Introduction to Bioinformatics for Medical Research

Gideon Greenspan
gdg@cs.technion.ac.il

Lecture 14
Genetic Mapping
Genetic Mapping

• Background

• Linkage Analysis
  – Pedigrees
  – Probability model

• Association Analysis
  – Single markers
  – Haplotype blocks

• Haplotype Resolution
Why map genes?

• Many diseases are partially genetic
  – Also: environmental factors, randomness

• We want to identify these genes
  – Early diagnosis for abortion or regular checks
  – First step towards developing treatment

• Individual sequencing is too costly (today)
  – Sequence a small number of markers
  – Analyze statistically via biological principles
Meiotic Recombination
Linkage Analysis

• Identify families with disease
  – Many families, many members is best

• Construct pedigree of family members
  – Collect disease status for each individual

• Test alleles at multiple markers
  – Problems: dispersed families, death

• Create models for possible disease locations
  – Find model which best explains data
## Linkage Analysis Example

<table>
<thead>
<tr>
<th>position</th>
<th>LOD_score</th>
<th>information</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00</td>
<td>-1.254417</td>
<td>0.224384</td>
</tr>
<tr>
<td>1.52</td>
<td>2.836135</td>
<td>0.226379</td>
</tr>
</tbody>
</table>

...[other data skipped]...

<table>
<thead>
<tr>
<th>position</th>
<th>LOD_score</th>
<th>information</th>
</tr>
</thead>
<tbody>
<tr>
<td>18.58</td>
<td>13.688599</td>
<td>0.384088</td>
</tr>
<tr>
<td>19.92</td>
<td>14.238474</td>
<td>0.401992</td>
</tr>
<tr>
<td>21.26</td>
<td>14.718037</td>
<td>0.426818</td>
</tr>
<tr>
<td>22.60</td>
<td>15.159389</td>
<td>0.462284</td>
</tr>
<tr>
<td>22.92</td>
<td>15.056713</td>
<td>0.462510</td>
</tr>
<tr>
<td>23.24</td>
<td>14.928614</td>
<td>0.463268</td>
</tr>
<tr>
<td>23.56</td>
<td>14.754848</td>
<td>0.464387</td>
</tr>
</tbody>
</table>

...[other data skipped]...

<table>
<thead>
<tr>
<th>position</th>
<th>LOD_score</th>
<th>information</th>
</tr>
</thead>
<tbody>
<tr>
<td>81.84</td>
<td>1.939215</td>
<td>0.059748</td>
</tr>
<tr>
<td>90.60</td>
<td>-11.930449</td>
<td>0.087869</td>
</tr>
</tbody>
</table>

**Putative distance of disease gene from first marker in centi-Morgans**

**Most ‘likely’ position**

**Log likelihood of placing disease gene at distance, relative to it being unlinked.**

**Maximum log likelihood score**
Founder Probabilities

• Founders assumed to be unrelated
• Each founder marker assigned probability based on population allele frequencies
  – Linkage equilibrium
Inheritance Probabilities

- Markov chain of selector distributions for each marker
  - No interference
- Probability that allele source is different from previous due to recombination = RF
Disease Probabilities

• Disease dependent on allele at unseen locus
  – Many positions tested
• Probability of disease depends on dominance model, liability class and penetrance
Association Analysis

• Also: Linkage Disequilibrium (LD) analysis
• Collect disease cases and healthy controls
  – Small, inbred populations are best (e.g. us)
• Measure alleles at densely spaced markers
  – Single Nucleotide Polymorphisms are ideal
• Test marker–disease correlations
  – How well does each SNP predict disease?
  – P value under assumption of independence
Association Analysis Example

- 3 1 1
- 0 1 2

- 1 3 1
- 1 1 1
False Associations

• Population structure
  – Migration and admixture
  – Preferential mating

• Phenotypic interactions
  – Epistasis between distant sites
  – Selective sweeps

• Many individual tests
  – Raise significance level for multiple tests
Recombination Hotspots
Bottleneck Effects

10^6 years

10^5 years
Haplotype Blocks

<table>
<thead>
<tr>
<th></th>
<th>GAACCTGC</th>
<th>ATTCGACTGC</th>
<th>CCAGTAGC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>ACGTACA</td>
<td>GATGAGCTG</td>
<td>CCAGTAGC</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>99</td>
<td>ACGTACA</td>
<td>AACCAGGCTG</td>
<td>TGTACTAA</td>
</tr>
<tr>
<td>100</td>
<td>GAACCTGC</td>
<td>GATGAGCTG</td>
<td>TGTGCTAA</td>
</tr>
</tbody>
</table>

Recombination hotspot separates blocks

Few block variants due to bottlenecks, drift

Mutation hotspot
HaploBlock Analysis

• Treat haplotype as single unified marker
  – Increased information, fewer tests
• Consider recombination events
  – Consider haplotypes in each block separately
• Consider mutation events
  – Cluster haplotypes into similarity clades
• Test disease correlation with each block
  – P values or prediction ability
Haplotype Resolution

Variable Loci

Maternal Chromosome

Hidden Haplotypes

Paternal Chromosome

Observed Genotypes

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>G</th>
<th>T</th>
<th>C</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>T</td>
<td>C</td>
<td>T</td>
<td>A</td>
<td>G</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>A/T</th>
<th>C/G</th>
<th>T/T</th>
<th>A/C</th>
<th>G/T</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Linkage vs Association

<table>
<thead>
<tr>
<th>Linkage analysis</th>
<th>Association analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Family pedigrees</td>
<td>Unrelated individuals</td>
</tr>
<tr>
<td>Observed recombination</td>
<td>Historical recombination</td>
</tr>
<tr>
<td>Widely spaced markers</td>
<td>Closely spaced SNPs</td>
</tr>
<tr>
<td>Mendelian diseases</td>
<td>Complex diseases</td>
</tr>
<tr>
<td>$10^6$ bp accuracy</td>
<td>$10^4$ bp accuracy</td>
</tr>
<tr>
<td>Many successes</td>
<td>Many false positives</td>
</tr>
</tbody>
</table>
Stages of Mapping a Gene

• Demonstrate disease is hereditary
  – Show it runs in families
• Linkage analysis to identify region
  – Widely-spaced markers, e.g. RFLPs
• Association analysis to narrow region
  – Closely-spaced markers, usually SNPs
• Clone the gene within found region
  – Investigate its metabolic relevance
Resources

• Genetic Analysis Software
  – http://linkage.rockefeller.edu/soft/list.html

• Introduction to Genetic Analysis
  – http://www2.qimr.edu.au/davidD/Course/

• Genetic Analysis Resources

• NCBI Human Genes and Disease

• International Haplotype Map Project
  – http://www.genome.gov/10001688