Genetic Association
Complex Traits - Multifactorial Inheritance

- Genetic Variants
- Non-genetic factors

Examples:
- Some cancers
- Type 1 diabetes
- Type 2 diabetes
- Alzheimer disease
- Inflammatory bowel disease
- Schizophrenia
- Cleft lip/palate
- Hypertension
- Rheumatoid arthritis
- Asthma
B  Disease gene network
**EPIGENETIC MECHANISMS**

- Development (in utero, childhood)
- Environmental chemicals
- Drugs/Pharmaceuticals
- Aging
- Diet

**HEALTH ENDPOINTS**

- Cancer
- Autoimmune disease
- Mental disorders
- Diabetes

---

**DNA methylation**

Methyl group (an epigenetic factor found in some dietary sources) can tag DNA and activate or repress genes.

**Histone modification**

The binding of epigenetic factors to histone "tails" alters the extent to which DNA is wrapped around histones and the availability of genes in the DNA to be activated.

**Histones** are proteins around which DNA can wind for compaction and gene regulation.
Genetic Association Studies

- Case-control designs
- Family based designs
- Quantitative trait association
  - continuous trait
  - longitudinal trait
Can't use RR, can only use OR because researcher sets the prevalence within the study. Good for rare diseases. In rare diseases, OR approximates RR. In non-rare diseases, the direction of OR and RR are the same, but the actual number obtained for OR and RR are different. You CANNOT obtain a RR for this. It makes no sense to.

RR and OR are both relevant for this. This is sometimes used to test out a new intervention/treatment.
Population Stratification

- **Cases**: Pop 1 (light green) + Pop 2 (dark green)
- **Controls**: Pop 1 (light green) + Pop 2 (dark green)
- **Genotype** (right): TT (white), AT (yellow), AA (green)
Population Structure Example

❖ Cases drawn from males in Africa
❖ Controls drawn from females in America

Differences in allele frequencies may be due to sex, ethnicity, or the disease of interest.

Use cohorts and possibly adjust for population stratification.
Population Stratification

Case

Population 1

Population 2

Control
Confounding

True Risk Factor → Exposure of Interest → Disease

Population Stratification

Ethnicity ← Genotype of Interest → Disease

True Risk Factor ← Genotype of Interest → Disease
The diagram illustrates the prevalence of NIDDM (Non-Insulin Dependent Diabetes Mellitus) in two populations: Full heritage American Indian and Caucasian populations.

### Full heritage American Indian Population
- **Gm3;5,13,14**: + ∼1%  ~99%
- (NIDDM prevalence ∼40%)

### Caucasian Population
- **Gm3;5,13,14**: + ∼66%  ~34%
- (NIDDM prevalence ∼15%)

#### Study without knowledge of genetic background:

<table>
<thead>
<tr>
<th>Gm3;5,13,14 haplotype</th>
<th>Cases</th>
<th>Controls</th>
</tr>
</thead>
<tbody>
<tr>
<td>+</td>
<td>7.8%</td>
<td>29.0%</td>
</tr>
<tr>
<td>-</td>
<td>92.2%</td>
<td>71.0%</td>
</tr>
</tbody>
</table>

Odds ratio 0.27, 95% CI 0.18–0.40

#### Proportion with NIDDM by heritage and marker status

<table>
<thead>
<tr>
<th>Index of Indian heritage</th>
<th>Gm3;5,13,14 haplotype</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>+</td>
</tr>
<tr>
<td>0</td>
<td>17.8%</td>
</tr>
<tr>
<td>4</td>
<td>28.3%</td>
</tr>
<tr>
<td>8</td>
<td>35.9%</td>
</tr>
</tbody>
</table>
Family Based Designs: Transmission Disequilibrium Test (TDT)

Family based association designs aim to avoid the potential confounding effects of population stratification by using the parents as controls for the case, which is their affected offspring.

Application of McNemar’s Test

<table>
<thead>
<tr>
<th></th>
<th>Non-transmitted allele</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transmitted allele</td>
<td>$M_1$</td>
</tr>
<tr>
<td>$M_1$</td>
<td>a</td>
</tr>
<tr>
<td>$M_2$</td>
<td>c</td>
</tr>
<tr>
<td>Total</td>
<td>a + c</td>
</tr>
</tbody>
</table>
Aldehyde dehydrogenase 2 (ALDH2)

- acute alcohol intoxication
- Most Caucasians have two major isozymes, while approximately 50% of Asians have one normal copy of the ALDH2 gene and one mutant copy
Types of Population Association Studies

- Candidate gene
- Candidate polymorphism
- Fine mapping
- Genome-wide association studies
Hapmap

- 30 trios from Nigeria
- 30 trios from US with northern/western European ancestry
- 44 unrelated individuals from Tokyo, Japan
- 45 unrelated Han Chinese from Beijing, China

Data is publicly available: www.hapmap.org
Data Quality

• Hardy-Weinberg equilibrium
  - imbreeding, population stratification, selection
  - symptom of the disease association
  - miscall heterozygotes as homozygotes
  ➡ discard at level $\alpha=10^{-3}$ or $\alpha=10^{-4}$

• Missing genotype data

• single/multiple imputation (be weary of missing not at random; case samples are often collected and analyzed separately from controls)
  - remove the sample

• Minor allele frequency
  - 5% is a common threshold

• remove individuals with high and low levels of heterozygotes

• Identify related individuals and population substructure
(a) Full genotype table for a general genetic model

<table>
<thead>
<tr>
<th></th>
<th>AA</th>
<th>AB</th>
<th>BB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cases</td>
<td>a</td>
<td>b</td>
<td>c</td>
</tr>
<tr>
<td>Controls</td>
<td>d</td>
<td>e</td>
<td>f</td>
</tr>
</tbody>
</table>

(b) Dominant model: allele B increases risk

<table>
<thead>
<tr>
<th></th>
<th>AA+BB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cases</td>
<td>a+b+c</td>
</tr>
<tr>
<td>Controls</td>
<td>d+e+f</td>
</tr>
</tbody>
</table>

(c) Recessive model: two copies of allele B required for increased risk

<table>
<thead>
<tr>
<th></th>
<th>AA+AB</th>
<th>BB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cases</td>
<td>a+b</td>
<td>c</td>
</tr>
<tr>
<td>Controls</td>
<td>d+e</td>
<td>f</td>
</tr>
</tbody>
</table>
(d) Multiplicative model: $r$-fold increased risk for AB, $r^2$ increased risk for BB. Analysed by allele, not by genotype

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cases</td>
<td>$2a + b$</td>
<td>$b + 2c$</td>
</tr>
<tr>
<td>Controls</td>
<td>$2d + e$</td>
<td>$e + 2f$</td>
</tr>
</tbody>
</table>

(e) Additive model: $r$-fold increased risk for AB, $2r$ increased risk for BB. Genotypes analysed by Armitage’s test for trend

<table>
<thead>
<tr>
<th></th>
<th>AA</th>
<th>AB</th>
<th>BB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cases</td>
<td>$a$</td>
<td>$b$</td>
<td>$c$</td>
</tr>
<tr>
<td>Controls</td>
<td>$d$</td>
<td>$e$</td>
<td>$f$</td>
</tr>
</tbody>
</table>
Cochran-Armitage test for trend

<table>
<thead>
<tr>
<th></th>
<th>Genotype aa</th>
<th>Genotype Aa</th>
<th>Genotype AA</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cases</td>
<td>$r_0$</td>
<td>$r_1$</td>
<td>$r_2$</td>
<td>$R$</td>
</tr>
<tr>
<td>Controls</td>
<td>$s_0$</td>
<td>$s_1$</td>
<td>$s_2$</td>
<td>$S$</td>
</tr>
<tr>
<td>Sum</td>
<td>$n_0$</td>
<td>$n_1$</td>
<td>$n_2$</td>
<td>$N$</td>
</tr>
</tbody>
</table>

```r
anova(lm(freq ~ score, weights = w))
library(GeneticsBase)
ArmitageTest
prop.trend.test
```
> ArmitageTest(x=c(2,2,1,1,0,0,0,0,0,0), mem=c(1,1,1,1,1,0,0,0,0,0))

<table>
<thead>
<tr>
<th>stat</th>
<th>pvalue</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.62500000</td>
<td>0.01770607</td>
</tr>
</tbody>
</table>

> prop.trend.test(c(2,2,1),c(2,2,6))

Chi-squared Test for Trend in Proportions

data:  c(2, 2, 1) out of c(2, 2, 6) , using scores: 1 2 3
X-squared = 5.625, df = 1, p-value = 0.01771
Which Test Should I Use?

- Additive
- Dominant
- Recessive

  ➡ maximum test statistic of all three (what about overdominance?)

- Fisher

  ➡ put more weight on the additive when MAF is small, otherwise more weight on Fisher
Bayesian Framework

- Frequentist approach: p-values
  - significance depends on $H_1$ and the power of the test but must be inferred from other quantities such as MAF, etc.

- Bayesian approach: posterior probability
  - alleviates the limitations of p-values at the cost of some additional modeling assumptions
Three Steps

• Prior probability for $H_1$
  - $\pi$ can vary across SNPs, vary depending on MAF, or same for all SNPs (The probability for $H_0$ is then $1 - \pi$); e.g. $10^{-4}$ to $10^{-6}$

• Compute a Bayes factor (BF) for each SNP
  - ratio between the probabilities of the data under $H_1$ and under $H_0$

• Calculate the posterior odds
  - $PO = BF \times \pi/(1 - \pi)$
  - $PPA = PO/(1 + PO)$
Posterior Probability of Association (PPA)

- Interpreted directly as a probability, irrespective of power, sample size, and number of tests/SNPs
- Can be computed directly from BF for any given $\pi$
- Same ranking as BF when $\pi$ is constant
- Has natural connections to the frequentist FDR
Bayes Factor

Let \( \theta_{\text{het}} \) denote the log-odds between the heterozygote and the common homozygote.

Let \( \theta_{\text{hom}} \) denote the log odds between the rare and common homozygotes. The null and alternative hypotheses:

\[
H_0 : \theta_{\text{het}} = \theta_{\text{hom}} = 0 \\
H_1 : \theta_{\text{het}} = t_1, \quad \theta_{\text{hom}} = t_2
\]

\[
BF = \frac{P(\text{data}|\theta_{\text{het}} = t_1, \theta_{\text{hom}} = t_2)}{P(\text{data}|\theta_{\text{het}} = \theta_{\text{hom}} = 0)}
\]

Then average (integrate) over \( t_1 \) and \( t_2 \) with \( \Pi \).
Parameter Distributions

- Additive model: $\theta_{\text{het}} = 2\theta_{\text{hom}} \quad \theta \sim \mathcal{N}(0, \sigma^2)$

- Strict dominant or recessive model:

  $\theta_{\text{het}} = \theta_{\text{hom}} \quad \text{or} \quad \theta_{\text{het}} = 0$

- Nuisance parameters (intercept, covariates)
## Bayesian vs Frequentist on WTCCC Data

| Trait | SNP*    | p-values‡ | log\(_{10}\)(BF)§ | PPA  
|-------|---------|-----------|------------------|------
|       |         | Trend test| General test     | π = 10^{-4} | π = 10^{-5} |
| BD    | rs420259| 2.2 × 10^{-4} | 6.3 × 10^{-5} | 4.1 | 0.56 | 0.11 |
| CD    | rs9858542| 7.7 × 10^{-7} | 3.6 × 10^{-5} | 4.7 | 0.83 | 0.33 |
| T2D   | rs9939609| 5.2 × 10^{-8} | 1.9 × 10^{-7} | 5.3 | 0.95 | 0.67 |
| CD    | rs17221417| 9.4 × 10^{-12} | 4.0 × 10^{-11} | 8.9 | 0.999999 | 0.99987 |
| T1D   | rs17696736| 2.2 × 10^{-15} | 1.5 × 10^{-14} | 12.5 | 1.000000 | 1.000000 |
Continuous outcomes: linear regression

ANOVA (2 degrees of freedom)
Regression (1 degree of freedom)
Case-Control: Logistic Regression

- $\text{logit } (\pi) = \log (\pi / (1 - \pi))$
- $\beta_0 = \beta_1 = \beta_2$ -- 2 degrees of freedom
- $\beta_1$ is half-way between $\beta_1$ and $\beta_2$ -- one degree of freedom
- recessive or dominant effects: $\beta_0 = \beta_1$ or $\beta_1 = \beta_2$
- fast score test implementations
Categorical Outcomes

• Un-ordered outcomes: multinomial regression

• Ordered outcomes (e.g. mild, moderate, severe): ordinal multinomial regression
  - proportional odds: the odds of an individual having a disease state in or above a given category is the same for all categories
  - adjacent categories regression: risk of category $k$ relative to $k-1$ is the same for all $k$
Other Considerations

• population stratification
• multiple SNPs
• epistatic and gene-environment interactions
• multiple testing
The World is Adaptive

It is not the strongest of the species that survive, nor the most intelligent, but the one most responsive to change.

Charles Darwin

*On the Origin of Species*
Nov 24, 1859

*On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life.*