Background

- Definitions
  - Inheritance of mutations at a locus
  - Alleles
  - SNPs (single nucleotide polymorphisms): A DNA sequence variation occurring when a single nucleotide differs
Abstraction of a causal mutation

Goal is to identify the causal mutations (the genomic loci) & causal genes

Looking for the causal mutation in populations

A possible strategy is to collect cases (affected) and control individuals, and look for a mutation that consistently separates the two classes. Next, identify the gene.
Looking for the causal mutation in populations

Problem 1: many unrelated common mutations, around one every 1000bp

Looking for the causal mutation in populations

Problem 2: We may not sample the causal mutation.
How to hunt for disease genes?

- We are guided by two simple facts governing these mutations
  1. Nearby mutations are correlated
  2. Distal mutations are not

Association mapping

- Sample a population of individuals at variant locations across the genome. Typically, these variants are single nucleotide polymorphisms (SNPs).
- Create a new bi-allelic variant corresponding to cases and controls, and test for correlations.
- By our assumptions, only the proximal variants will be correlated.
- Investigate genes near the correlated variants.
Consider a fixed population (of chromosomes) evolving in time.
• Each individual arises from a unique, randomly chosen parent from the previous generation.

Genealogy of a chromosomal population

Current (extant) population
Adding mutations

Infinite sites assumption: A mutation occurs at most once at a site.

Bafna, BIB

SNPs

The collection of acquired mutations in the extant population describe the SNPs

Bafna, BIB
Fixation and elimination

- Not all mutations survive.
- Some mutations get fixed, and are no longer polymorphic

Bafna, BIB

Removing extinct genealogies

Bafna, BIB
Removing fixed mutations

A coalescent tree: A phylogenetic tree based on the coalescent theory
Disease mutation

• We drop the ancestral chromosomes, and place the mutations on the internal branches.

Bafna, BIB

Disease mutation

• A causal mutation creates a clade of affected descendants.
• The underlying genealogy creates a correlation between SNPs.

Bafna, BIB
In the idealized model, we assume that each individual chromosome chooses two parental chromosomes from the previous generation.
Change of the local genealogy

Correlation of SNPs

- Proximal SNPs are correlated, distal SNPs are not.
  (The correlation decays rapidly after 20-50kbp)
# Basic Statistics

Bafna, BfB

## Correlation Examples

### Positive Correlation

<table>
<thead>
<tr>
<th></th>
<th>Pepsi</th>
<th>No Pepsi</th>
<th>SUM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coke</td>
<td>50</td>
<td>20</td>
<td>70</td>
</tr>
<tr>
<td>No Coke</td>
<td>10</td>
<td>20</td>
<td>30</td>
</tr>
<tr>
<td>SUM</td>
<td>60</td>
<td>40</td>
<td>100</td>
</tr>
</tbody>
</table>

### Negative Correlation

<table>
<thead>
<tr>
<th></th>
<th>Tea</th>
<th>No Tea</th>
<th>SUM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coffee</td>
<td>20</td>
<td>50</td>
<td>70</td>
</tr>
<tr>
<td>No Coffee</td>
<td>20</td>
<td>10</td>
<td>30</td>
</tr>
<tr>
<td>SUM</td>
<td>40</td>
<td>60</td>
<td>100</td>
</tr>
</tbody>
</table>

### No Correlation

<table>
<thead>
<tr>
<th></th>
<th>Swim</th>
<th>No Swim</th>
<th>SUM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Football</td>
<td>25</td>
<td>25</td>
<td>50</td>
</tr>
<tr>
<td>No Football</td>
<td>25</td>
<td>25</td>
<td>50</td>
</tr>
<tr>
<td>SUM</td>
<td>50</td>
<td>50</td>
<td>100</td>
</tr>
</tbody>
</table>
Correlation Formula

- Joint Probability

- Conditional Probability

- Generalized Correlation Formula
  - Correlation between two data sets, X and Y:
    \[ C(X,Y) = P_{xy} - P_x P_y \]

Testing for correlation

- In the absence of correlation
  \[ \text{Pr}[\text{snp} = \bullet \text{ AND} \text{ Disease}] \approx \text{Pr}[\text{snp} = \bullet \text{]} \text{Pr}[\text{Disease}] \]

\[ \text{Pr[\bullet AND \bullet]} = \frac{2}{8} = 0.125 \]
\[ \text{Pr[\bullet] Pr[\bullet]} = \frac{3}{8} \times \frac{4}{8} = 0.1875 \]

\[ \text{Difference} = 0.06 \]

Bafna, Bib
Testing for correlation

- When correlated

\[ \Pr[\text{snp} = \bullet \text{ AND Disease}] \neq \Pr[\text{snp} = \bullet] \Pr[\text{Disease}] \]

\[ \Pr[\text{AND } \bullet] = \frac{4}{8} = 0.5 \]

\[ \Pr[\bullet] \Pr[\bullet] = \frac{4 \times 4}{8 \times 8} = 0.25 \]

\text{Difference} = 0.25

Chi-Square Test

\( \chi^2 \) Test (\( \chi^2 \) Statistic)

- Evaluates whether an observed distribution in a sample differs from a theoretical distribution (i.e., hypothesis).
- Where \( E_i \) is an expected frequency and \( O_i \) is an observed frequency,

\[ \chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} \]

- The larger \( \chi^2 \), the more likely the variables are related (positively or negatively).

<table>
<thead>
<tr>
<th></th>
<th>Pepsi</th>
<th>No Pepsi</th>
<th>Sum (row)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coke</td>
<td>250 (90)</td>
<td>200 (360)</td>
<td>450</td>
</tr>
<tr>
<td>No Coke</td>
<td>50 (210)</td>
<td>1000 (840)</td>
<td>1050</td>
</tr>
<tr>
<td>Sum (col.)</td>
<td>300</td>
<td>1200</td>
<td>1500</td>
</tr>
</tbody>
</table>
Assigning confidence (validation of correlation)

Pr[ AND ] = \frac{4}{8} = 0.5

Pr[ ] Pr[ ] = \frac{4 \cdot 4}{8 \cdot 8} = 0.25

Difference = 0.25

<table>
<thead>
<tr>
<th>Expected</th>
<th>Observed</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 2</td>
<td>4 0</td>
</tr>
<tr>
<td>2 2</td>
<td>0 4</td>
</tr>
</tbody>
</table>

Bafna, BIB

Assigning confidence (validation of correlation)

\[ \chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i} = 2 \cdot \frac{(4 - 2)^2}{2} + 2 \cdot \frac{(-2)^2}{2} = 8 \]

<table>
<thead>
<tr>
<th>Expected</th>
<th>Observed</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 2</td>
<td>4 0</td>
</tr>
<tr>
<td>2 2</td>
<td>0 4</td>
</tr>
</tbody>
</table>

Bafna, BIB
Assigning confidence (validation of correlation)

\[ \chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i} = 25 \left( \frac{1}{2.5} + \frac{1}{2.5} + \frac{1}{1.5} + \frac{1}{1.5} \right) = 0.53 \]

Expected | Observed
--- | ---
2.5 | 3
2.5 | 2
1.5 | 1
1.5 | 2

Pearson Coefficient

- **Pearson Coefficient**
  - Evaluates correlation between two multi-dimensional data sets.
  - Given two data sets \( X = \{x_1, x_2, \ldots, x_n\} \) and \( Y = \{y_1, y_2, \ldots, y_n\} \),

  \[ r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}} \]

  - Co-variance between \( X \) and \( Y \)
  - Individual variance (standard deviation) of \( X \) and \( Y \)

- If \( r > 0 \), \( X \) and \( Y \) are positively correlated.
- If \( r = 0 \), \( X \) and \( Y \) are independent.
- If \( r < 0 \), \( X \) and \( Y \) are negatively correlated.
Conclusion

- Causal mutations and causal genes can be predicted by sampled mutations.
- We can measure correlation between a sampled mutation and the phenotype using statistical methods.
- Higher correlation indicates that the causal mutations and causal genes are close to the sampled mutation.