Multi-loci association test in genetic association study using similarity between individuals

Indranil Mukhopadhyay
Human Genetics Unit
Indian Statistical Institute, India
Acknowledgement

• Anbupalam Thalamuthu
• Eleanor Feingold
• Daniel Weeks
• Kushal Dey
• Pronoy Kanti Mondal
• Sarmistha Das
Outline ...

- Some prelims
- Disease … Genetics … ??
- Finding a disease gene
- A new test for multi-loci association
The Human Genome

- Human genome is *diploid*, meaning we have two copies of each chromosome (one from each parent)

  ![Diagram of Father, Mother, and Child chromosomes]

- 22 pairs of chromosomes + 1 pair of sex chromosome
Prelims …

• Gene: Fundamental unit of genetic information that passes from generation to generation

• Allele: One of two or more states in which either copy of a gene can exist

• Marker: A polymorphic entity with known physical location
Genetic Markers

• Known location in genome
  – Human Genome Project tells us precisely where the markers are

• Unchanged from generation to generation

• Follow transmission from parents to offspring

• Be able to distinguish alleles
  – Polymorphic- having more than one state (alleles)
Complex disease

Marker locus

Disease gene 1

Other disease genes

Environment and culture

Phenotype

Slide by S Ghosh
SNP  
*Single Nucleotide Polymorphism*

1. ATCGCGGTAAATAGCTACGATACGCTGACTAGCATG
2. ATCGCGATAATAATAGCTACGATACGCTGATTAGCATG
So an SNP has only two alleles

Marker = SNP  Alleles: $a$ or $b$

Genotypes: $aa$, $ab$, $bb$

**Association**: A tendency for a particular genotype to occur more commonly in cases for a disease than expected by chance

**Association testing**: A testing method to test the possible existence of association between a phenotype and a candidate gene
### Basic methods of association

**Genotype-based Test**

<table>
<thead>
<tr>
<th></th>
<th>$aa$</th>
<th>$ab$</th>
<th>$bb$</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Case</strong></td>
<td>$n_1$</td>
<td>$n_2$</td>
<td>$n_3$</td>
<td>$S$</td>
</tr>
<tr>
<td><strong>Control</strong></td>
<td>$N_1$</td>
<td>$N_2$</td>
<td>$N_3$</td>
<td>$T$</td>
</tr>
</tbody>
</table>

Null hypothesis ($H_0$): no difference in the genotypic distributions of cases and controls.

$$
\chi^2 = \sum_{all \ cells} \frac{(O - E)^2}{E}
$$
An example

<table>
<thead>
<tr>
<th></th>
<th>aa</th>
<th>ab</th>
<th>bb</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case</td>
<td>50</td>
<td>40</td>
<td>10</td>
<td>100</td>
</tr>
<tr>
<td>Control</td>
<td>130</td>
<td>60</td>
<td>10</td>
<td>200</td>
</tr>
</tbody>
</table>

H$_0$: no difference in genotypic distributions

• Observed frequencies are given

• Calculate expected frequencies under H$_0$

\[
\begin{align*}
\text{Case} & : 60, 33, 7, 100 \\
\text{Control} & : 120, 66, 14, 200
\end{align*}
\]

• Calculate chi-square statistic

\[
\chi^2 = \sum_{\text{all cells}} \frac{(O-E)^2}{E} = 6.96 > \chi^2_{1,0.05} = 3.84
\]

• P-value = 0.008 < 0.05
Genome-wide Association analysis (GWAS)

1) Collect cases and controls.

2) Genotype everyone at a marker.

3) Test genotype/phenotype association.

<table>
<thead>
<tr>
<th></th>
<th>aa</th>
<th>ab</th>
<th>bb</th>
</tr>
</thead>
<tbody>
<tr>
<td>cases</td>
<td>50</td>
<td>40</td>
<td>10</td>
</tr>
<tr>
<td>controls</td>
<td>130</td>
<td>60</td>
<td>10</td>
</tr>
</tbody>
</table>

P-value = 0.008 : small enough !!!

4) Genotype everyone at all markers.

- Test at each locus
- Check P-value < 0.05
- Hurray! Found causal locus
I have found one locus !!!

Write paper, have beer
... have fun!
But this ‘world is not enough’

Why??? : let’s look carefully …
Simple, good, … but…

- Millions of SNPs
- Need for multiple comparison
- … … …

- May miss some true signals
- Need extremely large sample
- many other issues …
Let’s give a fresh look …

**H₀: no association**
**Idea**

- Individuals belonging to control group form a class, those having the disease (cases) form another class.
- Use variation between cases and controls and variation within each class.
- Similarity scores or values based on the genotype of each marker.
- We study each marker separately and combine them to get a global statistic that is finally used to detect disease-marker association.
\(g_i \rightarrow g_j \rightarrow h_{\text{con}}(g_i, g_j) = y_{ij,\text{con}}\)

\(h_{\text{case}}(g_i, g_j) = y_{ij,\text{case}}\)
\[ y_{lij} = \mu + e_{lij} \quad i < j = 1, 2, \ldots, n_l; \quad l = \text{case, control} \]

**H**₀: no association

&

**H**₀ is true
\[ y_{lij} = \mu + \alpha_l + e_{lij} \quad i < j = 1,2,\ldots, n_l; \quad l = case, control \]

We are same!

We are different!!

additional effect over general effect
Model

Let \( y_{lij} = h_l(g_i, g_j) \) denote the kernel score between \((i,j)\)-th pair in the \(l\)-th group

\[
y_{lij} = h_l(g_i, g_j) : \text{not uncorrelated}
\]

**TABLE 1. Kernel scores corresponding to different choices of additive kernels associated with pair of genotypes \(g_i\) and \(g_j\).**

<table>
<thead>
<tr>
<th>(g_i)</th>
<th>(g_j)</th>
<th>Allele match</th>
<th>Allele share</th>
<th>Linear dosage</th>
<th>Recessive</th>
<th>Quadratic</th>
</tr>
</thead>
<tbody>
<tr>
<td>a/a</td>
<td>a/a</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>a/a</td>
<td>a/b</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>a/a</td>
<td>b/b</td>
<td>0</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>a/b</td>
<td>a/a</td>
<td>4</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>a/b</td>
<td>a/b</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>a/b</td>
<td>b/b</td>
<td>0</td>
<td>4</td>
<td>2</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>b/b</td>
<td>a/a</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>b/b</td>
<td>a/b</td>
<td>4</td>
<td>0</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>b/b</td>
<td>b/b</td>
<td>0</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>5</td>
</tr>
</tbody>
</table>
Model

\[ y_{lij} = \mu + \alpha_l + e_{lij} \quad i < j = 1,2,\ldots,n_l; \quad l = 1,2 \]

(i) \( \alpha_1 + \alpha_2 = 0 \)

(ii) \( V(y_{lij}) = \sigma^2 \)

(iii) \( \text{Cov}(y_{lij}, y_{l'i'j'}) = \begin{cases} 
\rho \sigma^2 & \text{for } i \neq i' \text{ or } j \neq j' \text{ if } l = l' \\
0 & \text{if } l \neq l' 
\end{cases} \)

\{l=1\} \Rightarrow \text{case}, \quad \{l=2\} \Rightarrow \text{control}
• Consider *each* marker separately

• Combine them to get a statistic

• $SSW_k = \text{Within class variation}$

• $SSB_k = \text{Between class variation}$
\[ y_{lij} = \mu + \alpha_l + e_{lij} \quad i < j = 1,2,\ldots,n_l; \quad l = 1,2 \]

\[ SSB_k / SSW_k : \text{ for a single marker} \]

Test statistic:

\[ \mathcal{I} = \frac{\sum_{k=1}^{K} SSB_k}{\sum_{k=1}^{K} SSW_k} \]

- If observed \( \mathcal{I} \) is small we can think that \( H_0 \) is true
- If observed \( \mathcal{I} \) is large we can think that \( H_0 \) is not true

\[ P(\mathcal{I} > \mathcal{I}_\gamma | H_0) = \gamma = P(\text{Type I error}) \]

\[ P - \text{value} = P(\mathcal{I} > \text{Obsd.}\mathcal{I} | H_0) \]
\[ P(\mathcal{S} > \mathcal{S}_\gamma | H_0) = \gamma = P(\text{Type I error}) \]

\[ Power = P(\mathcal{S} > \mathcal{S}_\gamma | H_1) \]

- The test is one-sided to the right

- The distribution of the test statistic is not known

- We calculate \textbf{Power} by simulation/permutation
Simulation

- Genotypes of 10 independent markers

- Number of markers associated with disease ranges from 1 to 5
• High-risk allele frequency is 0.05

• Relative risk is 1.5 and assume multiplicative model

• Sample size for each group is 500

• $\Theta_\gamma$ is calculated based on 10000 simulations

• Power is calculated based on 1000 simulations
POWER STUDY

Additive model: RR=1.25, MAF=0.05

Additive model: RR=1.5, MAF=0.05

Multiplicative model: RR=1.25, MAF=0.05

Multiplicative model: RR=1.5, MAF=0.05

KBAT (Mukhopadhyay et al (2010))

Zglobal (Schaid et al (2005))

MDMR (Wessel & Schork (2006))

MDMR+ (Modified MDMR)

Other competitive tests
Asymptotic distribution of KBAT statistic

\[
T = \beta(n_1, n_2) \frac{K(1 + v^2)}{2v(1 + v)} \left( \sum_{k=1}^{K} \frac{SSB_k}{\hat{\sigma}_{1k}^2} \right) \rightarrow \chi_k^2 \quad \text{as } (n_1, n_2) \rightarrow \infty
\]

where \( \beta(n_1, n_2) = \frac{n_1(n_1 - 1) + n_2(n_2 - 1)}{2n_1} \)
Family based KBAT
Notations

- SNP marker: $aa, ab, bb$
- No. of markers in a gene: $L$
- Phenotype: qualitative – affected or unaffected
- Nuclear families with at least one affected sib
- No. of families: $n$
Towards test statistic...

Consider $l$-th locus, $r$-th family

\[ h_r(g_{P1}^l, g_{P2}^l) \]

\[ \frac{1}{2n_r} \sum_{j=1}^{n_r} h_r(g_{P1}^l, g_{Sj}^l) \]

\[ + \frac{1}{2n_r} \sum_{j=1}^{n_r} h_r(g_{P2}^l, g_{Sj}^l) \]

\[ \frac{2}{n_r(n_r-1)} \sum_{i<j} h_r(g_{Si}^l, g_{Sj}^l) \]
Towards test statistic…

• Propose a 3-dimensional statistic using three statistics:

\[
U_{rl} = \sum_{rl}^{-\frac{1}{2}} (T_{rl} - \mu_l)
\]

where \( T_{rl} = (T_{1,rl}, T_{2,rl}, T_{3,rl})' \) and \( \Sigma_{rl} \) is the var-cov matrix of \( T_{rl} \); \( r = 1, ..., n; \ l = 1, ..., L \).

• Combine genetic information from \( L \) loci at a time for all \( n \) families to get the final statistic:
**Kernel based association test for family data**

**F-KBAT:**

\[ U_n = \hat{U}_n \hat{U}_n' \]

**Theorem:** Let \( \hat{U}_n \) be the mean of all estimated scaled score vectors \( \hat{U}_{rl} \) over all families and for all \( l \), replace \( \mu_l \) and \( \Sigma_{rl} \) by their consistent estimators. Assume \( \forall r \forall l \), \( j = (1,1,1)' \), \( \| \Sigma_{rl}^{-\frac{1}{2}} j \| \leq M < \infty \). Then under \( H_0 \) (no assoc.),

\[ \tilde{U}_n \tilde{U}_n' \overset{d}{\longrightarrow} \chi^2_3 \text{ as } n \to \infty. \]
Simulation

- 10 SNPs; causal markers $k=1,2,3,4,5$
- MAF = $0.1+i/100$, $i=1,2,\ldots,10$
- Genetic model: recessive, dominant
- No. of sibs per family ($X$) $\sim$ Poisson$(3|X>1)$
- $n = 200$ families
- Average p-value over 1000 simulations
- Disease model:
  - Model 1: affected if at least one of $k$ causal loci has risk genotype
  - Model 2: affected if all $k$ causal loci have risk genotypes
Power against no. of causal loci

- Allele match kernel & Recessive model
- Allele match kernel & Dominant model
Qt-KBAT
QT-KBAT: using quantitative trait

- Phenotype similarity $\iff$ Genotype similarity
- People who have similar phenotype trait values should have higher sharing of genetic material near the genes that influence those traits

We are different!!

We are same!

But are we genetically same (with respect to trait)??
MODEL

Phenotype similarity: \[ P_{ij} = |z_i - z_j| \]

Genotype similarity: 3 possible groups based on 3 possible similarity values

\[ G_1 = \{(g_i, g_i) : g_i = a/a, a/b \text{ & } b/b\} \]

\[ G_2 = \left\{ (g_i, g_j) : [g_i = a/a \text{ & } g_j = a/b] \right. \]
\[ \hspace{1cm} \text{or } [g_i = a/b \text{ & } g_j = b/b] \left. \right\} \]

\[ G_3 = \{(g_i, g_j) : g_i = a/a \text{ and } g_j = b/b\} \]

Total Number of markers: K
Model

\[ P_{l(ij)} = \mu + \beta_l G_{l(ij)} + e_{l(ij)}; \ i < j = 1, \ldots, n; \ l = 1, \ldots, K \]

(i) \( V(e_{l(ij)}) = \sigma^2 \)
(ii) Errors \( (e_{l(ij)}) \) are correlated
(iii) Errors are not Normally distributed

Test Statistic

\[ \mathcal{Z} = \sum_{l=1}^{K} \mathcal{Z}_l \text{ where } \mathcal{Z}_l = \frac{SSE_{\beta_l=0} - SSE}{SSE} \]
Asymptotic distribution of Qt-KBAT statistic

\[ \Im = \sum_{k=1}^{K} \Im_k \xrightarrow{L} \sum_{k=1}^{K} w_k \chi_1^2 \text{ as } n \to \infty \]
Conclusion, Future & ongoing works

• Our method is generally more powerful
• Significance may be determined by permutation
• Asymptotic distn helps in computing p-value fast
• Choice / effects of kernels and models
• Asymptotic distn when markers are not independent
Conclusion, Future & ongoing works

• KBAT for case-control data & Qt-KBAT for quantitative phenotype
• KBAT for family data
• Develop gene-gene interaction test
• Develop gene-environment test
• Asymptotic distns in all above cases …