Some newly developed analysis methods motivated by a genetics study

國家衛生研究院
生物統計與生物資訊研究組
組主任 熊昭
I. Motivation — An International Genetic Study
   SAPPHIRE

II. Improving the Power of Linkage Analysis in Sibpair Design

III. Distribution of Crossover Points
遺傳研究 - SAPPHIRe

Stanford Asian Pacific Program in Hypertension and Insulin Resistance
Objectives

- To map and identify major genetic loci underlying hypertension in Chinese and Japanese
- To study interactions between genetic and non-genetic determinants of hypertension in defined populations
Study Population

- Hawaii, San Francisco Bay Area, Taiwan
- 1318 sib pairs who are Japanese or Chinese, 35 to 60 years old and who both have hypertension or one is hypertensive and one is hypotensive
- Pedigrees with hypertension
Measured Variables

- Demographic: age, gender, ethnicity, education
- Anthropometry: height, weight, hip and waist circumference
- Clinical/Biochemical: SBP, DBP, HDL, triglycerides, total cholesterol, OGTT, medications
- Behavioral: physical activity, smoking, alcohol intake
- Genetic: 390 evenly spaced microsatellite markers for linkage analysis and candidate genes for association studies (this comes to ~600,000 genotypes for this study)
Candidate genes approach
Studying SNP genotyping data for promising spots

Genome-wide scan approach
Improving the power of linkage analysis in a sibpair design
Study Design

♦ The study design is based on “ascertaining sibpairs with extremely discordant or highly concordant phenotypes to obtain greater power to detect linkage” (Risch & Zhang 1995, 96)

♦ We suggest a systematic way to decide how to select discordant sibpairs or concordant sibpairs in linkage analysis based on their phenotype trait values.
Sib-pair Design

Identity by Descent (IBD):
DNA at the same locus on 2 homologous chromosomes is said to be IBD if it originated from the same ancestral chromosome.

♦ Inheritance vectors of a sib-pair at the locus $L$:

$\chi(L) = \chi = (\chi_1, \chi_2, \chi_3, \chi_4)$

$\chi_1$ = label of paternal chromosome ($\chi_3$) from which sib 1 (2) inherited DNA at $L$
$= 1$ or $2$

$\chi_2$ = label of maternal chromosome ($\chi_4$) from which sib 1 (2) inherited DNA at $L$
$= 3$ or $4$

Father                  Mother

$\chi_1$ $\parallel$ $\chi_2$  $\parallel$ $\chi_4$

$\chi_3$ $\parallel$ $\chi_4$
### Sib-pair IBD

<table>
<thead>
<tr>
<th>NO. of IBD</th>
<th>(Prob.)</th>
<th>Inheritance vectors</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1/4</td>
<td>(1,3,2,4) , (2,3,1,4) , (1,4,2,3) , (2,4,1,3)</td>
</tr>
<tr>
<td>1</td>
<td>1/2</td>
<td>(1,3,1,4) , (2,3,2,4) , (1,4,1,3) , (2,4,2,3) , (1,3,2,3) , (2,3,1,3) , (1,4,2,4) , (2,4,1,4)</td>
</tr>
<tr>
<td>2</td>
<td>1/4</td>
<td>(1,3,1,3) , (1,4,1,4) , (2,3,2,3) , (2,4,2,4)</td>
</tr>
</tbody>
</table>

Expectation = \( 0 \times \frac{1}{4} + 1 \times \frac{1}{2} + 2 \times \frac{1}{4} = 1 \)

When IBD information is incomplete, partial information may be summarized by the conditional distribution of the possible inheritance vectors at the marker loci.
QTL: Quantitative Trait Loci

- Examine phenotypes conditional on IBD to detect linkage to loci influencing complex trait.

- Unify the linkage analysis of qualitative and quantitative traits by considering IBD data conditional on the phenotypes of the sibpairs.
  Deduoit & Speed (2000): parametric approach
  Chang, Chen, Hsiao & Hsiung (2002): Non-parametric approach
Is there linkage between a marker locus and a gene influencing the trait of interest?

Inference based on IBD Process and Phenotype

. IBD process on a particular chromosome region \( \{ 0, 1 \} \)

\[
I(t) = \sum_{i=1}^{n+1} X_i 1_{[r_{i-1}, r_i]}(t)
\]

. Phenotype of a sibpair

\( \phi \in \mathbb{R}^2 \)

. For \( k^{th} \) sibpair,

\( (I_k, \phi_k) \)
Test hypotheses:

\( H_0 : \) there is no region of enriched or diluted IBD on the chromosome (no linkage)

\( H_1 : \) there is linkage

Under \( H_0 : \quad E \left( I(t) \phi \right) = 1 \)

\[
G_K(t) = \frac{1}{\sqrt{K}} \sum_{k=1}^{K} \sum_{j=1}^{J} w_j(\phi_k) \left( I_k(t) - 1 \right)^H_0 \rightarrow \text{Mean 0 Gaussian}
\]

If \( \sup_t |G_k| \) is bigger than a threshold, then there is linkage on the chromosome.
Select sibpairs based on their trait values and analyze only those pairs having the most informative phenotype.
In reality, IBD process is hardly observable.

- Genotype data are available at each locus in a dense set of markers
- The distribution of IBD can be derived from the distribution of crossover process.
Multipoint Allele-Sharing Method for Sib-pairs

\[
\frac{1}{\sqrt{K}} \sum_{k=1}^{K} \sum_{j=1}^{J} w_j(\phi_k) \left[ E \left( I_k(t) \mid M_k \right) - 1 \right] \sim N(0, \cdot)
\]

The conditional IBD distribution at a locus given the genotypes at a set of markers if no linkage.

\[M_k\] : genotype data taken in a set of markers.

\[I_k(\cdot)\] : IBD process.

\[\phi_k\] : phenotype values of the sibpair.
• Choose $W$ to increase the power of the test.

• If $\sup_t |G_k|$ is bigger than a threshold, then there is linkage on the chromosome.

• We also provide an estimate of the location of a gene influencing the trait.
Advantages of the Method

♦ The method can provide an estimate of the location of a gene influencing the trait if it is on the chromosome.

♦ One can compare the relative informativeness of extremely discordant sib-pairs and highly concordant sib-pairs by studying the variance of the respective estimates.

♦ One can also study linear combination of these two groups to gain more efficiency.
Assumptions Revisit

• The IBD process was assumed to be stationary, Markov chain with independent increment; neglecting effects of crossover interference.

• Strong positive crossover interference, and sex-specific variation in recombination were found. (Broman, K.W., Weber, J.L. (2000), etc)

• The distribution of IBD process can be obtained from the distribution of crossover points.
• Our approach does not use any genetic map function (such as Haldane map function, etc.) and assumes no knowledge of the allele frequency at markers.

• We use count location point process model in which the distribution of number of crossover points and the distribution of the location of crossover points are estimated.
Example: Distribution of Crossover Points in Chromosome 19
choose \( \alpha = 20 \) centiMorgan so that \( P(N_1(D) \geq 2) \leq 0.02 \) if \( D \) has genetic length < 20 cM.

- 229 markers for chromosome 19
- For male 57 markers
- For female 57 markers

In chromosome 19: consider 3 regions divided by 4 markers.

<table>
<thead>
<tr>
<th>Marker</th>
<th>AFMa310wd9</th>
<th>AFM256yc9</th>
<th>AFMa283yh1</th>
<th>AFMa351xe5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>0.00 cM.</td>
<td>41.63 cM.</td>
<td>54.46 cM.</td>
<td>96.04 cM.</td>
</tr>
<tr>
<td>Female</td>
<td>0.00 cM.</td>
<td>43.34 cM.</td>
<td>96.87 cM.</td>
<td>114.28 cM.</td>
</tr>
</tbody>
</table>

The probability of having more than 2 crossover points is negligible.
## Distribution of number of crossover points

<table>
<thead>
<tr>
<th></th>
<th>$B^{(1)}$</th>
<th>$B^{(2)}$</th>
<th>$B^{(3)}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.5849</td>
<td>0.8948</td>
<td>0.5250</td>
</tr>
<tr>
<td>1</td>
<td>0.4151</td>
<td>0.1052</td>
<td>0.4750</td>
</tr>
<tr>
<td>2</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

Probability of having exactly, 0, 1 or 2 crossover points in each of $B^{(1)}$, $B^{(2)}$ and $B^{(3)}$ during a male meiosis.

<table>
<thead>
<tr>
<th></th>
<th>$B^{(1)}$</th>
<th>$B^{(2)}$</th>
<th>$B^{(3)}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.5331</td>
<td>0.4880</td>
<td>0.8494</td>
</tr>
<tr>
<td>1</td>
<td>0.4669</td>
<td>0.5120</td>
<td>0.1506</td>
</tr>
<tr>
<td>2</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

Probability of having exactly, 0, 1 or 2 crossover points in each of $B^{(1)}$, $B^{(2)}$ and $B^{(3)}$ during a female meiosis.
### Distribution of the Crossover Point (Male)

<table>
<thead>
<tr>
<th>$B^{(1)}$</th>
<th>$B^{(2)}$</th>
<th>$B^{(3)}$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>cM.</strong></td>
<td><strong>Cmu. Pr.</strong></td>
<td><strong>cM.</strong></td>
</tr>
<tr>
<td>0.00</td>
<td>0.0000</td>
<td>41.63</td>
</tr>
<tr>
<td>6.51</td>
<td>0.1520</td>
<td>42.04</td>
</tr>
<tr>
<td>8.31</td>
<td>0.2195</td>
<td>42.45</td>
</tr>
<tr>
<td>13.77</td>
<td>0.3553</td>
<td>42.70</td>
</tr>
<tr>
<td>16.14</td>
<td>0.4205</td>
<td>42.76</td>
</tr>
<tr>
<td>19.45</td>
<td>0.4637</td>
<td>43.07</td>
</tr>
<tr>
<td>24.46</td>
<td>0.5891</td>
<td>43.20</td>
</tr>
<tr>
<td>25.53</td>
<td>0.6297</td>
<td>43.35</td>
</tr>
<tr>
<td>29.82</td>
<td>0.7224</td>
<td>43.43</td>
</tr>
<tr>
<td>30.89</td>
<td>0.7533</td>
<td>43.54</td>
</tr>
<tr>
<td>32.55</td>
<td>0.8026</td>
<td>43.63</td>
</tr>
<tr>
<td>34.13</td>
<td>0.8459</td>
<td>43.69</td>
</tr>
<tr>
<td>36.27</td>
<td>0.9077</td>
<td>43.71</td>
</tr>
<tr>
<td>37.33</td>
<td>0.9386</td>
<td>43.76</td>
</tr>
<tr>
<td>38.40</td>
<td>0.9464</td>
<td>44.83</td>
</tr>
<tr>
<td>39.85</td>
<td>0.9700</td>
<td>45.29</td>
</tr>
<tr>
<td>41.63</td>
<td>1.0000</td>
<td><strong>45.89</strong></td>
</tr>
</tbody>
</table>

For male meiosis, relative to the genetic distance from Marshfield (in the odd columns), the even columns give the conditional distribution of the location of the crossover point, given there is one crossover point in each of $B^{(1)}$, $B^{(2)}$ and $B^{(3)}$. 
For female meiosis, relative to the genetic distance from Marshfield (in the odd columns), the even columns give the conditional distribution of the location of the crossover point, given there is one crossover point in each of \( B^{(1)} \), \( B^{(2)} \) and \( B^{(3)} \).

The distribution of crossover point process varies from region to region, and from gender to gender.
There are more (less) crossover points in the middle of the chromosome for female (male) meioses.
\[ E(\ I(t) \ | M) \]

\( I(\cdot) \): Independent Increment

\( (I(\cdot), M(\cdot)) \) is a hidden Markov Model

Kruglyak and Lander (1995)

Genehunter
CEPH – families

884, 1331, 1332, 1347, 1362, 1413, 1416
8000 STRPs (short tandem repeat polymorphisms)
Center for Medical Genetics, Marshfield Medical Research Foundation
10 markers, 6 families
Conclusion

- Linkage analysis for complex disease is a challenging task.
- We propose a method to enhance the power of linkage analysis in sibpair design.
- We also propose a method to estimate the distribution of crossover points, which can take care the problem of positive crossover interference.
- The distribution of crossover point process varies between males and females.
- Drug complications also show gender difference in some clinical trials.