Of Sea Urchins, Birds and Men
Introduction

SNPs, HAPLOTYPES
A SNP is a position in a genome at which two or more different bases occur in the population, each with a frequency >1%.

The two alleles at the site are G and T

- The most abundant type of polymorphism
Human Genome contains ~ 3 G basepairs arranged in 46 chromosomes.

Two individuals are 99.9% the same. I.e. differ in ~ 3 M basepairs.

SNPs occur once every ~600 bp

Average gene in the human genome spans ~27Kb

~50 SNPs per gene
Haplotypes

C A G
T T G

Two individuals

SNP SNP SNP
Infinite Sites Assumption:

Each site mutates at most once.
Haplotype Pattern

At each SNP site label the two alleles as 0 and 1. The choice which allele is 0 and which one is 1 is arbitrary.
Recombination

Synopsis: Pairing of homologous chromosomes

Paternal

Maternal

Crossing over

AGTCTCGACCATATAC
ACGTTATTATTTA
Recombination

The two alleles are linked, i.e., they are "traveling together".

Recombination disrupts the linkage.
Linkage Disequilibrium (LD)

Emergence of Variations Over Time

Variations in Chromosomes Within a Population

Common Ancestor

Disease Mutation

time

present
Extent of Linkage Disequilibrium

Disease-Causing Mutation

2,000 gens. ago

1,000 gens. ago

Time = present
A Data Compression Problem

- Select SNPs to use in an association study
  - Would like to associate single nucleotide polymorphisms (SNPs) with disease.

- Very large number of candidate SNPs
  - Chromosome wide studies, whole genome-scans
  - For cost effectiveness, select only a subset.

- Closely spaced SNPs are highly correlated
  - It is less likely that there has been a recombination between two SNPs if they are close to each other.
Disease Associations
Association studies

Disease Responder
Allele 0

Control Non-responder
Allele 1

Marker A is associated with Phenotype

Marker A:
Allele 0 =
Allele 1 =
**Association studies**

Evaluate whether nucleotide polymorphisms associate with phenotype

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>A</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>A</td>
<td>C</td>
<td>G</td>
<td>A</td>
<td>T</td>
</tr>
<tr>
<td>A</td>
<td>T</td>
<td>A</td>
<td>A</td>
<td>G</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>T</td>
<td>A</td>
<td>G</td>
<td>T</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>T</td>
<td>G</td>
<td>G</td>
<td>T</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>T</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td></td>
</tr>
</tbody>
</table>
Association studies

<table>
<thead>
<tr>
<th>A</th>
<th>C</th>
<th>G</th>
<th>A</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>C</td>
<td>G</td>
<td>A</td>
<td>T</td>
</tr>
<tr>
<td>A</td>
<td>T</td>
<td>A</td>
<td>A</td>
<td>G</td>
</tr>
<tr>
<td>C</td>
<td>T</td>
<td>A</td>
<td>G</td>
<td>T</td>
</tr>
<tr>
<td>A</td>
<td>T</td>
<td>G</td>
<td>G</td>
<td>T</td>
</tr>
<tr>
<td>A</td>
<td>T</td>
<td>G</td>
<td>G</td>
<td>G</td>
</tr>
</tbody>
</table>
Association studies

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>
Data Compression

ACGATCGATCATGAT
ACGATCGATCATGAT
GGTGATTGCATCGAT
GGTGATTGCATCGAT
ACGATCGGGCTTCCG
ACGATCGGGCTTCCG
GGTGATATTATCATGAT
GGTGATATTATCATGAT

Haplotype Blocks based on LD
(Method of Gabriel et al. 2002)

Selecting Tagging SNPs in blocks
Real Haplotype Data

A region of Chr. 22
45 Caucasian samples

Two different runs of the Gabriel et al Block Detection method +
Zhang et al SNP selection algorithm

Our block-free algorithm