Statistical Genetics – Part I

Lecture 3: Haplotype reconstruction

Su-In Lee, CSE & GS, UW
suinlee@uw.edu

Outline

- Basic concepts
  - Allele, allele frequencies, genotype frequencies
  - Haplotype, haplotype frequency
  - Recombination rate
  - Linkage disequilibrium

- Haplotype reconstruction
  - Parsimony-based approach
  - EM-based approach

- Next topic
  - Disease association studies
Alleles

- Alternative forms of a particular sequence

- Each allele has a frequency, which is the proportion of chromosomes of that type in the population

```
C, G and -- are alleles

...ACTCGGTGGCCTTAAATCGGC...CCTACCTCGGTGGCCTTAAATCGGC... 
...ACTCGGTGGCCTTAAATCGGC...CGGACTCGGTGGCCTTAAATCGGC... 
...ACCCGGAAGCCCTTAATCGGC...AGCACCCTAATCGGC... 
...ACCCGGAAGCCCTTAATCGGC...--GGACCCTAATCGGC... 
...ACCCGGAAGCCCTTAATCGGC...GACCCTAATCGGC... 
...ACCCGGAAGCCCTTAATCGGC...CACCCTAATCGGC... 
```

single nucleotide polymorphism (SNP) allele frequencies for C, G, --

Allele Frequency Notations

- For two alleles
  - Usually labeled $p$ and $q = 1 - p$
  - e.g. $p$ = frequency of C, $q$ = frequency of G

- For more than 2 alleles
  - Usually labeled $p_A$, $p_B$, $p_C$ ...
  - ... subscripts A, B and C indicate allele names
Genotype

- The pair of alleles carried by an individual
  - If there are \( n \) alternative alleles ...
  - ... there will be \( n(n+1)/2 \) possible genotypes
  - In most cases, there are 3 possible genotypes

- **Homozygotes**
  - The two alleles are in the same state
  - (e.g. CC, GG, AA)

- **Heterozygotes**
  - The two alleles are different
  - (e.g. CG, AC)

Genotype Frequencies

- Since alleles occur in pairs, these are a useful descriptor of genetic data.

- However, in any non-trivial study we might have a lot of frequencies to estimate.

- \( p_{AA}, p_{AB}, p_{AC}, \ldots, p_{BB}, p_{BC}, \ldots, p_{CC}, \ldots \)
The Simple Part

- Genotype frequencies lead to allele frequencies.

- For example, for two alleles:
  - \( p_1 = p_{11} + \frac{1}{2} p_{12} \)
  - \( p_2 = p_{22} + \frac{1}{2} p_{12} \)

- However, the reverse is also possible!
  - We just need an additional assumption

Hardy-Weinberg Equilibrium (HWE)

- Relationship described in 1908
  - Hardy, British mathematician
  - Weinberg, German physician

- Shows \( n \) allele frequencies determine \( n(n+1)/2 \) genotype frequencies
  - Large populations

- Random union of the two gametes produced by two individuals
Random Mating: Mating Type Frequencies

- Denoting the genotype frequency of $A_iA_j$ by $p_{ij}$, and the allele frequency $A_i$ by $p_i$ ($i, j \in \{1, 2\}$),
  - $p_i = p_{1i} + \frac{1}{2} p_{12}$; $p_j = p_{2j} + \frac{1}{2} p_{12}$

<table>
<thead>
<tr>
<th>Mating</th>
<th>Frequency</th>
<th>$A_1A_1$</th>
<th>$A_1A_2$</th>
<th>$A_2A_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A_1A_1^*A_1A_1$</td>
<td>$p_{11}^2$</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$A_1A_1^*A_1A_2$</td>
<td>$2p_{11}p_{12}$</td>
<td>0.5</td>
<td>0.5</td>
<td>0</td>
</tr>
<tr>
<td>$A_1A_1^*A_2A_2$</td>
<td>$2p_{11}p_{22}$</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$A_1A_2^*A_1A_2$</td>
<td>$p_{12}^2$</td>
<td>0.25</td>
<td>0.5</td>
<td>0.25</td>
</tr>
<tr>
<td>$A_1A_2^*A_2A_2$</td>
<td>$2p_{12}p_{22}$</td>
<td>0</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>$A_2A_2^*A_2A_2$</td>
<td>$p_{22}^2$</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Mendelian Segregation: Offspring Genotype Frequencies

<table>
<thead>
<tr>
<th>Mating</th>
<th>Frequency</th>
<th>$A_1A_1$</th>
<th>$A_1A_2$</th>
<th>$A_2A_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A_1A_1^*A_1A_1$</td>
<td>$p_{11}^2$</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$A_1A_1^*A_1A_2$</td>
<td>$2p_{11}p_{12}$</td>
<td>0.5</td>
<td>0.5</td>
<td>0</td>
</tr>
<tr>
<td>$A_1A_1^*A_2A_2$</td>
<td>$2p_{11}p_{22}$</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$A_1A_2^*A_1A_2$</td>
<td>$p_{12}^2$</td>
<td>0.25</td>
<td>0.5</td>
<td>0.25</td>
</tr>
<tr>
<td>$A_1A_2^*A_2A_2$</td>
<td>$2p_{12}p_{22}$</td>
<td>0</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>$A_2A_2^*A_2A_2$</td>
<td>$p_{22}^2$</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
### Mendelian Segregation: Offspring Genotype Frequencies

<table>
<thead>
<tr>
<th>Mating</th>
<th>Frequency</th>
<th>Offspring</th>
<th>A1A1</th>
<th>A1A2</th>
<th>A2A2</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1A1*A1A1</td>
<td>p11^2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>A1A1*A1A2</td>
<td>2p11p12</td>
<td>0.5</td>
<td>0.5</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>A1A1*A2A2</td>
<td>2p11p22</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>A1A2*A1A2</td>
<td>p12^2</td>
<td>0.25</td>
<td>0.5</td>
<td>0.25</td>
<td></td>
</tr>
<tr>
<td>A1A2*A2A2</td>
<td>2p12p22</td>
<td>0</td>
<td>0.5</td>
<td>0.5</td>
<td></td>
</tr>
<tr>
<td>A2A2*A2A2</td>
<td>p22^2</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

\[ \text{Offspring} = p_{11}^2 + 2p_{11}(0.5p_{12}) + (0.5p_{12})^2 \]
\[ = (p_{11} + 0.5p_{12})^2 \]
\[ = p_{11}^2 \]

---

### Mendelian Segregation: Offspring Genotype Frequencies

<table>
<thead>
<tr>
<th>Mating</th>
<th>Frequency</th>
<th>Offspring</th>
<th>A1A1</th>
<th>A1A2</th>
<th>A2A2</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1A1*A1A1</td>
<td>p11^2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>A1A1*A1A2</td>
<td>2p11p12</td>
<td>0.5</td>
<td>0.5</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>A1A1*A2A2</td>
<td>2p11p22</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>A1A2*A1A2</td>
<td>p12^2</td>
<td>0.25</td>
<td>0.5</td>
<td>0.25</td>
<td></td>
</tr>
<tr>
<td>A1A2*A2A2</td>
<td>2p12p22</td>
<td>0</td>
<td>0.5</td>
<td>0.5</td>
<td></td>
</tr>
<tr>
<td>A2A2*A2A2</td>
<td>p22^2</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

\[ \text{Offspring} = p_{11}^2 \]
\[ = 2p_{11}p_{22} \]
Mendelian Segregation: Offspring Genotype Frequencies

<table>
<thead>
<tr>
<th>Mating</th>
<th>Frequency</th>
<th>Offspring</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A_1A_1^*A_1A_1$</td>
<td>$p_{11}^2$</td>
<td>$1$</td>
</tr>
<tr>
<td>$A_1A_1^*A_1A_2$</td>
<td>$2p_{11}p_{12}$</td>
<td>$0.5$</td>
</tr>
<tr>
<td>$A_1A_1^*A_2A_2$</td>
<td>$2p_{11}p_{22}$</td>
<td>$0$</td>
</tr>
<tr>
<td>$A_1A_2^*A_1A_2$</td>
<td>$p_{12}^2$</td>
<td>$0.25$</td>
</tr>
<tr>
<td>$A_1A_2^*A_2A_2$</td>
<td>$2p_{12}p_{22}$</td>
<td>$0$</td>
</tr>
<tr>
<td>$A_2A_2^*A_2A_2$</td>
<td>$p_{22}^2$</td>
<td>$0$</td>
</tr>
</tbody>
</table>

$= p_1^2 \quad = 2p_1p_2 \quad = p_2^2$

Frequency of $A_1$ in offspring $= p_1^2 + \frac{1}{2} 2p_1p_2 = p_1(p_1 + p_2) = p_1$

Conclusion: HWE

- Allele frequencies and genotype ratios in a randomly-breeding population *remain constant* from generation to generation.

- Genotype frequencies are function of allele frequencies.
  - Equilibrium reached in one generation
  - Independent of initial genotype frequencies
  - Random mating, etc. required
Review: Genetic Variation

- Single nucleotide polymorphism (SNP)
  - Each variant is called an allele; each allele has a frequency

- Hardy Weinberg equilibrium (HWE)
  - Relationship between allele frequency and genotype frequencies

- How about the relationship between alleles of neighboring SNPs?
  - We need to know about linkage (dis)equilibrium

Let’s consider the history of two neighboring alleles...
History of Two Neighboring Alleles

- Alleles that exist today arose through ancient mutation events...

**Before mutation**

```
A
```

**After mutation**

```
A
```

```
C Mutation
```

History of Two Neighboring Alleles

- One allele arose first, and then the other...

**Before mutation**

```
A
```
```
C
```
```
G
```
```
G
```

**After mutation**

```
A
```
```
C
```
```
G
```
```
G
```
```
C Mutation
```

Haplotype: combination of alleles present in a chromosome
Recombination Can Create More Haplotypes

- No recombination (or 2n recombination events)
  - Without recombination:
    - A
    - C
    - G
    - C
  - With recombination:
    - A
    - C
    - G
    - C

- Recombination
  - Recombinant haplotype
Haplotype

- A combination of alleles present in a chromosome
- Each haplotype has a frequency, which is the proportion of chromosomes of that type in the population

- Consider N binary SNPs in a genomic region
- There are $2^N$ possible haplotypes
  - But in fact, far fewer are seen in human population

More On Haplotype

- What determines haplotype frequencies?
  - Recombination rate ($r$) between neighboring alleles
  - Depends on the population
  - $r$ is different for different regions in genome

- Linkage disequilibrium (LD)
  - Non-random association of alleles at two or more loci, not necessarily on the same chromosome.

- Why do we care about haplotypes or LD?
Useful Roles For Haplotypes

- Linkage disequilibrium studies
  - Summarize genetic variation
  - Learn about population history

- Selecting markers to genotype
  - Identify haplotype tag SNPs

What is genotyping?

- Genome-wide sequencing is still too expensive
- There are sites that are known to vary across individuals (e.g. SNPs)
- “genotyping” means determining the alleles in each SNP for a certain individual.

Exploiting LD – Tag SNPs

- In a typical a few distinct SNPs
- Carefully selected SNPs
  
<table>
<thead>
<tr>
<th>Haplotype 1</th>
<th>Haplotype 2</th>
<th>Haplotype 3</th>
<th>Haplotype 4</th>
<th>Haplotype 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>T T</td>
<td>T T</td>
<td>T T</td>
<td>T T</td>
<td>T T</td>
</tr>
<tr>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>30%</td>
<td>20%</td>
<td>20%</td>
<td>20%</td>
<td>10%</td>
</tr>
</tbody>
</table>

Different alleles of each SNP
Association Studies and LD

- Why is LD important for disease association studies?

- If all polymorphisms were independent at the population level, association studies would have to examine every one of them...

- Linkage disequilibrium makes tightly linked variants strongly correlated producing cost savings for genotyping in association studies

The Problems...

- Haplotypes are hard to measure directly
  - X-chromosome in males
  - Sperm typing
  - Hybrid cell lines
  - Other molecular techniques

- Often, statistical reconstruction required
Goal

- Haplotype reconstruction

![Haplotype reconstruction diagram with genetic markers and DNA sequences]

Typical Genotype Data

- Two alleles for each individual
  - Chromosome origin for each allele is unknown

- Multiple haplotype pairs can fit observed genotype
Use Information on Relatives?

- Family information can help determine phase at many markers

- Still, many ambiguities might not be resolved
  - Problem more serious with larger numbers of markers

- Can you propose examples?

Example – Inferring Haplotypes

- Genotype: AT//AA//CG
  - Maternal genotype: TA//AA//CC  →  TAC/AAC
  - Paternal genotype: TT//AA//CG  →  TAC/TAG
  - Then the haplotype is AAC/TAG

- Genotype: AT//AA//CG
  - Maternal genotype: AT//AA//CG
  - Paternal genotype: AT//AA//CG
  - Cannot determine unique haplotype

- Problem
  - Determine Haplotypes without parental genotypes
What If There Are No Relatives?

- Rely on linkage disequilibrium
- Assume that population consists of small number of distinct haplotypes

Haplotype Reconstruction

- Also called, *phasing, haplotype inference* or *haplotyping*

Data
- Genotypes on N markers from M individuals

Goals
- Frequency estimation of all possible haplotypes
- Haplotype reconstruction for individuals
- How many out of all possible haplotypes are plausible in a population?
Clark’s Haplotyping Algorithm


- One of the first haplotyping algorithms
  - Computationally efficient
  - Very fast and widely used in 1990’s
  - More accurate methods are now available

---

Clark’s Haplotyping Algorithm

- Find unambiguous individuals
  - What kinds of genotypes will these have?
  - Initialize a list of known haplotypes

  - Unambiguous individuals
    - Homozygous at every locus (e.g. TT//AA//CC)
      Haplotypes: TAC
    - Heterozygous at just one locus (e.g. TT//AA//CG)
      Haplotypes: TAC or TAG
Unambiguous vs. Ambiguous

- Haplotypes for 2 SNPs (alleles: A/a, B/b)

Clark’s Haplotyping Algorithm

- Find unambiguous individuals
  - What kinds of genotypes will these have?
  - Initialize a list of known haplotypes

- Resolve ambiguous individuals
  - If possible, use two haplotypes from list
  - Otherwise, use one known haplotype and augment list

- If unphased individuals remain
  - Assign phase randomly to one individual
  - Augment haplotype list and continue from previous step
**Parsimonious Phasing - Example**

- **Notation (more compact representation)**
  - 0/1: homozygous at each locus (00,11)
  - h: heterozygous at each locus (01)

\[
\begin{array}{c}
10100h \\
h01h00 \\
0h1h0 \\
\end{array}
\quad
\begin{array}{c}
10100 \\
101001 \\
101000 \\
001100 \\
010110 \\
\end{array}
\]

**Notes ...**

- **Clark’s Algorithm is extremely fast**

- **Problems**
  - No homozygotes or single SNP heterozygotes in the sample
  - Many unresolved haplotypes at the end
  - Error in haplotype inference if a crossover of two actual haplotypes is identical to another true haplotype
  - Frequency of these problems depend on average heterozygosity of the SNPs, no of loci, recombination rate, sample size
The EM Haplotyping Algorithm


- Why EM for haplotyping?
  - EM is a method for MLE with hidden variables.

- What are the hidden variables, parameters?
  - **Hidden variables**: haplotype state of each individual
  - **Parameters**: haplotype frequencies

Assume That We Know Haplotype Frequencies

- **Probability of first outcome:**
  - \(2P_{A\bar{b}}P_{\bar{a}B} = 0.06\)

- **Probability of second outcome:**
  - \(2P_{A\bar{b}}P_{\bar{a}b} = 0.18\)
Conditional Probabilities Are ...

For example, if
\[ P_{AB} = 0.3 \]
\[ P_{ab} = 0.3 \]
\[ P_{Ab} = 0.3 \]
\[ P_{aB} = 0.1 \]

- Conditional probability of first outcome:
  \[ \frac{2P_{AB}P_{ab}}{2P_{AB}P_{ab} + 2P_{Ab}P_{aB}} = 0.25 \]

- Conditional probability of second outcome:
  \[ \frac{2P_{AB}P_{ab}}{2P_{AB}P_{ab} + 2P_{Ab}P_{aB}} = 0.75 \]

Assume That We Know The Haplotype State Of Each Individual

- Computing haplotype frequencies is straightforward

Individual 1

\[ \frac{P_{AB}}{2P_{AB} + 2P_{Ab}} = ? \]
\[ \frac{P_{ab}}{2P_{AB} + 2P_{Ab}} = ? \]

Individual 2

\[ \frac{P_{AB}}{2P_{AB} + 2P_{Ab}} = ? \]
\[ \frac{P_{ab}}{2P_{AB} + 2P_{Ab}} = ? \]

Individual 3

\[ \frac{P_{AB}}{2P_{AB} + 2P_{Ab}} = ? \]
\[ \frac{P_{ab}}{2P_{AB} + 2P_{Ab}} = ? \]

Individual 4

\[ \frac{P_{AB}}{2P_{AB} + 2P_{Ab}} = ? \]
\[ \frac{P_{ab}}{2P_{AB} + 2P_{Ab}} = ? \]
Phasing By EM

- EM: Method for maximum-likelihood parameter inference with hidden variables

  ![Diagram showing EM algorithm for phasing]

  - **Parameters** (haplotype frequencies)
    - Maximize Likelihood
  - **Hidden variables** (haplotype states of individuals)
    - Find expected values
  - **Estimating haplotype frequencies**

  **EM Algorithm For Haplotyping**

  1. “Guesstimate” haplotype frequencies
  2. Use current frequency estimates to replace ambiguous genotypes with fractional counts of phased genotypes
  3. Estimate frequency of each haplotype by counting
  4. Repeat steps 2 and 3 until frequencies are stable
Phasing by EM

Data:

\[
\begin{array}{c}
10h1h1 \\
h001h \\
h1h11 \\
\end{array}
\begin{array}{cccccc}
1 & 0 & 0 & 0 & 1 & \frac{1}{4} \\
1 & 0 & 1 & 1 & 1 & \frac{1}{4} \\
1 & 0 & 0 & 1 & 1 & \frac{1}{4} \\
1 & 0 & 1 & 0 & 1 & \frac{1}{4} \\
0 & 0 & 0 & 1 & 0 & \frac{1}{4} \\
1 & 0 & 0 & 1 & 1 & \frac{1}{4} \\
0 & 0 & 0 & 1 & 1 & \frac{1}{4} \\
1 & 0 & 0 & 1 & 0 & \frac{1}{4} \\
1 & 0 & 0 & 1 & 1 & \frac{1}{4} \\
1 & 1 & 1 & 1 & 1 & \frac{1}{4} \\
1 & 0 & 1 & 1 & 1 & \frac{1}{4} \\
1 & 1 & 0 & 1 & 1 & \frac{1}{4} \\
\end{array}
\]

Frequencies

\[
\begin{array}{cccc}
00010 & 1/12 \\
00011 & 1/12 \\
10001 & 1/12 \\
10010 & 1/12 \\
10011 & 3/12 \\
10101 & 1/12 \\
10111 & 2/12 \\
11011 & 1/12 \\
11111 & 1/12 \\
\end{array}
\]
Phasing by EM

Data:

\[
\begin{array}{c|c|c|c|c|c|c}
\text{Haplotypes} & 00010 & 00011 & 10001 & 10010 & 10011 & 10101 \\
\hline
1 0 h h 1 & 0.4 & 0.6 & 0.125 & 0.067 & 0.042 & 0.042 \\
\hline
h 0 0 1 h & 0.75 & 0.25 & 0.325 & 0.75 & 0.25 & 1 \\
\hline
1 h h 1 & 0.6 & 0.4 & 0.67 & 1 & 0.67 & 1 \\
\end{array}
\]

Phasing by EM

\[
\begin{array}{c|c|c|c|c|c|c}
\text{Frequencies} & 00010 & 00011 & 10001 & 10010 & 10011 & 10101 \\
\hline
0 0 0 1 0 & 1/12 & 1/12 & 1/12 & 1/12 & 3/12 & 1/12 \\
0 0 0 1 1 & 1/12 & 1/12 & 1/12 & 1/12 & 2/12 & 1/12 \\
1 0 0 1 0 & 1/12 & 1/12 & 1/12 & 1/12 & 1/12 & 1/12 \\
1 0 0 1 1 & 1/12 & 1/12 & 1/12 & 1/12 & 1/12 & 1/12 \\
1 0 1 0 1 & 1/12 & 1/12 & 1/12 & 1/12 & 1/12 & 1/12 \\
1 0 1 1 1 & 1/12 & 1/12 & 1/12 & 1/12 & 1/12 & 1/12 \\
1 1 0 1 1 & 1/12 & 1/12 & 1/12 & 1/12 & 1/12 & 1/12 \\
1 1 1 1 1 & 1/12 & 1/12 & 1/12 & 1/12 & 1/12 & 1/12 \\
\end{array}
\]
Phasing by EM

Data:

<table>
<thead>
<tr>
<th>Haplotypes</th>
<th>Frequencies</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\frac{1}{4}$</td>
<td>1/6</td>
</tr>
<tr>
<td>$\frac{1}{4}$</td>
<td>0</td>
</tr>
<tr>
<td>$\frac{1}{4}$</td>
<td>0</td>
</tr>
<tr>
<td>$\frac{1}{4}$</td>
<td>0</td>
</tr>
<tr>
<td>$\frac{1}{4}$</td>
<td>1/2</td>
</tr>
<tr>
<td>$\frac{1}{4}$</td>
<td>0</td>
</tr>
<tr>
<td>$\frac{1}{4}$</td>
<td>1/6</td>
</tr>
<tr>
<td>$\frac{1}{4}$</td>
<td>0</td>
</tr>
<tr>
<td>$\frac{1}{4}$</td>
<td>0</td>
</tr>
<tr>
<td>$\frac{1}{4}$</td>
<td>1/6</td>
</tr>
</tbody>
</table>

Computational Cost (for SNPs)

- Consider sets of $m$ unphased genotypes
  - Markers 1..$m$

- If markers are bi-allelic
  - $2^m$ possible haplotypes
  - $2^{m-1} (2^m + 1)$ possible haplotype pairs
  - $3^m$ distinct observed genotypes
  - $2^{n-1}$ reconstructions for $n$ heterozygous loci

For example, if $m=10$

- $2^m = 1024$
- $2^{m-1} (2^m + 1) = 524,800$
- $3^m = 59,049$
- $2^{n-1} = 512$
EM Algorithm For Haplotyping

- Cost grows rapidly with number of markers

- Typically appropriate for < 25 SNPs
  - Fewer microsatellites

- More accurate than Clark’s method

- Fully or partially phased individuals contribute most of the information

Enhancements to EM

- List only haplotypes present in sample

- Gradually expand subset of markers under consideration, eliminating haplotypes with low estimated frequency from consideration at each stage
  - SNPHAP, Clayton (2001)
  - HAPLOTyper, Qin et al (2002)
Divide-And-Conquer Approximation

- Number of potential haplotypes increases exponentially
  - Number of observed haplotypes does not

- Approximation
  - Successively divide marker set
  - Locally phase each segment through EM
  - Prune haplotype list as segments are ligated
  - Merge by phasing vectors of haplotype pairs

- Computation order: $\sim m \log m$
  - Exact EM is order $\sim 2^m$

Next Topic:
DISEASE ASSOCIATION STUDIES
Why are we so different?

- Human genetic diversity

Different “phenotype”
- Appearance
- Disease susceptibility
- Drug responses

Different “genotype”
- Individual-specific DNA
- 3 billion-long string

Motivation

- Which sequence variation affects a trait?
  - Better understanding disease mechanisms
  - Personalized medicine

Sequence variations

Obese? 15%
Bold? 30%
Diabetes? 6.2%
Parkinson’s disease? 0.3%
Heart disease? 20.1%
Colon cancer? 6.5%
Detecting Genetic Basis for Disease

- Genome-wide association study ("GWAS")
  - P-value: The probability that we see that much correlation given that the SNP is not relevant to the disease

![Diagram showing genetic markers on 0.1-1M SNPs]

<table>
<thead>
<tr>
<th>Diabetes patients</th>
<th>Normal individuals</th>
</tr>
</thead>
<tbody>
<tr>
<td>...ACTCGGTAGGCATAAATTCGGCCCGGTCAGATTCCATACAGTTTGTACCATGG...</td>
<td>...ACTCGGTAGGCATAAATTCGGCCCGGTCAGATTCCATACAGTTTGTACCATGG...</td>
</tr>
<tr>
<td>...ACTCGGTGGGCATAAATTCGGCCCGGTCAGATTCCATCCAGTTTGTTCCATGG...</td>
<td>...ACTCGGTGGGCATAAATTCTGCCCGGTCAGATTCCATCCAGTTTGTTCCATGG...</td>
</tr>
<tr>
<td>...ACTCGTGACCCGAGAGTCAGTTCCATACAGTTTGTACCATGG...</td>
<td>...ACTCGTGACCCGAGAGTCAGTTCCATACAGTTTGTACCATGG...</td>
</tr>
<tr>
<td>P-value = 0.2</td>
<td>P-value = 1.0e-7</td>
</tr>
</tbody>
</table>

Outline

- Disease association studies
  - Single marker based association tests
  - Haplotype-based approach
  - Indirect association – predicting unobserved SNPs
  - Selection of tag SNPs
A single marker association test

- **Data**
  - Genotype data from case/control individuals
    - e.g. case: patients, control: healthy individuals

- **Goals**
  - Compare frequencies of particular alleles, or genotypes, in set of cases and controls
  - Typically, relies on standard contingency table tests
    - Chi-square goodness-of-fit test
    - Likelihood ratio test
    - Fisher’s exact test

---

Construct contingency table

- Organize genotype counts in a simple table
  - Rows: one row for cases, another for controls
  - Columns: one of each genotype (or allele)
  - Individual cells: count of observations

<table>
<thead>
<tr>
<th>i: case, control</th>
<th>j: 0/0, 0/1, 1/1</th>
<th>j=1</th>
<th>j=2</th>
<th>j=3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0/0</td>
<td>0/1</td>
<td>1/1</td>
<td></td>
</tr>
<tr>
<td>i=1 Case (affected)</td>
<td>0_{1,1}</td>
<td>0_{1,2}</td>
<td>0_{1,3}</td>
<td>(O_{1.1} = O_{0,1} + O_{1,1} + O_{1,3})</td>
</tr>
<tr>
<td>i=2 Control (unaffected)</td>
<td>0_{2,1}</td>
<td>0_{2,2}</td>
<td>0_{2,3}</td>
<td>(O_{2.1} = O_{2,1} + O_{2,2} + O_{2,3})</td>
</tr>
</tbody>
</table>

- **Notation**
  - Let \(O_{ij}\) denote the observed counts in each cell
  - Let \(E_{ij}\) denote the expected counts in each cell
    - \(E_{ij} = O_{i.} \cdot O_{.j} / O_{..}\)
Goodness of fit tests (1/2)

- Null hypothesis
  - There is no statistical dependency between the genotypes and the phenotype (case/control)

- P-value
  - Probability of obtaining a test statistic at least as extreme as the one that was actually observed

- Chi-square test
  \[ \chi^2 = \sum_{i,j} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}} \]
  - If counts are large, compare statistic to chi-squared distribution
    - \( p = 0.05 \) threshold is 5.99 for 2 df (degrees of freedom, e.g. genotype test)
    - \( p = 0.05 \) threshold is 3.84 for 1 df (e.g. allele test)
    - If counts are small, exact or permutation tests are better

Goodness of fit tests (2/2)

- Likelihood ratio test
  - The test statistics (usually denoted D) is twice the difference in the log-likelihoods:
    \[ D = -2 \ln \left( \frac{\text{likelihood for null model}}{\text{likelihood for alternative model}} \right) \]
    \[ D = -2 \ln \left( \prod_{i,j} \frac{E_{i,j}}{O_{i,j}} \right) = 2 \sum_{i,j} O_{i,j} \ln \frac{O_{i,j}}{E_{i,j}} \]

- How about we do this for haplotypes?
  - When does it out-perform the single marker association test?