causative marker.
- Study the influence of LD: squared correlation coefficient $r^2 = 0.5, 0.7, \text{and} 1$.
- LD constraint $p(A_i|A_{-i}) = 1$ and $p(A_i|a_{-i}) = q$, $i = 1, 2$, where $A_{-i}$ is the disease-causing allele at the unobserved locus indexed by $-i$, $A_i$ is the disease allele at the genotyped locus of marker $i$, which is in LD with the causative locus $-i$.
- Assume $p(D) = 0.01$, $n_1 = n_0 = 2000$, MAF $p_j = 0.05, 0.1, 0.2, \text{and} 0.5$. Bonferroni type I error rate $\alpha = 0.05$. 

When Marginal Effects Fixed
When Marginal Effects Fixed

Weedon MN, et al. (Nat Genet, 2008):

- GWAS data: \(n = 16,482, p = 402,951\).
- Human height: \(s.d. = 6.82cm\).
- SNV rs11107116: marginal effect = 0.045s.d., MAF \(q_1 = 0.23\).
- SNV rs10906982: marginal effect = 0.046s.d., MAF \(q_2 = 0.48\).
- Consider a series possible interaction effect \(b_3\).
- Estimate coefficients \(b_1\) and \(b_2\), and error variance \(\sigma^2\) using the two-marker epistasis model.
When Marginal Effects Fixed

Finding both SNPs ($R=20$)

Finding either SNP ($R=20$)

$b_3 = 0.3$: Adjusted p-value for average interaction term $> 0.9$.

$b_3 = 0.6$: Adjusted p-value for average interaction term $= 0.003$. 
Remarks

- A novel method for power calculation of model selection
  - Relief computationally intensive simulations
  - Look into the mechanism of discovering genetic signal
  - Implemented R package: `markerSearchPower`

- Marginal selection is powerful for finding at least one SNVs in a large genetic parameter space.
- Exhaustive search is preferred for finding true genetic model.
- Forward selection based on only one SNV in the first stage is less preferred.
Remarks

- Binary trait vs. quantitative trait:
  - The correlation is stronger among score statistics for binary traits than that among F test statistics for quantitative traits.
  - Exhaustive search is more powerful for binary traits than for quantitative traits, especially in finding the joint association.

- Discovery number control vs. Bonferroni control.
  - The $\alpha$-control is more stringent than the $R$-control, and the $R$-control leads to higher statistical power.
  - With the Bonferroni control, we expect more joint associations to be found when applying exhaustive scan, especially for finding the joint associations.
Whole Genome Association Tests and Missing Heritability

Interactive Genetic Effects: Power of SNV Searching Strategies

Weak and Sparse Genetic Effect: Detection Boundary and Optimal SNV-set Method
Challenges of Weak and Sparse Genetic Effects

Genetic factors to be discovered have

- **weak** association at population level due to small effects and/or small variations, and is
- **sparse** among candidates.
SNV-Set Analysis

**SNV-set methods:** Study the association of multiple SNVs, usually grouped in functional segments.


- Minimal P-value method (Min_P)
- Multiple-covariate linear model (MulLm)
- Ridge regression (Ridge)
- Principal Component Analysis (PCA)
- Linear combination test (LCT)
- Quadratic Test (QT)
- Decorrelation Test (DT)
- Kernal-machine-based test (KMT)
- Fisher’s combination test
- False discovery rate method
- ...
Find the boundary that separates detectable and non-detectable genetic effects by *any* statistical method.

Find the *optimal* procedures: *asymptotically powerful*, i.e.,

\[ \text{Type I error} + \text{Type II error} \rightarrow 0, \]

whenever the genetic effects are above the detection boundary.

Genetic Model

Consider a set of \( L \) correlated SNVs of \( n \) independent individuals. Additive genetic model:

\[
Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_L X_L + \varepsilon,
\]

- \( Y = (Y_1, \ldots, Y_n)' \) is the trait vector.
- \( X_j = (X_{1j}, \ldots, X_{nj})' \) is the random genotype vector of the \( j \)-th SNV, \( 1 \leq j \leq L \), with \( X_{kj} \sim \text{binomial}(2, q_j) \), and LD structure: correlation matrix \( \Sigma \) among \( X_{k1}, \ldots, X_{kL} \): each row of \( \Sigma \) has no more than \( \Delta \) elements exceeding \( \gamma \) in magnitude.
- Error term \( \varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)' \sim N \left( 0, \sigma^2 I \right) \) with \( \sigma^2 \) is unknown.
Detection of Associated SNV Sets

- The $j$-th SNV is associated (i.e., a signal) if $\beta_j \neq 0$.
- The support of $\beta$, i.e., $\beta_j \neq 0$ for $j \in M^* \equiv \{j_1, \ldots, j_K\}$ is uniformly distributed over $\{1, 2, \ldots, L\}$.
- Given $j \in M^*$, $\text{sgn}(\beta_j) = \pm 1$, with equal probabilities.

The problem of detecting an associated SNV set can be reformulated as a hypothesis testing problem.

$$H_0 : \beta_j = 0, \text{ for all } 1 \leq j \leq L,$$
$$H_1 : \beta_j \neq 0 \text{ only for a small fraction of } j, 1 \leq j \leq L.$$
Detection Boundary of Genetic Effects

**Detection boundary:** A curve that separates the region of impossibility from the region of possibility, in the 2-D phase space calibrating the following:

- **Signal sparsity:** The number of associated SNVs is
  
  \[ K = L^{1 - \alpha}, \]
  
  where \( \alpha \) is called **sparse parameter**. Consider very sparse case \( \alpha \in (1/2, 1) \).

- **Signal strength:** The genetic effect is
  
  \[ |\beta_j| = \tau_j \frac{\sigma}{\sqrt{2nq_j(1 - q_j)}}, \]
  
  where
  
  \[ \tau_j = \sqrt{2r_j \log(L)}, \]
  
  \( r_j \) is called **strength parameter**.
Detection Boundary of Genetic Effects

- Assumptions:
  - The sample size \( n = O(L^a) \) for some constant \( a > 0 \).
  - The number of large correlations in each row of \( \Sigma \) is assumed to be \( \Delta = O(L^\varepsilon) \) for all \( \varepsilon > 0 \).

**Theorem**

*For the above defined the genetic model and assumptions, if the correlation \( \gamma \) and the L-n relative value \( \gamma' = \sqrt{\frac{\log L}{n}} \) satisfy\*(\( \gamma + \gamma' \)) \( L^{1-\alpha} (\log L)^4 \to 0 \), then all tests are asymptotically powerless if \( r_j < r^*(\alpha) \), \( j \in M^* \), where*\[
  r^*(\alpha) = \begin{cases} \alpha - 1/2, & 1/2 < \alpha \leq 3/4 \\ (1 - \sqrt{1 - \alpha})^2, & 3/4 < \alpha < 1 \end{cases}
\]*
Detection Boundary of Genetic Effect

Detection boundary on the planes of the proportion of true SNVs and the genetic effect (left) and Heritability (right) $L = 10,000$, $n = 1,000$, $\sigma = 1$, and $q_j = 0.3$. 
The HC test statistic for a set of SNVs with genotypes $X_1, \ldots, X_L$ is

$$HC_L = \max_{1 \leq j \leq L} \sqrt{L} \frac{i - p(j)}{\sqrt{p(j)(1 - p(j))}}.$$

where $p_j$, $j = 1, \ldots, L$, are the marginal P-values, and $p(1) \leq \ldots \leq p(L)$. 
Higher Criticism Procedure

- Quantitative traits: R-test or T-test:

\[ R_j = \sqrt{n - 1} \rho_j \text{ and } T_j = \sqrt{n - 2} \frac{\rho_j}{\sqrt{1 - \rho_j}}, \]

where \( \rho_j \) is the Pearson correlation between \( X_j \) and \( Y \). The P-values
\[ p_j = P (|N(0,1)| > |R_j|) \text{ (or) } \]
\[ p_j = P (|N(0,1)| > |T_j|) \]

- Binary traits: Z-test (Zuo and Zhao 2006)

\[ D_j = \sqrt{n} \frac{\hat{p}_{\text{case}} - \hat{p}_{\text{control}}}{\sqrt{2\hat{p}_{\text{all}} (1 - \hat{p}_{\text{all}})}}, \]

where \( \hat{p}_{\text{case}} \), \( \hat{p}_{\text{control}} \) and \( \hat{p}_{\text{all}} \) are the empirical allele frequencies in cases, controls, and the combined group, respectively. The P-values
\[ p_j = P (|N(0,1)| > |D_j|). \]
Theorem

Consider above genetic model. Under some mathematical assumptions, $HC_L$ based on $R_j$ has asymptotically full power if $r_j > r^* (\alpha)$, $j \in M^*$. 

![Graph showing detectable and undetectable regions for genetic effect and heritability.](image)
Other Procedures

Let $\mathbf{T} = (T_1, ..., T_L)'$ be a vector of marginal test statistics, $\hat{\Sigma}$ be the Pearson correlation coefficients among the SNVs

- Minimal P-value method: The association of the SNV set is measured by the minimal marginal P-value $p^{(1)}$.
- Linear combination test (LCT) statistic (let $\mathbf{e}$ be a vector of 1)
  \[
  T^L = T^L = \mathbf{e}' \mathbf{T} / \sqrt{\mathbf{e}' \hat{\Sigma} \mathbf{e}}.
  \]
- Quadratic test (QT) statistic
  \[
  T^Q = \mathbf{T}' \hat{\Sigma}^{-1} \mathbf{T}.
  \]
- Decorrelation test (DT) / Fisher’s combination test
  \[
  T^D = -2 \sum_{j=1}^{L} \log p_j,
  \]
  where the P-values $p_j = P \{|N(0,1)| > |W_j|\}$, and $W_j$ is the $j$th element of $\mathbf{W} = \mathbf{D}^{-1} \mathbf{T}$ with $\hat{\Sigma} = \mathbf{D}\mathbf{D}'$. 
SNV-Set Procedures
Numerical Analysis

Data generation

- MAF $q_j = 0.4$
- SNV-set size $L = 100$
- Number of true SNVs $K = 3$ (i.e., sparsity parameter $\alpha = 0.76$)
- LD structures:
  - Case 1: $\Sigma = I$
  - Cases 2-4: The 1st off-diagonal elements of $\Sigma$ are 0.3, 0.25, or 0.2.
  - Cases 5-6: The 1st and the 2nd off-diagonal elements of $\Sigma$ are 0.25 and 0.3, or 0.25 and 0.2.
Data Generation

- **Quantitative traits:**
  - Sample size $n = 1000$.
  - Strength parameter $r$ from 0.4 to 0.9 (i.e., $\beta$ from 0.088 to 0.131, or the heritability from 0.011 to 0.024).

- **Binary Traits:**
  - Genetic risk of disease
    \[
    \text{logit} \left( \frac{P(Y_k = 1|X_k)}{P(Y_k = 0|X_k)} \right) = \beta_0 + \beta_1 X_{k1} + \beta_2 X_{k2} + \ldots + \beta_L X_{kL},
    \]
  - 1000 cases and 1000 controls
  - $\beta_0 = -2$, non-zero $\beta_j$ from 0.1 to 0.24 (i.e., disease odds ratio from 1.11 to 1.27).
Statistical Power: Quantitative Traits
Statistical Power: Binary Traits

![Diagrams showing power as a function of covariates' coefficients](image)

- **Legend**:
  - HC
  - MinP
  - LKMT
  - QT
  - DT
  - LCT
  - Ridge
  - PCA
HC Applied to GWAS of Crohn’s Disease

- Data from NIDDK-IBDGC (National Institute of Diabetes, Digestive and Kidney Diseases - Inflammatory Bowel Disease Genetics Consortium, Duerr et al 2006 Science).
- 417 cases and 434 controls from Jewish population.
- 307,964 SNVs were grouped into 15,860 genes.
- The gene length (number of SNPs) ranges from 1 to 844 and is highly skewed to the right: the lower quartile, median and upper quartile are 3, 7 and 19, respectively.
- The top genes found by HC procedure were not obtained by minP methods: \textit{PFAAP5, AGTR1, CDA08, NXPH1, LCN10, OR51G1, FDXR, KIAA1904, and EDG1}.
- Catalogue of Somatic Mutations in Cancer (COSMIC) and literature indicate that some of these genes are likely to be relevant according to their functionalities.
Remarks

- Detection boundary for weak and sparse genetic effects.
- Minimal P-value (single-SNV) method does not reach the whole boundary.
- Some SNV-set methods may not be superior to single-SNV method.
- HC procedure reaches the boundary and thus is among the most powerful methods.
- For the defined genetic model, there is no need to consider joint association test (e.g., multiple regression, KMT, etc.)