INTRODUCTION TO THE HAPMAP PROJECT

Zané Lombard
Wits Bioinformatics

Adapted from tutorials on the HapMap website (www.hapmap.org) & slide material from L. Ouko

Adapted from tutorials on the HapMap website (www.hapmap.org) & slide material from L. Ouko
Human genetic variation

- Human Genome $\rightarrow$ $\sim$3 billion bp
  - Contain information that influences our physical traits, our likelihood of suffering from disease, and the responses of our bodies to substances that we encounter in the environment.
  - 99.9% similarity among individuals
  - 0.1% difference impacts variable response to environment, pathogens etc among individuals
  - only about 1.5% of the genome codes for proteins
Single nucleotide polymorphisms (SNPs)

- Most common genetic variant
- SNPs are used as markers to locate genes in DNA sequences - useful in disease mapping
- Testing 12 million common SNPs would be extremely expensive

- For a case-control study with 1,000 cases & 1,000 controls
- Genotype all DNAs for all SNPs
- That adds up to 24 billion genotypes
- Imagine, this approach cost 50 cents a genotype.
- That’s R12 billion for each disease – completely out of the question!!
The International HapMap Project aims to identify a large fraction of the genetic diversity in the human species.

The development of the HapMap will enable geneticists to take advantage of how SNPs and other genetic variants are organized on chromosomes.

- Genetic variants that are near each other tend to be inherited together.
- E.g. all of the people who have an A rather than a G at a particular location in a chromosome can have identical genetic variants at other SNPs in the chromosomal region surrounding the A.
- These regions of linked variants are known as haplotypes. This phenomenon is influenced by recombination & linkage disequilibrium.
Recombination between Homologous Chromosomes

Alleles: A and a
B and b
C and c

Red and Blue are homologous Chromosomes, one from each parent
Origins of haplotypes

- The non-random association between alleles in a population

No LD

2 SNPs = 4 Haplotypes

Hi LD

2 SNPs = 2 Haplotypes
SNPs, Haplotypes & tagSNPs

**a** SNPs

<table>
<thead>
<tr>
<th>Chromosome</th>
<th>SNPs</th>
</tr>
</thead>
</table>

**b** Haplotypes

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

**c** Tag SNPs

<table>
<thead>
<tr>
<th>Tag SNPs</th>
</tr>
</thead>
<tbody>
<tr>
<td>A / G</td>
</tr>
<tr>
<td>T / C</td>
</tr>
<tr>
<td>C / G</td>
</tr>
</tbody>
</table>
Haplotypes

- SNPs that occur together suggests underlying structure to genome
- SNPs occur in blocks of which there are common varieties
- ~65% to 85% of the human genome is organized in haplotypes
- If blocks easily identified could be important tool for studying genetic variation in relation to disease, drug response etc..
- Founded in 2002

- Participating institutions and funding from Japan, UK, Canada, China, USA and Nigeria

- “...develop a haplotype map of the human genome, which will describe the common patterns of human DNA sequence variation”
Strategy

1. Recruit individuals that represent global diversity
2. Genotype SNPs for all individuals
3. Identify chromosomal regions with groups of strongly associated SNPs – haplotypes
4. Determine linkage disequilibrium between SNPs
5. Identify tagSNPs for the haplotypes
Populations sampled

- Yoruba people in Ibadan, Nigeria
  - 30 both-parent-and-adult-child trios
- Japanese in Tokyo
  - 45 unrelated individuals
- Han Chinese in Beijing
  - 45 unrelated individuals
- The U.S. Utah residents of northern and western European ancestry
  - 30 trios
  - Residents with ancestry from Northern and Western Europe
Genotyping

- 11 Centers for typing: Canada, China, Japan, UK, USA
- Genotyped at least one common SNP every 5 kb
- The Phase I HapMap contained 1,007,329 SNPs that passed a set of quality control filters
  - SNPs at $f > 0.05$ MAF chosen
- The HapMap Project contributed ~6 million new SNPs to dbSNP
  - In 2005 dbSNP contained 9.2 million candidate human SNPs, of which 3.6 million have been validated by both alleles having been seen two or more times during discovery (‘double-hit’ SNPs), and 2.4 million have genotype data
Haplotyping

- Phased haplotypes were generated using the program PHASE version 2.0

- Each allele in a genotype is assigned to one or the other parental chromosome using computer algorithms

- The numbers and size of possible haplotypes are limited because of recombination events
Haplotype output

Nature 426 Dec 2003
Linkage disequilibrium

- If two alleles tend to be inherited together more often than would be predicted, then the alleles are in linkage disequilibrium.

- The basis of measuring Linkage Disequilibrium is the difference between the observed and expected frequencies of pairs of alleles.

- LD in all genotyped SNPs calculated
LD Measures

- **D prime (D’)**
  - D’ is the difference between the expected and the observed haplotype frequency.
  - D' (normalised LD) is the only measure of LD not sensitive to allele frequencies.
  - A score of 1 = LD

- **R square (r²)**
  - The square of the correlation coefficient r, a measure of the effect of X in reducing the uncertainty in predicting Y.
  - Gives information on sample size required to detect association.
  - A score of 1 = LD

- **Likelihood of Odds (LOD) Score**
  - The logarithm of odds - a statistical measure of the likelihood that two genetic markers occur together on the same chromosome and are inherited as a single unit of DNA (co-segregation).
  - A score of >2 = LD
The triangle plot is constructed by connecting every pair of SNPs along lines at 45 degrees to the horizontal track line. The colour of the diamond at the position that two SNPs intersect indicates the amount of LD: more intense colours indicate higher LD. A grey diamond indicates missing data.
LD AND tagSNPs

- Reduce the number of SNPs needed to genotype region (use few tagSNPs)
  - High LD - few SNPs sampled
  - Low LD – more SNPs sampled
Interesting findings

A: Similarity of allele frequencies in CHB/JPT samples.
   - These were subsequently analyzed jointly

B: Identification of recombination hot spots
   - 21,617 identified recombination hotspots
   - ~1 per 122 kb
Interesting findings

C: Haplotype sizes vary across populations due to migrational history

- Haplotypes in non-African populations tend to be longer than in African populations

D: LD correlates to genomic features

- Areas of very high and very low LD have the highest density of genes
- LD low
  - associated with immune and neuro-physiological genes
- LD elevated
  - associated with cell cycle regulators, DNA damage responses, DNA/RNA metabolism.
How HAPMAP could benefit human health

- Provide an extensive resource that researchers can use to discover the genetic variants involved in disease and individual responses to therapeutic agents

- Learn much more about the origins of illnesses and about ways to prevent, diagnose and treat

- Association studies

- Customizable treatment, new therapies
Critique

- How to define the haplotype block boundaries
- How universal are the blocks (from population to population)
- Genetic variations may reinforce racial stereotypes
- Groups studied do not represent human diversity
- Different methods for selecting tag SNPs give different answers
Retort

- It's meant as a tool to study genetic variation at unprecedented levels of accuracy and detail.
- Offers a direct route to testing ideas about the genetics of common diseases.
- Unsuccessful track-record genetics has in dissecting complex disease traits.
- All association studies need to be replicated in different groups of people and with rigorous statistical tests.
# HAPMAP – Phase Comparison

<table>
<thead>
<tr>
<th></th>
<th>Phase 1</th>
<th>Phase 2</th>
<th>Phase 3</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Samples &amp; POP panels</strong></td>
<td>269 samples (4 panels)</td>
<td>270 samples (4 panels)</td>
<td>1,184 samples (11 panels)</td>
</tr>
<tr>
<td><strong>Genotyping centers</strong></td>
<td>HapMap International Consortium</td>
<td>Perlegen</td>
<td>Broad &amp; Sanger</td>
</tr>
<tr>
<td><strong>Unique SNPs</strong></td>
<td>1.1 M</td>
<td>3.8 M (phase I+II)</td>
<td>1.6 M (Affy 6.0 &amp; Illumina 1M)</td>
</tr>
<tr>
<td><strong>Sequence Data</strong></td>
<td>---</td>
<td>---</td>
<td>Sequenced ten 100-kb regions (n=692)</td>
</tr>
</tbody>
</table>
# HAPMAP Phase III

<table>
<thead>
<tr>
<th>LABEL</th>
<th>POPULATION SAMPLE</th>
<th># Samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASW</td>
<td>African ancestry in Southwest USA</td>
<td>90</td>
</tr>
<tr>
<td>CEU</td>
<td>Utah residents with Northern and Western European ancestry from the CEPH collection</td>
<td>180</td>
</tr>
<tr>
<td>CHB</td>
<td>Han Chinese in Beijing, China</td>
<td>90</td>
</tr>
<tr>
<td>CHD</td>
<td>Chinese in Metropolitan Denver, Colorado</td>
<td>100</td>
</tr>
<tr>
<td>GIH</td>
<td>Gujarati Indians in Houston, Texas</td>
<td>100</td>
</tr>
<tr>
<td>JPT</td>
<td>Japanese in Tokyo, Japan</td>
<td>91</td>
</tr>
<tr>
<td>LWK</td>
<td>Luhya in Webuye, Kenya</td>
<td>100</td>
</tr>
<tr>
<td>MEX</td>
<td>Mexican ancestry in Los Angeles, California</td>
<td>90</td>
</tr>
<tr>
<td>MKK</td>
<td>Maasai in Kinyawa, Kenya</td>
<td>180</td>
</tr>
<tr>
<td>TSI</td>
<td>Toscans in Italy</td>
<td>100</td>
</tr>
<tr>
<td>YRI</td>
<td>Yoruba in Ibadan, Nigeria</td>
<td>180</td>
</tr>
</tbody>
</table>

1,301
HapMap 3 Samples

- 1,184 samples from diverse populations (N=11)
- Individual and community consent for thorough genetic ascertainment (up to complete resequencing) and public sharing of data on Internet
Interesting Outcomes

- Of the SNPs identified through sequencing, 77% were new (i.e. not previously in dbSNP) and 99% of those had a MAF < 5%
- Reveal that many more variants remain to be found, especially rare variants

The International HapMap 3 Consortium, Nature Sept 2010; 467:52-58
Interesting Outcomes

- Confirmed that non-African diversity is largely a subset of African diversity
- African samples provided a more complete discovery resource for variant sites in non-African than the converse

However, it does not work as well for rare variants
- Rare variants could likely be more NB in population-specific contributions to disease?
- Underscores the value of next-gen sequencing of whole genomes within various populations to find rare variants that contribute to disease.
Using HAPMAP data: Population substructure in Africans

Here is the result of running ADMIXTURE on the three African HapMap-3 populations, using about 440K SNPs, including Tuscans as a non-African group.