Terminology Review

• **Allele**: different forms of genetic variations at a given gene or genetic locus
  – Locus 1 has two alleles, A and C, and Locus 2 has two alleles, T and G

• **Genotype**: specific allelic make-up of an individual’s genome
  – Individual 1 has genotype AA at Locus 1 and genotype TG at Locus 2

• **Heterozygous/Homozygous**
  – Locus 1 of Individual 1 is homozygous, and Locus 2 is heterozygous
Single Nucleotide Polymorphism (SNP)

- SNP: “Binary” nucleotide substitutions at a single locus on a chromosome
  - each variant is called an "allele"
From SNPs to Haplotypes

- Haplotype: a stretch of consecutive nucleotides that lie on the same chromosome
- What are the alleles here?
Haplotype $h = (h_1, h_2)$

possible associations of alleles to chromosome
Why Haplotypes?

- Haplotypes have a greater power for discriminating genomic regions
  - Consider $J$ binary markers (e.g., SNPs) in a genomic region
  - There are $2^J$ possible haplotypes
  - SNPs have only two alleles, whereas haplotypes have a larger number of alleles
  - Good genetic marker for population, evolution and hereditary diseases
Haplotypes and SNPs

- **SNPs** can distinguish between two groups of individuals (a group with C, another group with T)
- **Haplotypes** can distinguish between three groups of individuals (each group with CTG, TGA, and CTA)
Haplotypes and SNPs

- Haplotypes can have a greater power to detect disease-related genome region
Inferring Haplotypes from SNP Array Data

- **Genotype:** AC//AA//TG
  - Maternal genotype: CA//AA//TT
  - Paternal genotype: CC//AA//TG
  - Then the haplotype is AAC/TAG.

- **Genotype:** AC//AA//TG
  - Maternal genotype: AC//AA//TG
  - Paternal genotype: AC//AA//TG
  - Cannot determine unique haplotype

- **Problem:** How can we determine haplotypes without parental genotypes
Phasing: Inferring Haplotypes from SNP Data

• Given multilocus genotypes at a set of SNPs for many individuals, phasing means
  – Reconstruct haplotypes for all individuals
  – Estimate frequencies of all possible haplotypes

• Haplotype reconstruction algorithm
  – Clark’s parsimony algorithm (Clark, Mol. Biol. Evol. 1990)
Identifiability

Genotype representations

0/0  →  0
1/1  →  1
0/1  →  2

Genotypes of 14 individual

21 2 222 02
02 1 111 22
11 0 000 01
02 1 111 22
21 2 222 02
02 1 111 22
11 0 000 01
02 1 111 22
21 2 222 02
22 2 222 21
21 1 222 02
02 1 111 22
22 2 222 21
21 2 222 02
|| | ||| ||
Identifiability

Parsimonious solution

\[
\begin{array}{cccccc}
01 & 1 & 111 & 00 \\
11 & 0 & 000 & 01 \\
01 & 1 & 111 & 00 \\
00 & 1 & 111 & 11 \\
11 & 0 & 000 & 01 \\
11 & 0 & 000 & 01 \\
01 & 1 & 111 & 00 \\
00 & 1 & 111 & 11 \\
01 & 1 & 111 & 00 \\
11 & 0 & 000 & 01 \\
11 & 0 & 000 & 01 \\
01 & 1 & 111 & 00 \\
00 & 1 & 111 & 11 \\
01 & 1 & 111 & 00 \\
00 & 1 & 111 & 11 \\
01 & 1 & 111 & 00 \\
00 & 1 & 111 & 11 \\
01 & 1 & 111 & 00 \\
11 & 0 & 000 & 01 \\
\end{array}
\]

\[
\begin{array}{cccccc}
01 & 1 & 101 & 00 \\
11 & 0 & 010 & 01 \\
01 & 1 & 111 & 00 \\
00 & 1 & 111 & 11 \\
11 & 0 & 000 & 01 \\
11 & 0 & 000 & 01 \\
01 & 1 & 111 & 00 \\
00 & 1 & 111 & 11 \\
01 & 1 & 111 & 00 \\
01 & 1 & 111 & 00 \\
01 & 1 & 111 & 00 \\
00 & 1 & 111 & 11 \\
11 & 0 & 000 & 01 \\
00 & 1 & 111 & 11 \\
01 & 1 & 111 & 00 \\
00 & 1 & 111 & 11 \\
01 & 1 & 111 & 00 \\
11 & 0 & 000 & 01 \\
00 & 1 & 111 & 11 \\
01 & 1 & 111 & 00 \\
11 & 0 & 000 & 01 \\
00 & 1 & 111 & 11 \\
01 & 1 & 111 & 00 \\
01 & 1 & 111 & 00 \\
01 & 1 & 111 & 00 \\
00 & 1 & 111 & 11 \\
11 & 0 & 000 & 01 \\
00 & 1 & 111 & 11 \\
\end{array}
\]
Haplotype Reconstruction Algorithm by Clark (1990)

- Choose individuals that are homozygous at every locus (e.g. TT/AA/CC)
  - Haplotype: TAC
- Choose individuals that are heterozygous at just one locus (e.g. TT/AA/CG)
  - Haplotypes: TAC or TAG
- Tally the resulting known haplotypes.
- For each known haplotype, look at all remaining unresolved cases: is there a combination to make this haplotype?
  - Known haplotype: TAC
    - Unresolved pattern: AT/AA/CG
    - Inferred haplotype: TAC/AAG. Add to list.
  - Known haplotype: TAC and TAG
    - Unresolved pattern: AT/AA/CG
    - Inferred haplotypes: TAC and TAG. Add both to list.
- Continue until all haplotypes have been recovered or no new haplotypes can be found this way.
Problems: Clark (1990)

• Many unresolved haplotypes at the end
• Ignores recombination
  – Error in haplotype inference if a crossover of two actual haplotypes is identical to another true haplotype
  – Frequency of such errors depends on recombination rate
• Clark (1990): algorithm "performs well" even with small sample sizes.
RECOMBINATION & LINKAGE
DISEQUILIBRIUM
**Morgan’s Fruitfly Experiment**

*Morgan’s fruitfly data (1909):* 2,839 flies

Eye color
- A: red
- a: purple

Wing length
- B: normal
- b: vestigial

AABB x aaBB

AaBb x aaBB

<table>
<thead>
<tr>
<th></th>
<th>Exp</th>
<th>Obs</th>
</tr>
</thead>
<tbody>
<tr>
<td>AaBb</td>
<td>710</td>
<td>1,339</td>
</tr>
<tr>
<td>Aabb</td>
<td>710</td>
<td>151</td>
</tr>
<tr>
<td>aaBb</td>
<td>710</td>
<td>154</td>
</tr>
<tr>
<td>aabb</td>
<td>710</td>
<td>1,195</td>
</tr>
</tbody>
</table>
“Linked” Genes

- When two genes lies on the same chromosome, they are transmitted to offspring in a non independent manner
Morgan’s Explanation: Recombination

\( F_1: \)

\( \times \)

\( F_2: \)

\( \otimes \) Recombination has taken place
Recombination

- Parental types: AaBb, aabb
- Recombinants: Aabb, aaBb
  - The proportion of recombinants between the two genes (or characters) is called the recombination fraction between these two genes.
Review: Correlation

• “GPA” and “TV in hours per week” are negatively correlated

<table>
<thead>
<tr>
<th>Participant</th>
<th>GPA</th>
<th>TV in hours per week</th>
</tr>
</thead>
<tbody>
<tr>
<td>#1</td>
<td>3.1</td>
<td>14</td>
</tr>
<tr>
<td>#2</td>
<td>2.4</td>
<td>10</td>
</tr>
<tr>
<td>#3</td>
<td>2.0</td>
<td>20</td>
</tr>
<tr>
<td>#4</td>
<td>3.8</td>
<td>7</td>
</tr>
<tr>
<td>#5</td>
<td>2.2</td>
<td>25</td>
</tr>
<tr>
<td>#6</td>
<td>3.4</td>
<td>9</td>
</tr>
<tr>
<td>#7</td>
<td>2.9</td>
<td>15</td>
</tr>
<tr>
<td>#8</td>
<td>3.2</td>
<td>13</td>
</tr>
<tr>
<td>#9</td>
<td>3.7</td>
<td>4</td>
</tr>
<tr>
<td>#10</td>
<td>3.5</td>
<td>21</td>
</tr>
</tbody>
</table>

Mean GPA: 3.02
Mean TV: 13.8

How can we quantify the level of correlation?
Covariance and Correlation

• Degree of association between two variables x and y

• Given observations $x_1, ..., x_n$ and $y_1, ..., y_n$
  – Covariance

$$\frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{(n - 1)}$$

  – Correlation coefficient:

$$\frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{(n - 1)s_{xy}} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

  (Variance of $x_i$’s) x (n-1)

  (Variance of $y_i$’s) x (n-1)

• Falls between -1 and +1, with sign indicating direction of association
Correlation between $X_1$ and $X_2$

- Correlation $r = 0$
- Correlation $r = -0.3$
- Correlation $r = 0.5$
- Correlation $r = -0.7$
- Correlation $r = 0.9$
- Correlation $r = -0.99$
**Basic Concepts**

**High LD** -> No Recombination

\[ r^2 = 1 \]

SNP1 “tags” SNP2

**Low LD** -> Recombination

Many possibilities

\[ \text{etc...} \]
Linkage Disequilibrium (LD)

- LD reflects the relationship between alleles at different loci.
- Often, $r^2$ (squared correlation coefficient) is used as a measure of LD.
How to Compute $r^2$ on SNP Data

<table>
<thead>
<tr>
<th>Individuals</th>
<th>SNP1</th>
<th>SNP2</th>
<th>SNP3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

$r^2$ matrix

<table>
<thead>
<tr>
<th>SNP1</th>
<th>SNP2</th>
<th>SNP3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>1.0</td>
<td>0.0</td>
</tr>
<tr>
<td>1.0</td>
<td>1.0</td>
<td>0.0</td>
</tr>
<tr>
<td>0.0</td>
<td>0.0</td>
<td>1.0</td>
</tr>
</tbody>
</table>
Linkage Disequilibrium in SNP Data

• $r^2$ in SNP data from a population of individuals (Black: $r^2=1$, white: $r^2=0$)
Reducing Genotyping Costs with Tag SNPs

• Nearby SNPs in the genome are in linkage disequilibrium (LD), and thus contain redundant information.

• If we knew which SNPs are in LD, we can pre-select the representative SNPs for each LD block of chromosome, and genotype only for those SNPs.

\( r^2 \) values
(black: \( r^2 = 1 \), white: \( r^2 = 0 \))

These two SNPs are in high LD and thus are redundant
Reducing Genotyping Costs with Tag SNPs

• Two-stage data collection process
  – Stage 1:
    • Collect genotype data for a dense set of SNPs for multiple individuals
    • Select a non-redundant set of tag SNPs by examining the LD pattern
  – Stage 2:
    • Collect genotype data only for the tagSNPs for a large number of individuals
Algorithm for Selecting Tag SNPs

• Greedy algorithm

Randomly select a tag SNP

Find the SNPs with a high LD with the previously selected tag SNP ($r^2 > 0.8$) and remove those SNPs from the set of candidate tag SNPs

Iterate until the set of candidate tag SNPs is empty
Recombination and Haplotypes

• Remember Clark’s method does not take into account recombination

• How can we find haplotypes from SNP data collected for a population of individuals under recombination?
  – Assume haplotypes of ancestor chromosomes and treat modern individuals’ chromosomes as a mosaic of ancestor chromosomes
  – However, ancestor chromosomes cannot be observed!

• Key idea:
  – Haplotype of each individual is a mosaic of other individuals’ haplotypes
  – unresolved haplotypes are similar to known haplotypes
Recombination and Haplotypes

• $h_1, h_2, h_3$: unobserved ancestral haplotypes
  – we have no SNP data

• $h_{4A}, h_{4B}$: unobserved haplotypes for modern individuals
  – Haplotypes are unobserved, however, we have SNP data

• Circles: mutations

Mosaic of ancestor chromosomes
PHASE Model as an HMM

- Inferring the unobserved state labels for each of the observed SNP amounts to haplotype reconstruction.
PHASE Model as an HMM

- **States:** $h_1$, $h_2$, $h_3$, unobserved ancestral haplotypes

  - State space with possible transitions

- **Transition probabilities** (from SNP $X_l$ to $X_{l+1}$) are dependent on
  - distance between adjacent SNPs $d_l$
  - Recombination rate between adjacent SNPs $\rho_l$

- **Emission probabilities:** mutation model

- **Task:** infer hidden state labels for each locus of each individual ($h_{4A}$, $h_{4B}$)
INTERNATIONAL HAPMAP PROJECT
(HAPMAP.ORG)
## HapMap Phase 3 Samples

<table>
<thead>
<tr>
<th>label</th>
<th>population sample</th>
<th># samples</th>
<th>QC+ Draft 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASW*</td>
<td>African ancestry in Southwest USA</td>
<td>90</td>
<td>71</td>
</tr>
<tr>
<td>CEU*</td>
<td>Utah residents with Northern and Western European ancestry from the CEPH collection</td>
<td>180</td>
<td>162</td>
</tr>
<tr>
<td>CHB</td>
<td>Han Chinese in Beijing, China</td>
<td>90</td>
<td>82</td>
</tr>
<tr>
<td>CHD</td>
<td>Chinese in Metropolitan Denver, Colorado</td>
<td>100</td>
<td>70</td>
</tr>
<tr>
<td>GIH</td>
<td>Gujarati Indians in Houston, Texas</td>
<td>100</td>
<td>83</td>
</tr>
<tr>
<td>JPT</td>
<td>Japanese in Tokyo, Japan</td>
<td>91</td>
<td>82</td>
</tr>
<tr>
<td>LWK</td>
<td>Luhya in Webuye, Kenya</td>
<td>100</td>
<td>83</td>
</tr>
<tr>
<td>MEX*</td>
<td>Mexican ancestry in Los Angeles, California</td>
<td>90</td>
<td>71</td>
</tr>
<tr>
<td>MKK*</td>
<td>Maasai in Kinyawa, Kenya</td>
<td>180</td>
<td>171</td>
</tr>
<tr>
<td>TSI</td>
<td>Toscans in Italy</td>
<td>100</td>
<td>77</td>
</tr>
<tr>
<td>YRI*</td>
<td>Yoruba in Ibadan, Nigeria</td>
<td>180</td>
<td>163</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1,301</td>
<td>1,115</td>
</tr>
</tbody>
</table>

* Population is made of family trios
Haplotype Structure and Recombination Rate Estimates: HapMap I vs. HapMap II
HapMap: Allele Frequencies in Different Populations

- Comparison of allele frequencies for individuals from pairs of populations

- The red regions show that there are many SNPs that have similar low frequencies in each pair of analysis panels/populations.

- CHB (Chinese) and JPT (Japanese) have similar allele frequencies
Why Haplotypes?

• Haplotypes have a greater power for discriminating genomic regions
  – Consider $J$ binary markers (e.g., SNPs) in a genomic region
  – There are $2^J$ possible haplotypes
    • but in fact, far fewer are seen in human population
  – SNPs have only two alleles, whereas haplotypes have a larger number of alleles
  – Good genetic marker for population, evolution and hereditary diseases
Summary

• Haplotype: a set of genetic markers that lie on the same chromosome

• How can we find haplotypes from SNPs?

• Recombination, linkage disequilibrium, and how to take advantage of them
  – Haplotypes as a set of linked SNPs with a greater discriminative power
  – Tag SNPs for saving the genotyping cost

• HapMap Project