Inferring Genetic Variation and Discovering Associations with Phenotypes

BMI/CS 776
www.biostat.wisc.edu/bmi776/
Spring 2022
Daifeng Wang
daifeng.wang@wisc.edu

These slides, excluding third-party material, are licensed under CC BY-NC 4.0 by Mark Craven, Colin Dewey, Anthony Gitter and Daifeng Wang
Outline

• Variation detection
  – Array technologies
  – Whole-genome sequencing

• GWAS and QTL basics
  – Testing SNPs for association
  – Correcting for multiple-testing
Variation detecting technologies

- **Array-based technologies**
  - Relies on hybridization of sample DNA to pre-specified probes
  - Each probe is chosen to measure a single possible variant: SNP, CNV, etc.

- **Sequencing-based technologies**
  - Whole-genome shotgun sequence, usually at low coverage (e.g., 4-8x)
  - Align reads to reference genome: mismatches, indels, etc. indicate variations
  - Long read sequencing
Array-based technologies

- Currently two major players
- Affymetrix Genome-Wide Human SNP Arrays
  - Used for HapMap project, Navigenics service
- Illumina BeadChips
  - Used by 23andMe, deCODEme services
Affymetrix SNP arrays

- Probes for ~900K SNPs
- Another ~900K probes for CNV analysis
- Differential hybridization – one probe for each possible SNP allele

![Diagram of SNP probes and fluorescent tagging](image)
Illumina BeadChips

• OmniExpress+
  – ~900K SNPs (700K fixed, 200 custom)

• Array with probes immediately adjacent to variant location

• Single base extension (like sequencing) to determine base at variant location

The Infinium HD Assay protocol features single-tube sample preparation and whole-genome amplification without PCR or ligation steps, significantly reducing labor and sample-handling errors.
Sequencing-based genotyping

compute $\text{argmax } P(\text{genotype} \mid \text{reads, reference})$ for each genomic position

genotype

sequencing error?

genotype = GA?

reads

reference

ACTCTACGTACGATCGTCGCTACGTGCTAGCTAGTCGCAC

GATCGTCGCTACGTGCTAG
CTCTACGTACGATCGTCGCT
CGATCATCGCTACGTGCTAGCT
TACGATCGTCTCTACGTGC
CTACGTAAGATCATCGCTA
ACGATCATCGCTACGTGCT
CGTACGATCGTCGCTACGT

7
Long read sequencing

- Pacific Biosciences SMRT
- MinION nanopore
- Illumina TruSeq Synthetic

*De novo* assembly of two Swedish genomes reveals missing segments from the human GRCh38 reference and improves variant calling of population-scale sequencing data

Adam Ameur, Huiwen Che, Marcel Martin, Ignas Bunikis, Johan Dahlberg, Ida Höijer, Susana Häggqvist, Francesco Vezzi, Jessica Nordlund, Pall Olason, Lars Feuk, Ulf Gyllensten

doi: https://doi.org/10.1101/267062

— “over 10 Mb of sequences absent from the human GRCh38 reference in each individual”
GWAS jargon

Locus - genetic position on a chromosome, and a single base pair position in the context of SNPs

SNP - a locus (single base pair) that exhibits variation (polymorphism) in a population

Allele (in the context of SNPs) - the alternative forms of a nucleotide at a particular locus

Genotype - the pair of alleles at a locus, one paternal and one maternal

Heterozygous - the two alleles differ at a locus

Homozygous - the two alleles are identical at a locus

Genotyped SNP - we have observed the genotype at a particular SNP, e.g. because the SNP is among the 1 million on the SNP array we used

Ungenotyped SNP - we have not observed the genotype at a particular locus

Causal SNP - a SNP that directly affects the phenotype, e.g. a mutation changes the amino acid sequence of a protein and changes the protein's function in a way that directly affects a biological process

Haplotype - a group of SNPs that are inherited jointly from a parent

Linkage disequilibrium - alleles at multiple loci that exhibit a dependence (nonrandom association)

### GWAS data

<table>
<thead>
<tr>
<th>Individual</th>
<th>Genotype at Position 1</th>
<th>Genotype at Position 2</th>
<th>Genotype at Position 3</th>
<th>...</th>
<th>Genotype at Position $M$</th>
<th>Disease?</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>CC</td>
<td>AG</td>
<td>GG</td>
<td></td>
<td>AA</td>
<td>N</td>
</tr>
<tr>
<td>2</td>
<td>AC</td>
<td>AA</td>
<td>TG</td>
<td></td>
<td>AA</td>
<td>Y</td>
</tr>
<tr>
<td>3</td>
<td>AA</td>
<td>AA</td>
<td>GG</td>
<td></td>
<td>AT</td>
<td>Y</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$N$</td>
<td>AC</td>
<td>AA</td>
<td>TT</td>
<td></td>
<td>AT</td>
<td>N</td>
</tr>
</tbody>
</table>

- $N$ individuals genotyped at $M$ positions
- Disease status (or other phenotype) is measured for each individual
GWAS task

• *Given*: genotypes and phenotypes of individuals in a population
• *Do*: identify which genomic positions are associated with a given phenotype
Can we identify causal SNPs?

- Typically only genotype at 1 million sites
- Humans vary at ~100 million sites
- Unlikely that an associated SNP is causal
- **Tag SNPs**: associated SNPs “tag” blocks of the genome that contain the causal variant

![Diagram of haplotype block]
Direct and indirect associations

Phenotype

direct association
indirect association

direct association (haplotype block)
SNP imputation

- Estimate the ungenotyped SNPs using reference haplotypes

1000 Genomes

SNP array

Nielsen *Nature* 2010
Genotype imputation

- Evaluate the evidence for association at genetic markers that are not directly genotyped
- Increases power of genome-wide association scans
- Useful for combining data from studies that rely on different genotyping platforms

*Genotype imputation.
A pipeline for genotype imputation

- Raw Genotype Calls
  - SNP QC using PLINK
    - Sample level quality control
      - examining call rate
      - Heterozygosity
      - relatedness between genotyped individuals
      - correspondence between sex chromosome genotypes and reported gender
    - Marker level quality control
      - examining call rates
      - excluding low frequency SNPs
      - Strand flip
  - Split data by chromosomes
  - Pre-Phasing using SHAPEIT2/Eagle2
    - haplotypes are estimated for all available samples
  - SNP imputation using IMPUTE2/Minimac3
    - missing alleles are imputed directly onto these phased haplotypes
  - Genotype filtering
  - Convert to SNP dosage
    - Genotype

Wang et al., Science, 2018
Basics of association testing

• Test each site individually for association with a statistical test
  – each site is assigned a \( p \)-value for the null hypothesis that the site is not associated with the phenotype

• Correct for the fact that we are testing multiple hypotheses
Basic genotype test

- Assuming binary phenotype (e.g., disease status)
- Test for significant association with Pearson’s Chi-squared test or Fisher’s Exact Test

<table>
<thead>
<tr>
<th>genotype</th>
<th>AA</th>
<th>AT</th>
<th>TT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disease</td>
<td>40</td>
<td>30</td>
<td>30</td>
</tr>
<tr>
<td>No disease</td>
<td>70</td>
<td>20</td>
<td>10</td>
</tr>
</tbody>
</table>

Chi-squared test $p$-value = 4.1e-5 (2 degrees of freedom)
Fisher’s Exact Test $p$-value = 3.4e-5
Armitage (trend) test

- Can gain more statistical power if we can assume that probability of trait is linear in the number of one of the alleles

Balding Nature Reviews Genetics 2006
Trend test example

<table>
<thead>
<tr>
<th>phenotype</th>
<th>genotype</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AA</td>
</tr>
<tr>
<td>Disease</td>
<td>40</td>
</tr>
<tr>
<td>No disease</td>
<td>70</td>
</tr>
</tbody>
</table>

| Disease proportion | 0.36 | 0.60 | 0.75 |

Trend in Proportions test $p$-value = $8.1 \times 10^{-6}$

(note that this is a smaller $p$-value than from the basic genotype test)
GWAS Versus QTL

• Both associate genotype with phenotype

• GWAS pertains to discrete phenotypes
  – For example, disease status is binary

• QTL pertains to quantitative (continuous) phenotypes
  – Height
  – Gene expression
  – Splicing events
  – Metabolite abundance
Expression QTL (eQTL)

- traits are expression levels of various genes

- Merge expression from all studies

- Filtered out the very lowly-expressed genes by minimum of 50 samples having an FPKM value of at least 0.1

- Quantile normalization

- PEER calculations

- Inverse quantile normalization
A pipeline for eQTLs

- MAF>=0.01
- R2>0.3
- Hardy-Weinberg Equilibrium p<1x10^{-6}

- 65 covariates
  - Top 3 genotyping principal components
  - Probabilistic Estimation of Expression Residuals (PEER) factors
  - Genotyping array platform
  - Gender
  - Disease status

Wang et al., Science, 2018
Challenges of association studies (e.g., GWAS)

- Population structure
- Interacting variants
- Multiple testing
- Interpreting hits
Population structure issues

- If certain populations disproportionately represent cases or controls, then spurious associations may be identified.
Interacting variants

• Most traits are complex: not the result of a single gene or genomic position
• Ideally, we’d like to test subsets of variants for associations with traits
  – But there are a huge number of subsets!
  – Multiple testing correction will likely result in zero association calls
• Area of research
  – Only test carefully selected subsets
  – Bayesian version: put prior on subsets
Multiple testing

• In the genome-age, we have the ability to perform large numbers of statistical tests simultaneously
  – SNP associations (~1 million)
  – Gene differential expression tests (~ 20 thousand)

• Do traditional $p$-value thresholds apply in these cases?
Expression in BRCA1 and BRCA2 Mutation-Positive Tumors

- 7 patients with BRCA1 mutation-positive tumors vs. 7 patients with BRCA2 mutation-positive tumors
- 5631 genes assayed

Expression in BRCA1 and BRCA2 Mutation-Positive Tumors

- Key question: which genes are differentially expressed in these two sets of tumors?

- Methodology: for each gene, use a statistical test to assess the hypothesis that the expression levels differ in the two sets.
Hypothesis testing

• Consider two competing hypotheses for a given gene
  – *null hypothesis*: the expression levels in the first set come from the same distribution as the levels in the second set
  – *alternative hypothesis*: they come from different distributions

• First calculate a test statistic for these measurements, and then determine its *p*-value

• *p*-value: the probability of observing a test statistic that is as extreme or more extreme than the one we have, assuming the null hypothesis is true
Calculating a \( p \)-value

1. Calculate test statistic (e.g. T statistic)

\[
T = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}
\]

where

\[
\bar{x}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} x_{ij}
\]

\[
s_j^2 = \frac{1}{n_j - 1} \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2
\]

2. See how much mass in null distribution with value this extreme or more

If test statistic is here, \( p = 0.034 \)
Multiple testing problem

• If we’re testing one gene, the $p$-value is a useful measure of whether the variation of the gene’s expression across two groups is significant

• Suppose that most genes are not differentially expressed

• If we’re testing 5000 genes that don’t have a significant change in their expression (i.e. the null hypothesis holds), we’d still expect about 250 of them to have $p$-values $\leq 0.05$

• Can think of $p$-value as the *false positive rate* over null genes
Family-wise error rate

• One way to deal with the multiple testing problem is to control the probability of rejecting at least one null hypothesis when all genes are null

• This is the family-wise error rate (FWER)

• Suppose you tested 5000 null genes and predicted that all genes with $p$-values $\leq 0.05$ were differentially expressed

$$FWER = 1 - (1 - 0.05)^{5000} \approx 1$$

– you are guaranteed to be wrong at least once!
– above assumes tests are independent
Bonferroni correction

• Simplest approach
• Choose a \( p \)-value threshold \( \beta \) such that the FWER is \( \leq \alpha \)

\[
\alpha = 1 - (1 - \beta)^g
\]

• where \( g \) is the number of genes (tests)

\[
\text{for } \beta g \ll 1, \quad \beta \approx \frac{\alpha}{g}
\]

• For \( g=5000 \) and \( \alpha=0.05 \) we set a \( p \)-value threshold of \( \beta=1e-5 \)
Loss of power with FWER

• FWER, and Bonferroni in particular, reduce our power to reject null hypotheses
  – As $g$ gets large, $p$-value threshold gets very small
• For expression analysis, FWER and false positive rate are not really the primary concern
  – We can live with false positives
  – We just don’t want too many of them relative to the total number of genes called significant
The False Discovery Rate

[Benjamini & Hochberg ‘95; Storey & Tibshirani ‘02]

<table>
<thead>
<tr>
<th>gene</th>
<th>p-value</th>
<th>rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>0.0001</td>
<td>1</td>
</tr>
<tr>
<td>F</td>
<td>0.001</td>
<td>2</td>
</tr>
<tr>
<td>G</td>
<td>0.016</td>
<td>3</td>
</tr>
<tr>
<td>J</td>
<td>0.019</td>
<td>4</td>
</tr>
<tr>
<td>I</td>
<td>0.030</td>
<td>5</td>
</tr>
<tr>
<td>B</td>
<td>0.052</td>
<td>6</td>
</tr>
<tr>
<td>A</td>
<td>0.10</td>
<td>7</td>
</tr>
<tr>
<td>D</td>
<td>0.35</td>
<td>8</td>
</tr>
<tr>
<td>H</td>
<td>0.51</td>
<td>9</td>
</tr>
<tr>
<td>E</td>
<td>0.70</td>
<td>10</td>
</tr>
</tbody>
</table>

- Suppose we pick a threshold, and call genes above this threshold “significant”
- The false discovery rate is the expected fraction of these that are mistakenly called significant (i.e. are truly null)
The False Discovery Rate

<table>
<thead>
<tr>
<th></th>
<th>Called significant</th>
<th>Called not significant</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Null true</td>
<td>$F$</td>
<td>$m_0 - F$</td>
<td>$m_0$</td>
</tr>
<tr>
<td>Alternative true</td>
<td>$T$</td>
<td>$m_1 - T$</td>
<td>$m_1$</td>
</tr>
<tr>
<td>Total</td>
<td>$S$</td>
<td>$m - S$</td>
<td>$m$</td>
</tr>
</tbody>
</table>

- **false positives (false discoveries)**
- **true positives**
- **total significant at threshold**
- **features (genes)**

Storey & Tibshirani *PNAS* 100(16), 2002
The False Discovery Rate

\[ F(t) = \#\{ \text{null } p_i \leq t; i = 1 \ldots m \} \]

\[ S(t) = \#\{ p_i \leq t; i = 1 \ldots m \} \]

\[ FDR(t) = E \left[ \frac{F(t)}{S(t)} \right] \approx \frac{E[F(t)]}{E[S(t)]} \]

<table>
<thead>
<tr>
<th>gene</th>
<th>p-value</th>
<th>rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>0.0001</td>
<td>1</td>
</tr>
<tr>
<td>F</td>
<td>0.001</td>
<td>2</td>
</tr>
<tr>
<td>G</td>
<td>0.016</td>
<td>3</td>
</tr>
<tr>
<td>J</td>
<td>0.019</td>
<td>4</td>
</tr>
<tr>
<td>I</td>
<td>0.030</td>
<td>5</td>
</tr>
<tr>
<td>B</td>
<td>0.052</td>
<td>6</td>
</tr>
<tr>
<td>A</td>
<td>0.10</td>
<td>7</td>
</tr>
<tr>
<td>D</td>
<td>0.35</td>
<td>8</td>
</tr>
<tr>
<td>H</td>
<td>0.51</td>
<td>9</td>
</tr>
<tr>
<td>E</td>
<td>0.70</td>
<td>10</td>
</tr>
</tbody>
</table>

# genes

p-value threshold
The False Discovery Rate

- To compute the FDR for a threshold $t$, we need to estimate $E[F(t)]$ and $E[S(t)]$

$$FDR(t) = E\left[ \frac{F(t)}{S(t)} \right] \approx \frac{E[F(t)]}{E[S(t)]}$$

estimate by the observed $S(t)$

$S(t) = \# \{ p_i \leq t; i = 1 \ldots m \}$

$F(t) = \# \{ \text{null } p_i \leq t; i = 1 \ldots m \}$

- So how can we estimate $E[F(t)]$?
Estimating $E[F(t)]$

- Two approaches we’ll consider
  - Benjamini-Hochberg (BH)
  - Storey-Tibshirani ($q$-value)

- Different assumptions about null features ($m_0$)
Benjamini-Hochberg

- Suppose the fraction of genes that are truly null is very close to 1 so $m_0 \approx m$
- Then
  \[ E[F(t)] = E[\#\{\text{null } p_i \leq t; i = 1 \ldots m\}] \approx mt \]
- Because \( p \)-values are uniformly distributed over \([0,1] \) under the null model
- Suppose we choose a threshold \( t \) and observe that \( S(t) = k \)

\[
FDR(t) \approx \frac{E[F(t)]}{S(t)} = \frac{mt}{k}
\]
Benjamini-Hochberg

• Suppose we want $FDR \leq \alpha$
• Observation:

$FDR(t) \leq \alpha$

$$\frac{mt}{k} \leq \alpha$$

$$t \leq \frac{k}{m} \alpha$$
Benjamini-Hochberg

• Algorithm to obtain FDR ≤ α
• Sort the \( p \)-values of the genes so that they are in increasing order

\[
P_{(1)} \leq P_{(2)} \ldots \leq P_{(m)}
\]

• Select the largest \( k \) such that

\[
P_{(k)} \leq \frac{k}{m} \alpha
\]

• where we use \( P_{(k)} \) as the \( p \)-value threshold \( t \)
What fraction of the genes are truly null?

- Consider the $p$-value histogram from Hedenfalk et al.
  - includes both null and alternative genes
  - but we expect null $p$-values to be uniformly distributed

\[
\pi_0 = \frac{m_0}{m}
\]

\[
\hat{\pi}_0(\lambda) = \frac{\#\{p_i > \lambda; i = 1 \ldots m\}}{m(1 - \lambda)}
\]

Storey & Tibshirani *PNAS* 100(16), 2002
Storey & Tibshirani approach

$$FDR(t) \approx \frac{\hat{\pi}_0 \times m \times t}{\# \{p_i \leq t\}}$$

<table>
<thead>
<tr>
<th>gene</th>
<th>p-value</th>
<th>rank</th>
<th>q-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>0.0001</td>
<td>1</td>
<td>0.0010</td>
</tr>
<tr>
<td>F</td>
<td>0.001</td>
<td>2</td>
<td>0.0050</td>
</tr>
<tr>
<td>G</td>
<td>0.016</td>
<td>3</td>
<td>0.0475</td>
</tr>
<tr>
<td>J</td>
<td>0.019</td>
<td>4</td>
<td>0.0475</td>
</tr>
<tr>
<td>I</td>
<td>0.030</td>
<td>5</td>
<td>0.0600</td>
</tr>
<tr>
<td>B</td>
<td>0.052</td>
<td>6</td>
<td>0.0867</td>
</tr>
<tr>
<td>A</td>
<td>0.10</td>
<td>7</td>
<td>0.1430</td>
</tr>
<tr>
<td>D</td>
<td>0.35</td>
<td>8</td>
<td>0.4380</td>
</tr>
<tr>
<td>H</td>
<td>0.51</td>
<td>9</td>
<td>0.5670</td>
</tr>
<tr>
<td>E</td>
<td>0.70</td>
<td>10</td>
<td>0.7000</td>
</tr>
</tbody>
</table>

$$\hat{q}(p_i) = \min_{t \geq p_i} FDR(t)$$

pick minimum FDR for all greater thresholds
**q-value example for gene J**

\[
m = 20 \quad t = 0.019
\]

\[
\hat{\pi}_0 = 0.5 \quad \#\{p_i \leq t\} = 4
\]

\[
FDR(t) \approx \frac{\hat{\pi}_0 \times m \times t}{\#\{p_i \leq t\}}
\]

\[
= \frac{0.5 \times 20 \times 0.019}{4} = 0.0475
\]

<table>
<thead>
<tr>
<th>gene</th>
<th>p-value</th>
<th>rank</th>
<th>q-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>0.0001</td>
<td>1</td>
<td>0.0010</td>
</tr>
<tr>
<td>F</td>
<td>0.001</td>
<td>2</td>
<td>0.0050</td>
</tr>
<tr>
<td>G</td>
<td>0.016</td>
<td>3</td>
<td>0.0475</td>
</tr>
<tr>
<td>J</td>
<td>0.019</td>
<td>4</td>
<td>0.0475</td>
</tr>
<tr>
<td>I</td>
<td>0.030</td>
<td>5</td>
<td>0.0600</td>
</tr>
<tr>
<td>B</td>
<td>0.052</td>
<td>6</td>
<td>0.0867</td>
</tr>
<tr>
<td>A</td>
<td>0.10</td>
<td>7</td>
<td>0.1430</td>
</tr>
<tr>
<td>D</td>
<td>0.35</td>
<td>8</td>
<td>0.4380</td>
</tr>
<tr>
<td>H</td>
<td>0.51</td>
<td>9</td>
<td>0.5670</td>
</tr>
<tr>
<td>E</td>
<td>0.70</td>
<td>10</td>
<td>0.7000</td>
</tr>
</tbody>
</table>

\[
\hat{q}(p_i) = \min_{t \geq p_i} FDR(t)
\]

In this case, already have minimum FDR for all greater thresholds.
q-values vs. p-values for Hedenfalk et al.

- p-value of 0.05 implies that 5% of all tests will result in false positives
- q-value of 0.05 implies that 5% of significant tests will result in false positives

Storey & Tibshirani *PNAS* 100(16), 2002
FDR summary

• In many high-throughput experiments, we want to know what is different across two sets of conditions/individuals (e.g. which genes are differentially expressed)

• Because of the multiple testing problem, \( p \)-values may not be so informative in such cases

• FDR, however, tells us which fraction of significant features are likely to be null

• \( q \)-values based on the FDR can be readily computed from \( p \)-values (see Storey’s R package qvalue)