Appropriate Selection of Tagging SNPs in Indirect Association Studies

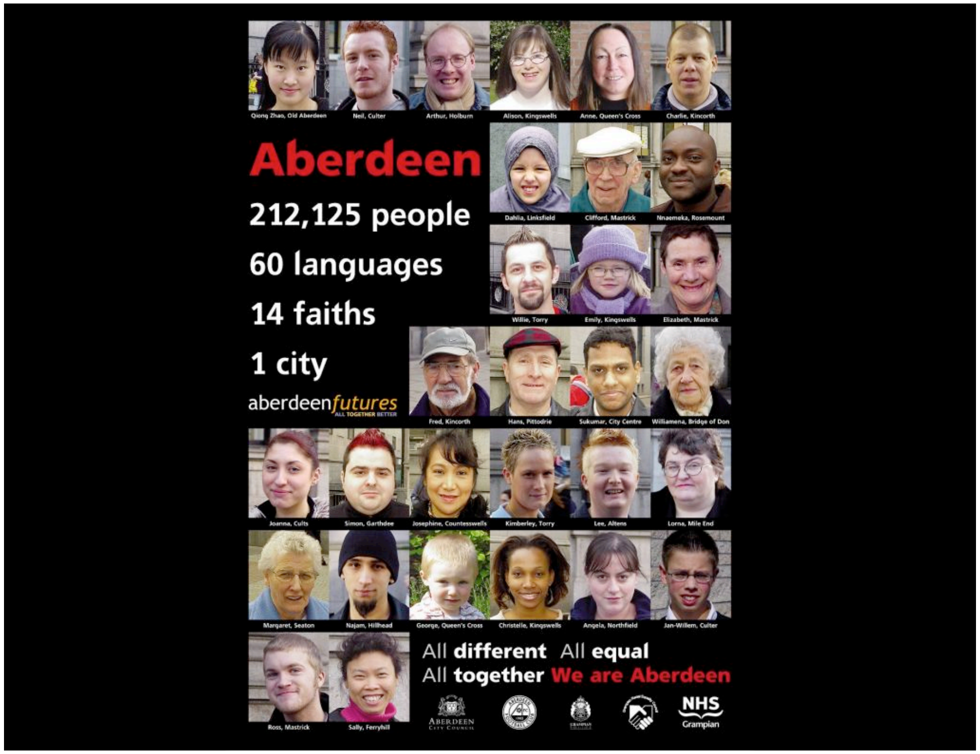
Ryan Roper

CSE 527 Final Presentation

December 15, 2004

Outline

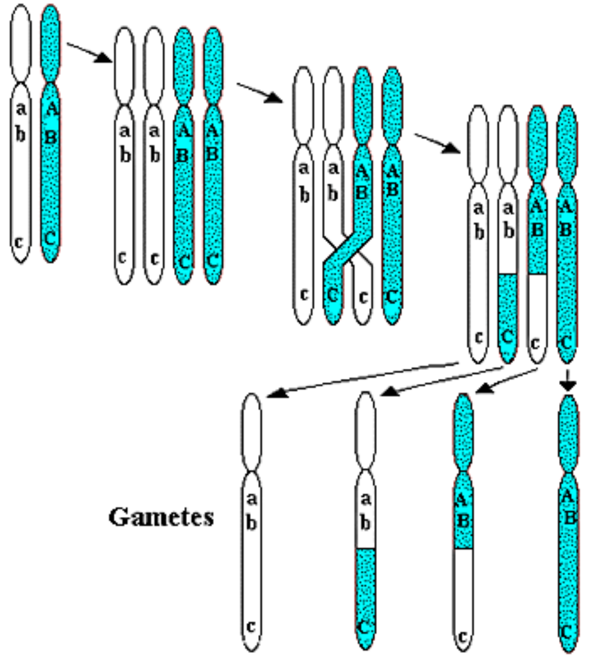
- Background
- Indirect Association Studies
- Selection Of Tagging SNPs For Genotyping
- Discussion of Two Approaches
- Summary



http://www.aberdeencity.gov.uk/acc/YourCity/default.asp

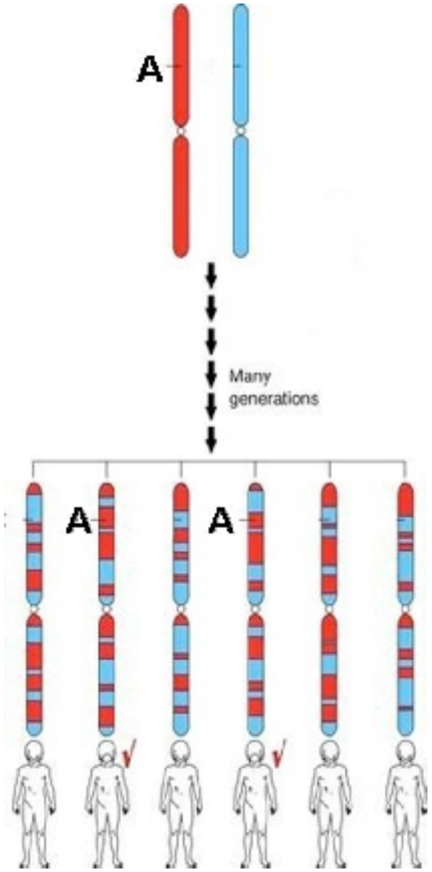
- 99.9% sequence conservation in the human population
- 80% of divergence occurs in the form of single-nucleotide polymorphisms (SNP)
- **Definition:** A SNP is a single base pair substitution that occurs with a frequency of >1 % at the site where it is located

Meiotic Recombination



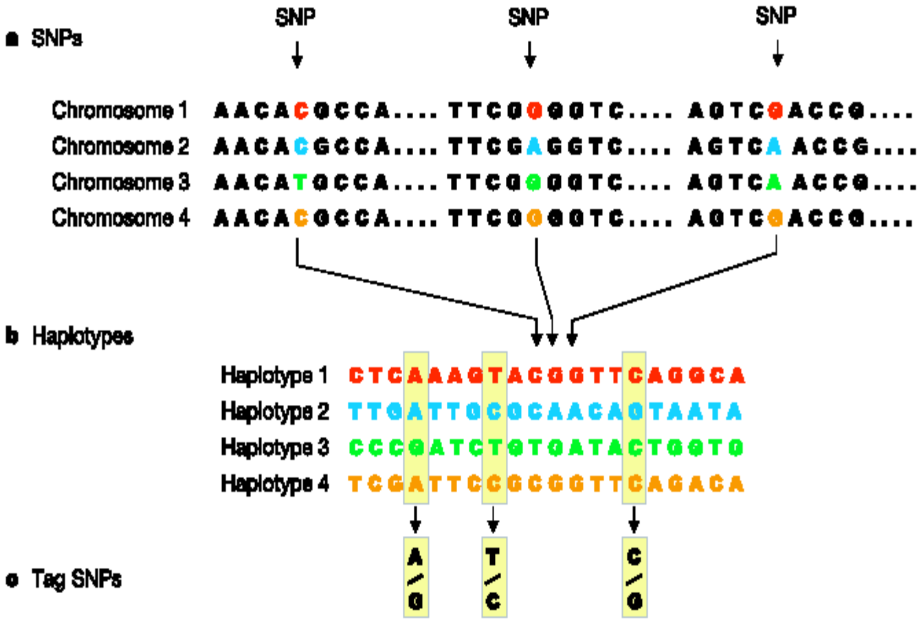
Crossing-over and recombination during meiosis

http://www.accessexcellence.org

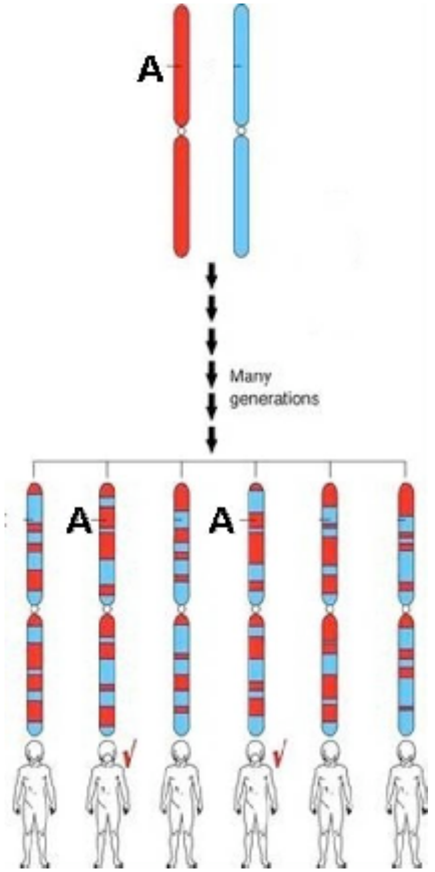


http://www.hapmap.org/originhaplotype.html

Recombination events during meiosis result in rearrangement or shuffling of chromosomal segments


a SNPs
SNP
SNP
SNP
Chromosome 1 AACACGCCA.... TTCGGGGTC.... AGTCGACCG....
Chromosome 2 AACACGCCA.... TTCGAGGTC.... AGTCAACCG....
Chromosome 3 AACATGCCA.... TTCGGGGTC.... AGTCAACCG....
Chromosome 4 AACACGCCA.... TTCGGGGTC.... AGTCGACCG....
b Haplotypes
Haplotype 1 CTCAAAGTACGGTTCAGGCA
Haplotype 2 TTGATTGCGCAACAGTAATA
Haplotype 3 CCCGATCTGTGATACTGGTG
Haplotype 4 TCGATTCGGCGGTTCAGACA
c Tag SNPs
A/G
T/C
C/G

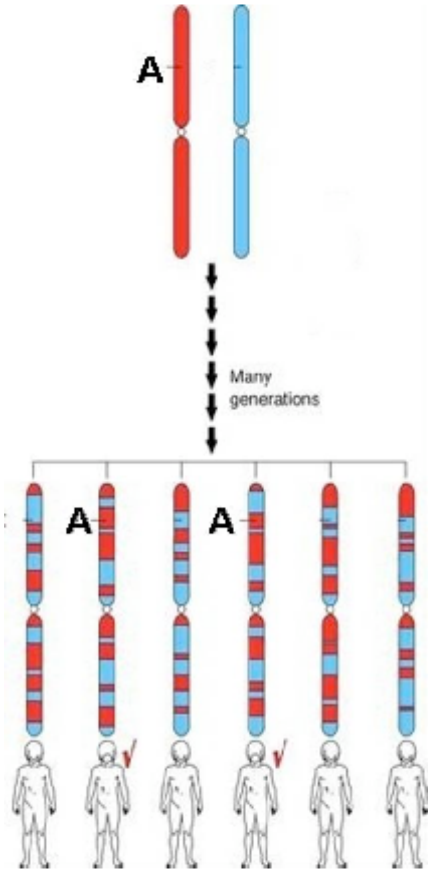
The International HapMap Project *Nature*. 2003 Dec 18;426(6968):789-96



http://www.hapmap.org/originhaplotype.html

- **Haplotypes** are sets of SNPs on the same chromosomal segment that tend to be transmitted as a block
- **Tagging SNPs** are a subset of SNPs that may be used to uniquely identify haplotypes

Association Studies



http://www.hapmap.org/originhaplotype.html

Objective: to identify allelic variants that tend to be correlated with the occurrence of a disease

Direct vs. Indirect Association Studies

- Direct
  - Hypothesis-driven approach requiring prior information about potential disease risk of a gene or set of genes
  - Genotyping over a relatively small portion of the genome
- Indirect
  - Discovery-based approach that does not require prior information about potential disease risk loci
  - Genotyping must be done with broad coverage of the genome

Indirect Association Studies

- There is an estimated 10 million SNPs in the human genome.
- Indirect association studies, therefore, require selection of tagging SNPs (tSNPs) that uniquely identify haplotypes.

Assumption: Risk-related polymorphic loci will either be directly typed or will be correlated with one or more typed polymorphisms.

An Important Issue in Indirect Association Studies:

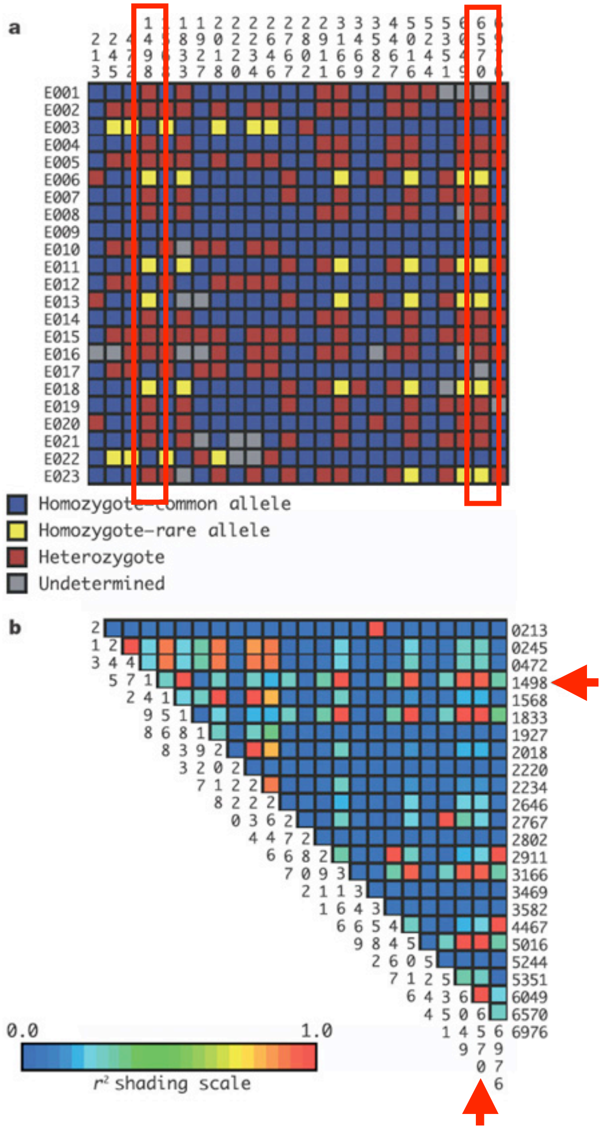
How to optimally select a set of markers (i.e. tag SNPs) such that the set will provide adequate information for association without requiring an excessive number of loci to be genotyped.

Linkage Disequilibrium (LD)

Definition: Two loci that are in linkage disequilibrium are inherited together more often than would be expected by chance

Parameters: Either r^2 or D'

Note: $r^2 = 1$ is stronger than $D' = 1$ in that it requires two loci to have identical allele frequencies



Linkage disequilibrium between two loci is determined by similarity in inheritance pattern across many individuals

LD in the Context of Haplotype Blocks

- A haplotype block is a group of SNPs showing a high degree of LD (high r^2 or D′) within the group and, ideally, a comparatively low LD with SNPs in other blocks
- Haplotype blocks (or groups) are also referred to as LD groups

One Example From the Literature

Calson et al. Selecting a Maximally Informative Set of Single-Nucleotide Polymorphisms for Association Analyses Using Linkage Disequilibrium. 2004. *Am. J. Hum. Genet*. 74:106-120.

Algorithm for LD-Grouping

1) Select from all SNPs exceeding a specified marker allele frequency (MAF)
2) Identify the SNP in linkage disequilibrium (above a specified r^2 threshold) with the most other sites above the specified MAF
3) Calculate all pairwise r^2 within this group or bin and specify those SNPs that exceed the r^2 threshold with all other sites in the bin as tagSNPs. Only one tSNP per bin is genotyped.
4) Iterate this process until all SNPs exceeding the MAF threshold are binned. A SNP not exceeding the r^2 threshold with any other SNP is placed alone in a bin.

Test Data for LD-Grouping

- 47 unrelated individuals – 24 African Americans and 23 European Americans
- 100 genes were resequenced

Results of Carlson et al.

LD threshold (r^2)	AA	EA
0.5	5.2 tSNPs per 10 kb (~500,000 genome-wide)	2.6 tSNPs per 10 kb (~250,000 genome-wide)
0.8	8.25 tSNPs per 10 kb (~800,000 genome-wide)	4.2 tSNPs per 10 kb (~400,000 genome-wide)

- Average number of tSNPs is higher in higher-diversity populations
- Desired threshold or correlation between typed and un-typed SNPs greatly influences number of tSNPs

Important Considerations

- Increased r^2 threshold provides greater statistical confidence in associations, but also requires more tSNPs
- Population stratification is important in selecting optimal tSNPs and should be taken into consideration

Summary

- Indirect association studies attempt to map disease loci by genotyping a subset of SNPs (tagging SNPs)
- Quality of results are dependent upon appropriately selecting tagging SNPs
- Some important considerations are LD threshold and population stratification