Linkage in Selected Samples

Boulder Methodology Workshop
2005

Michael C. Neale
Virginia Institute for Psychiatric & Behavioral Genetics
Basic Genetic Model

Pihat = p(IBD=2) + 0.5 p(IBD=1)

Q: QTL Additive Genetic  F: Family Environment  E: Random Environment
3 estimated parameters: q, f and e  Every sibship may have different model
Mixture distribution model

Each sib pair $i$ has different set of WEIGHTS

$rQ = 1$

$rQ = 0.5$

$rQ = 0$

weight$_j$ x Likelihood under model $j$

$p(\text{IBD}=2) \times P(\text{LDL}_1 \& \text{LDL}_2 \mid rQ = 1)\n
p(\text{IBD}=1) \times P(\text{LDL}_1 \& \text{LDL}_2 \mid rQ = 0.5)\n
p(\text{IBD}=0) \times P(\text{LDL}_1 \& \text{LDL}_2 \mid rQ = 0)\n
Total likelihood is sum of weighted likelihoods
QTL's are factors

- Multiple QTL models possible, at different places on genome
- A big QTL will introduce non-normality
  - Introduce mixture of means as well as covariances (27ish component mixture)
- Mixture distribution gets nasty for large sibships
## Biometrical Genetic Model

<table>
<thead>
<tr>
<th>Genotype</th>
<th>Genotype means</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
<td>m + a</td>
</tr>
<tr>
<td>Aa</td>
<td>m + d</td>
</tr>
<tr>
<td>aa</td>
<td>m – a</td>
</tr>
</tbody>
</table>

Diagram:  
- A: 0  
- B: -a  
- C: d  
- D: +a  

Courtesy Pak Sham, Boulder 2003
Mixture of Normal Distributions

Equal variances, Different means and different proportions according to allele frequencies
Implementing the Model

- Estimate QTL allele frequency $p$
- Estimate distance between homozygotes $2a$
- Compute QTL additive genetic variance as $2pq[a+d(q-p)]^2$
- Compute likelihood conditional on
  - IBD status
  - QTL allele configuration of sib pair (IBS)
27 Component Mixture

Sib1

AA

Aa

aa

Sib 2

AA

Aa

aa
19 Possible Component Mixture

Sib1

AA

Aa

aa

Sib 2

AA

Aa

aa
## Results of QTL Simulation

### 3 Component vs 19 Component

<table>
<thead>
<tr>
<th>Parameter</th>
<th>True</th>
<th>3 Component</th>
<th>19 Component</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q</td>
<td>0.4</td>
<td>0.414</td>
<td>0.395</td>
</tr>
<tr>
<td>A</td>
<td>0.08</td>
<td>0.02</td>
<td>0.02</td>
</tr>
<tr>
<td>E</td>
<td>0.6</td>
<td>0.56</td>
<td>0.58</td>
</tr>
<tr>
<td>Test Q=0 (Chisq)</td>
<td>---</td>
<td>13.98</td>
<td>15.88</td>
</tr>
</tbody>
</table>

Information in selected samples

Concordant or discordant sib pairs

• Deviation of pihat from .5
  • Concordant high pairs > .5
  • Concordant low pairs > .5
  • Discordant pairs < .5

• How come?
Pihat deviates > .5 in ASP

Larger proportion of IBD=2 pairs

Sib 2

Sib 1
Pihat deviates < .5 in DSP's

Larger proportion of IBD=0 pairs

Sib 1

Sib 2

IBD=2

IBD=1

IBD=0
Sibship informativeness: sib pairs

unequal allele frequencies

rare recessive
dominance

Courtesy Shaun Purcell, Boulder Workshop 03/03
Two sources of information

Forrest & Feingold 2000

- Phenotypic similarity
  - IBD 2 > IBD 1 > IBD 0
  - Even present in selected samples

- Deviation of pihat from .5
  - Concordant high pairs > .5
  - Concordant low pairs > .5
  - Discordant pairs < .5

- These sources are independent
Implementing F&F

- Simplest form test mean pihat = .5
- Predict amount of pihat deviation
  - Expected pihat for region of sib pair scores
  - Expected pihat for observed scores
- Use multiple groups in Mx
Predicting Expected Pihat deviation

Sib 2

Sib 1

IBD=2

IBD=1

IBD=0
Expected Pihats: Theory

- IBD probability conditional on phenotypic scores $x_1, x_2$

- $E(\text{pihat}) = p(\text{IBD}=2|(x_1, x_2)) + .5p(\text{IBD}=1|(x_1, x_2))$  

\[
p(\text{IBD}=2 |(x_1, x_2)) + p(\text{IBD}=1 |(x_1, x_2)) + p(\text{IBD}=0 |(x_1, x_2))
\]

- $p(\text{IBD}=2|(x_1, x_2)) = \frac{\sum_{IBD=2}(x_1, x_2)}{\sum_{IBD=2}(x_1, x_2) + 2\sum_{IBD=1}(x_1, x_2) + \sum_{IBD=0}(x_1, x_2)}$
Expected Pihats

- Compute Expected Pihats with \texttt{pdfnor}

- \texttt{pdfnor}(X\_M\_C)
  - Observed scores $X$ (row vector 1 x nvar)
  - Means $M$ (row vector)
  - Covariance matrix $C$ (nvar x nvar)
How to measure covariance?
Ascertainment

- Critical to many QTL analyses
- Deliberate
  - Study design
- Accidental
  - Volunteer bias
  - Subjects dying
Exploiting likelihood

- Correction not always necessary
  - ML MCAR/MAR

- Simulate bivariate normal data $X, Y$
  - $\Sigma = 1 \ 0.5$
  - $0.5 \ 1$
  - $\mu = 0, 0$

- Make some variables missing
  - Generate independent random normal variable, $Z$, if $Z > 0$ then $Y$ missing
  - If $X > 0$ then $Y$ missing
  - If $Y > 0$ then $Y$ missing

- Estimate elements of $\Sigma$ & $\mu$
- Constrain elements to population values 1, 0.5, 0 etc
- Compare fit
- Ideally, repeat multiple times and see if expected 'null' distribution emerges
# Results of simulation

Population covariance 1 .5 1 Means 0, 0

<table>
<thead>
<tr>
<th>Missingness</th>
<th>mean x</th>
<th>mean y</th>
<th>var x</th>
<th>cov xy</th>
<th>var y</th>
<th>LR Chisq</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>MCAR (rand)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MLE</td>
<td>-0.0116</td>
<td>-0.1</td>
<td>1.0505</td>
<td>0.4998</td>
<td>0.8769</td>
<td>6.492</td>
</tr>
<tr>
<td>sample</td>
<td>-0.0116</td>
<td>-0.0919</td>
<td>1.0505</td>
<td></td>
<td>0.8839</td>
<td></td>
</tr>
<tr>
<td><strong>MAR (on x)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MLE</td>
<td>0.0048</td>
<td>0.0998</td>
<td>1.0084</td>
<td>0.4481</td>
<td>1.1025</td>
<td>5.768</td>
</tr>
<tr>
<td>sample</td>
<td>0.0014</td>
<td>0.4437</td>
<td>1.0084</td>
<td></td>
<td>0.9762</td>
<td></td>
</tr>
<tr>
<td><strong>NMAR (on y)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MLE</td>
<td>-0.0204</td>
<td>0.6805</td>
<td>0.9996</td>
<td>0.1356</td>
<td>0.2894</td>
<td>227.262</td>
</tr>
<tr>
<td>sample</td>
<td>0.0448</td>
<td>0.7373</td>
<td>0.9996</td>
<td>0.2851</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Weighted likelihood approach

- Usual nice properties of ML remain
- Flexible
- Simple principle
  - Consideration of possible outcomes
  - Re-normalization
- May be difficult to compute
Example: Two Coin Toss

3 outcomes

Outcome

HH
HT/TH
TT

Probability $i = \frac{\text{freq } i}{\text{sum (freqs)}}$
Example: Two Coin Toss

3 outcomes

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>HH</td>
<td>1</td>
</tr>
<tr>
<td>HT/TH</td>
<td>2</td>
</tr>
<tr>
<td>TT</td>
<td>1</td>
</tr>
</tbody>
</table>

Probability $i = \frac{\text{freq} \ i}{\sum \text{(freqs)}}$
Non-random ascertainment

Example

- Probability of observing TT globally
  - 1 outcome from 4 = 1/4

- Probability of observing TT if HH is not ascertained
  - 1 outcome from 3 = 1/3
  - or 1/4 divided by ‘Ascertainment Correction' of 3/4 = 1/3
Correcting for ascertainment

Univariate case; only subjects > t ascertained

![Graph showing the correction for ascertainment with a normal distribution curve and a threshold at t.](image)
Ascertainment Correction

- Be / All you can be

\[ \text{tx} \int (x) \, dx \]
Affected Sib Pairs

\[(x,y) dy dx\]
Ascertainment Corrections for Sib Pairs

ASP ++  
\[ (x,y) \, dy \, dx \]

DSP +-  
\[ (x,y) \, dy \, dx \]

CUSP +-  
\[ (x,y) \, dy \, dx \]
Correcting for ascertainment

• Multivariate selection: multiple integrals
  • double integral for ASP
  • four double integrals for EDAC

• Use (or extend) weight formula

• Precompute in a calculation group
  • unless they vary by subject
Initial Results of Simulations

- **Null Model**
  - 50% heritability
  - No QTL
  - Used to generate null distribution
  - 0.05 empirical significance level at approximately 91 Chi-square

- **QTL Simulations**
  - 37.5% heritability
  - 12.5% QTL
  - Mx: 879 significant at nominal .05 p-value
  - Merlin: 556 significant at nominal .05 p-value
  - Some apparent increase in power
Measurement is KEY

Need continuous interval scales

Most complex traits not measured this way

Use latent trait instead

Factor model equivalent to Item response theory model

Can allow for non-normal Factors

Measurement of multiple Sx
Conclusion

- Quantifying QTL variation in selected samples can be done.
- Can be computationally challenging.
- May provide more power.
- Permits multivariate analysis of correlated traits.