Association Mapping in Families

Gonçalo Abecasis
University of Oxford
Linkage Analysis

- Sharing between relatives
  - Identifies large regions
    - Include several candidates

- Complex disease
  - Scans on sets of small families popular
  - No strong assumptions about disease alleles
  - Low power
  - Limited resolution
Association Analysis

- Sharing between *unrelated* individuals
  - Trait alleles originate in common ancestor
  - High resolution
    - Recombination since common ancestor
    - Large number of independent tests
- Powerful if assumptions are met
  - Same disease haplotype shared by many patients
- Sensitive to population structure
Stratification vs Disequilibrium

Population A

Population B

Ancestor

Present-day
Disequilibrium Mapping

- Control for possible population structure
  - Distinguish linkage disequilibrium from other types of association
- Family-based association analysis
  - Using families collected for linkage mapping
- Powerful if assumptions are met
  - Same disease haplotype shared by many patients
- High-resolution
### Essential Notation

- **$i$**: families
- **$j = 1 .. n_i$**: offspring in family $i$
- **$y_{ij}$**: quantitative phenotype
- **$g_{ij}$**: no. of ‘1’ alleles at marker
- **$g_iF, g_iM$**: parental genotypes, optional
- **$\pi_{ijk}$**: IBD between $j$ and $k$ in family $i$
- **$\varphi_{ijk}$**: kinship between $j$ and $k$ in family $i$
Controlling for Stratification

- If stratum were known...
  - For each individual genotype \( g_{ij} \)
  - Average number of alleles in a strata \( b_{ij} \)
  - Adjust for stratum differences \( w_{ij} = g_{ij} - b_{ij} \)

\[
\hat{y}_{ij} = \mu + \hat{\beta}_b b_{ij} + \hat{\beta}_w w_{ij}
\]

- How to define stratum then?
  - Use family data to estimate \( b_{ij} \)
b_{ij} as Family Control

- Expected genotype for each individual
  - Ancestors
  - Siblings
- Informative individuals
  - Genotype may differ from expected
  - Have heterozygous ancestor in pedigree
Allowable Family Structures
Nuclear Families

\[ b_{ij} = b_i = \begin{cases} \frac{g_{iF} + g_{iM}}{2} & \text{average of parental genotypes} \\ \sum_{k}^{\text{ship}} \frac{g_{ik}}{n_{sibs}} & \text{average of sibling genotypes} \end{cases} \]

\[ w_{ij} = g_{ij} - b_{ij} \]
Extended Families

\[ b_{ij} = \begin{cases} 
\frac{b_{iF_j} + b_{iM_j}}{2} & \text{average of parental controls} \\
\sum_{k} \frac{g_{ik}}{n_{sibs}} & \text{average of sibling genotypes} \\
g_{ij} & \text{self-genotype} \\
\text{undefined} & \text{otherwise}
\end{cases} \]
Allowing for Related Data

- Similarities between individuals
  - Variance–covariance matrix
- Major gene, polygenic, environment

\[
\Omega_{ijk} = \begin{cases} 
\sigma_a^2 + \sigma_g^2 + \sigma_e^2 & \text{if } j = k \\
\pi_{ijk} \sigma_a^2 + 2\varphi_{ijk} \sigma_g^2 & \text{if } j \neq k
\end{cases}
\]
Likelihood function

- Multivariate Normal Distribution
- Defines asymptotic significance levels

\[ L = \prod_i (2\pi)^{-n_i/2} \left| \hat{\Omega}_i \right|^{-1/2} e^{-1/2[(y_i - \hat{y}_i)'\hat{\Omega}_i^{-1}(y_i - \hat{y}_i)]} \]

\[ \chi^2 = 2 \ln \left( \frac{L_{\text{full model}}}{L_{\text{sub-model}}} \right) \]
Parameter Derivations

Model = (\mu, \beta_b, \beta_w, \sigma_e^2, \sigma_g^2, \sigma_a^2)

\[
\begin{bmatrix}
\beta_b \\
\beta_w
\end{bmatrix} = \frac{\sum_i n_i (p_i - q_i) \mu_i}{NV_b} + a
\]

\[a = \frac{D}{pq} a_{QTL}\]

\[\sigma_a^2 = V_{QTL} - 2 pq a\]

"I think you should be more explicit here in step two."

© 1998 Sidney Harris
Exact Permutation Test

- In each family, \( w_i = [w_{i1}, w_{i2}, \ldots] \) is the pattern of allelic transmission
  - \( w_i \) and \(-w_i\) are equally likely (\( H_0 \))
- Null distribution of the data
  - Randomly permute any set families by replacing each \( w_i \) with itself or \(-w_i\) with equal probability
  - The permuted data sets define the null distribution of the maximum likelihood statistic
- Empirical significance levels
Application: Angiotensin-1

- British population
- Circulating ACE levels
  - Normalized separately for males / females
- 10 di-allelic polymorphisms
  - 26 kb
  - Common
  - In strong linkage disequilibrium
- Keavney et al, HMG, 1998
Haplotype Analysis

- 3 clades
  - All common haplotypes
  - >90% of all haplotypes
- “B” = “C”
  - Equal phenotypic effect
  - Functional variant on right
Evidence for Linkage

\[ H_0 : (\mu, \sigma_g^2, \sigma_e^2) \quad H_1 : (\mu, \sigma_g^2, \sigma_e^2, \sigma_a^2) \]
Evidence for Association

\[ H_0 : (\mu, \sigma^2_g, \sigma^2_a, \sigma^2_e, \beta_b) \]

\[ H_1 : (\mu, \sigma^2_g, \sigma^2_a, \sigma^2_e, \beta_b, \beta_w) \]
Evidence Against Complete LD

\[ H_0 : (\mu, \beta_b, \beta_w, \sigma_g^2, \sigma_e^2) \quad H_1 : (\mu, \beta_b, \beta_w, \sigma_g^2, \sigma_e^2, \sigma_a^2) \]
Drawing Conclusions

- for Linkage
- for Association
- against Complete LD
Parameter Estimates

- Estimates
  - Total linkage ($\sigma^2_a$)
  - Linkage due to LD ($\sigma^2_{a*LD}$)
  - Effect size at marker ($\beta_w$)

- Depend on
  - QTL allele frequencies ($p,q$)
  - QTL effect (a)
  - Disequilibrium (D)
  - Marker allele frequencies ($r,s$)
Useful diagnostics

- A bit of algebra
- Provide indicator of distance
  - Minimum $D'$ ($D'_{\text{min}}$)
- Select next markers
  - Range for QTL alleles ($p_{\text{min}}$, $p_{\text{max}}$)

\[
D'_{\text{min}} = \sqrt{\frac{\sigma_{a-LD}^2}{\sigma_a^2}}
\]

\[
p_{\text{min}} = \frac{1}{1 + \frac{\sigma_a^2}{(2\beta_w^2r^2)}}
\]

\[
p_{\text{max}} = 1 - \frac{1}{1 + \frac{\sigma_a^2}{(2\beta_w^2s^2)}}
\]
Application to T-5991C

- LOD scores
  - ~7 linkage
  - ~9 association
  - ~1 linkage minus association

- Trait locus predictions
  - In greater than 78% disequilibrium
  - Minor allele frequency between .15 and .48

- Compare to I/D and neighbors
ACE: $D_{\text{min}}'$, $p_{\text{min}}$ and $p_{\text{max}}$

<table>
<thead>
<tr>
<th>Minor allele</th>
<th>Expected</th>
<th>Actual</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>T-5991C</td>
<td>G2215A</td>
</tr>
<tr>
<td>$D'$</td>
<td>&gt; 0.78</td>
<td>0.78</td>
</tr>
<tr>
<td></td>
<td>.15–.48</td>
<td>.45–.50</td>
</tr>
</tbody>
</table>
Finer mapping

- 3 mutations (inc. I/D) explain all linkage
  - UK population
- How to identify causal variant?
  - Population with more haplotype diversity
  - Jamaican sample
    - Colin McKenzie, University of West Indies, Jamaica
    - Same di-allelic polymorphisms
Jamaican Population Summary

LOD for Linkage, for Association, against Complete LD

- T-5991C
- T-3892C
- T-93C
- G2215A
- G2350A
- A-5466C
- A-240T
- T1237C
- I/D
- 4656(CT)3/2
Example Summary

- Agrees with haplotype analysis
- Distinguishes complete and incomplete disequilibrium
  - Measure of distance for incomplete LD
  - Indicator of trait allele frequencies
- Typical or fairy-tale?
Study design

- What markers?
  - Effect of disequilibrium
  - Effect of allele frequencies

- Family sample
  - What size families?
  - Parents or no parents?

- Effect of phenotypic selection
Sensitivity to Disequilibrium

<table>
<thead>
<tr>
<th></th>
<th>0%</th>
<th>25%</th>
<th>50%</th>
<th>75%</th>
<th>100%</th>
</tr>
</thead>
<tbody>
<tr>
<td>480 triads</td>
<td>0</td>
<td>2</td>
<td>20</td>
<td>70</td>
<td>97</td>
</tr>
<tr>
<td>240 sib-pairs</td>
<td>0</td>
<td>2</td>
<td>23</td>
<td>73</td>
<td>98</td>
</tr>
<tr>
<td>120 sib-quads</td>
<td>0</td>
<td>3</td>
<td>27</td>
<td>76</td>
<td>98</td>
</tr>
<tr>
<td>Estimate of $a$</td>
<td>0</td>
<td>1.1</td>
<td>2.2</td>
<td>3.4</td>
<td>4.5</td>
</tr>
</tbody>
</table>

Power for $\alpha=0.001$, $h^2 = .1$, $s^2 = .3$, $\theta = 0$.

Average additive genetic value estimated at the marker.
Effect of Family Structure

-Log(α) vs. Total Offspring

Legend:
- red: parents
- blue diamonds: sib-pair
- blue triangles: sib-triad
- blue squares: sib-quad
<table>
<thead>
<tr>
<th>Disease Allele Frequency</th>
<th>0.1</th>
<th>0.3</th>
<th>0.5</th>
<th>0.7</th>
<th>0.9</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>248</td>
<td>626</td>
<td>1306</td>
<td>2893</td>
<td>10830</td>
</tr>
<tr>
<td>0.3</td>
<td>1018</td>
<td>238</td>
<td>466</td>
<td>996</td>
<td>3651</td>
</tr>
<tr>
<td>0.5</td>
<td>2874</td>
<td>702</td>
<td>267</td>
<td>556</td>
<td>2002</td>
</tr>
<tr>
<td>0.7</td>
<td>9169</td>
<td>2299</td>
<td>925</td>
<td>337</td>
<td>1187</td>
</tr>
<tr>
<td>0.9</td>
<td>73783</td>
<td>18908</td>
<td>7933</td>
<td>3229</td>
<td>616</td>
</tr>
</tbody>
</table>

\[ \lambda_s = 1.5, \alpha = 5 \times 10^{-8}, \text{Spielman TDT} \]

(Müller-Myhsok and Abel, 1997)
Effect of Allele Frequencies

180 Random Sib-Pairs

Power

QTL allele frequency

Marker allele frequency
Effect of Selection

- Affected Pair
- Extreme Concordant Pair
- Discordant Pair
- Affected + Unselected Sib
- Extreme + Unselected Sib
- Extreme Concordant or Discordant Pair
References

QTDT

Linkage Disequilibrium Analyses for Quantitative Traits

QTDT provides a convenient one-stop interface for family based tests of linkage disequilibrium. The general model described by Abecasis (1999) applies to families with and without parental data, and includes an optional permutation framework for exact p-values. The tests described by Allison (CUTFQ, 1997), Efron's (1997, 1999), and Carter (1998) are also supported.

Background Information
- Brief about family based disequilibrium analysis and QTDT.

Quick-Start Tutorial
- Take a tour of the QTDT package. Recommended.

Browse documentation
- The online documentation includes information on the format and known bugs.

Download programs
- Download the latest version of the program. [Updated February 2000]

Registration form
- If you decide to use QTDT, please take a minute to let us know.

University of Oxford | Wellcome Trust Centre | Asthma Genetics