

# Technology and methodology for inferring genetic variation and discovering associations with phenotypes

BMI/CS 776

[www.biostat.wisc.edu/bmi776/](http://www.biostat.wisc.edu/bmi776/)

Colin Dewey

[cdewey@biostat.wisc.edu](mailto:cdewey@biostat.wisc.edu)

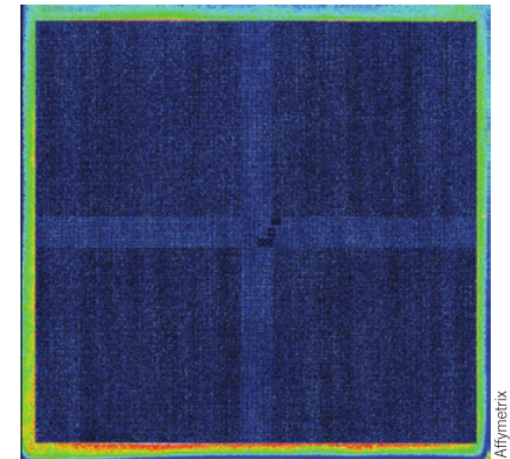
Spring 2015

# Outline

- Variation detection
  - Array technologies
  - Whole-genome sequencing
- The basics of GWAS
  - 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



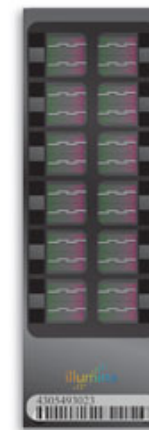
Affymetrix SNP chip



Illumina HiSeq sequencer

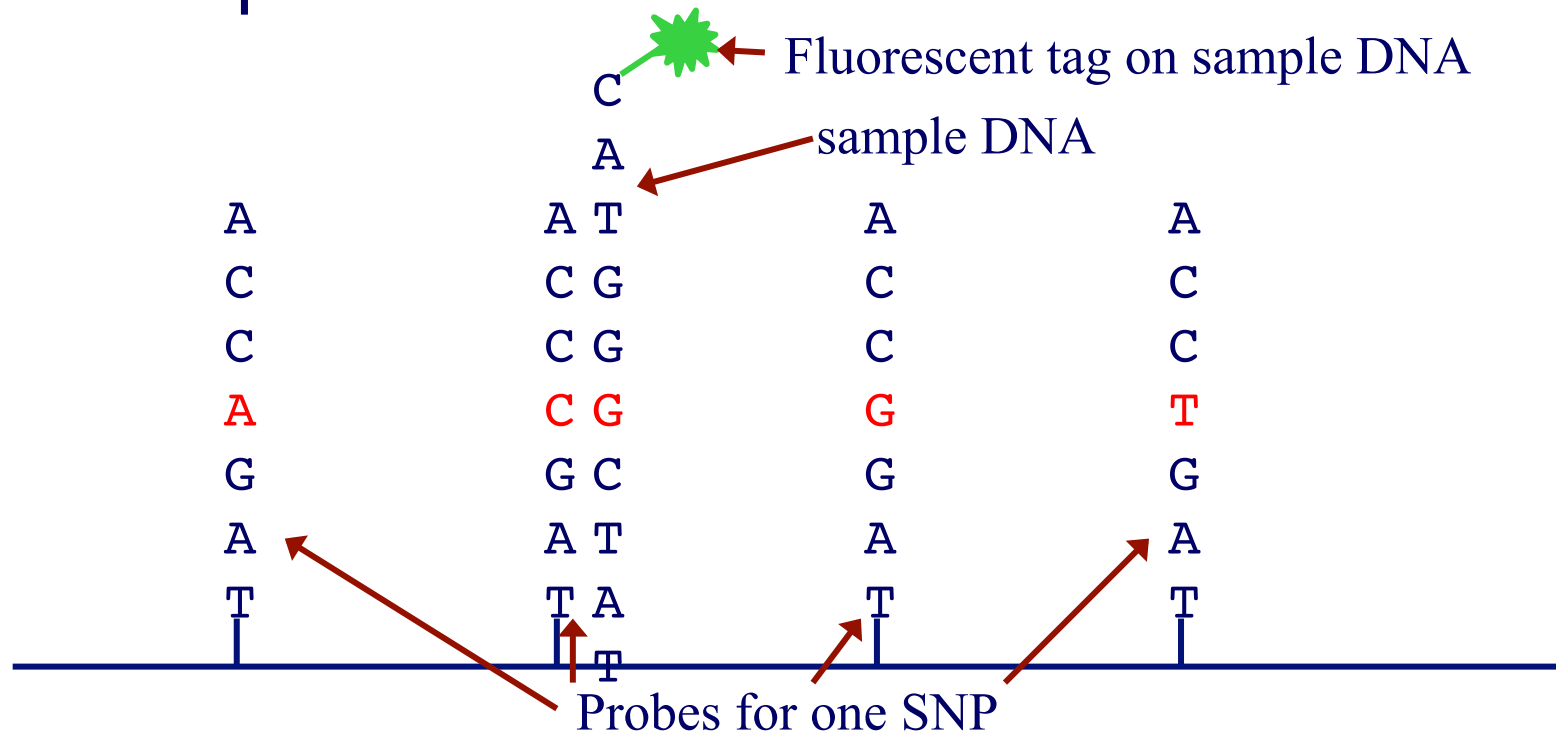
# Array-based technologies

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



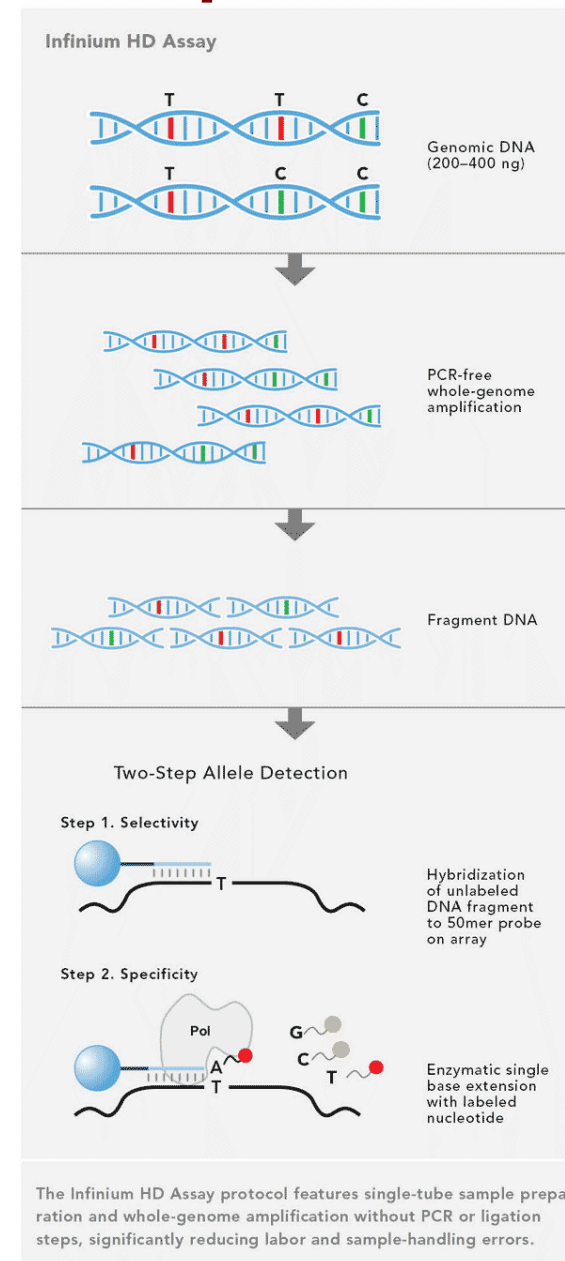
# Affymetrix SNP arrays

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



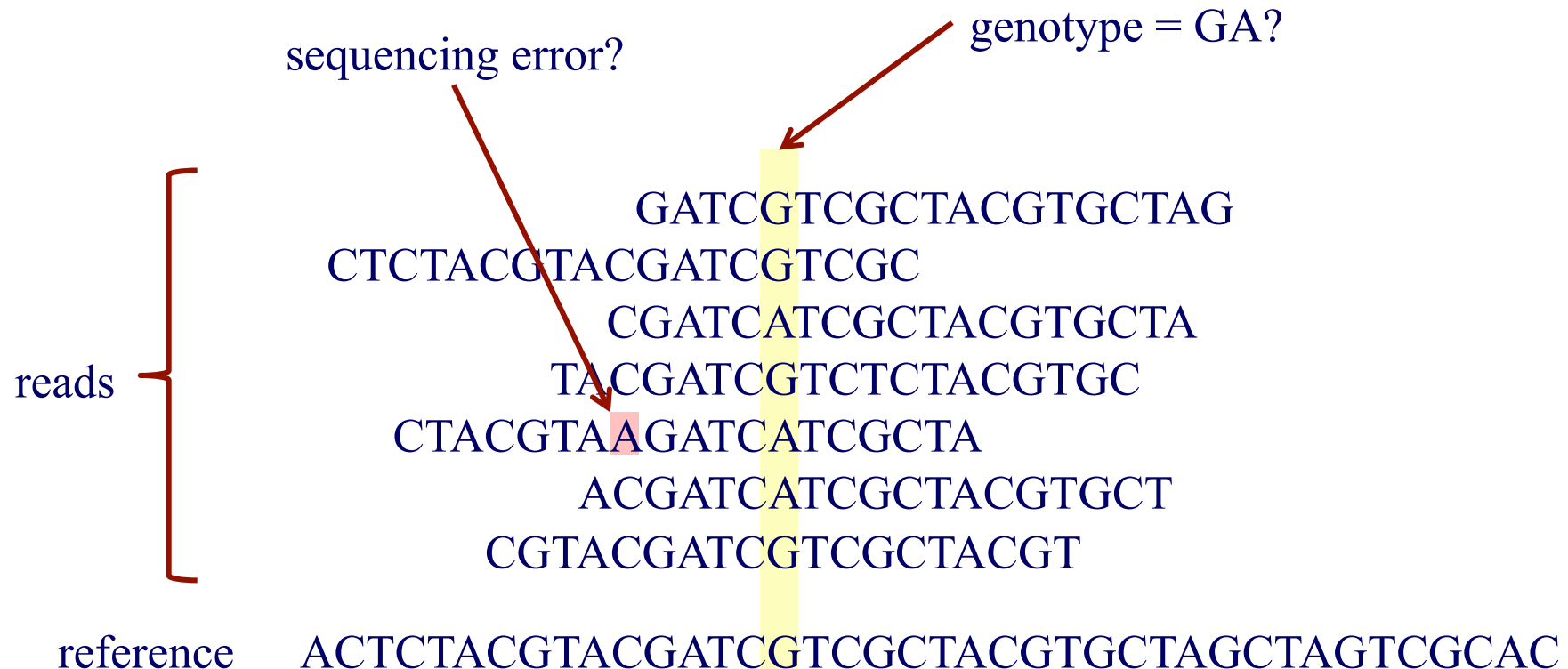
# Illumina BeadChips

- OmniExpress+
  - ~900K SNPs (700K fixed, 200 custom)
- Array with probes immediately adjacent to variant location
- One base extension (like sequencing) to determine base at variant location



# Sequencing-based genotyping

compute  $\arg \max_{genotype} P(genotype | reads, reference)$  for each genomic position



# GWAS data

Individual	Genotype at Position 1	Genotype at Position 2	Genotype at Position 3	...	Genotype at Position M	Disease?
1	CC	AG	GG		AA	N
2	AC	AA	TG		AA	Y
3	AA	AA	GG		AT	Y
...						
N	AC	AA	TT		AT	N

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

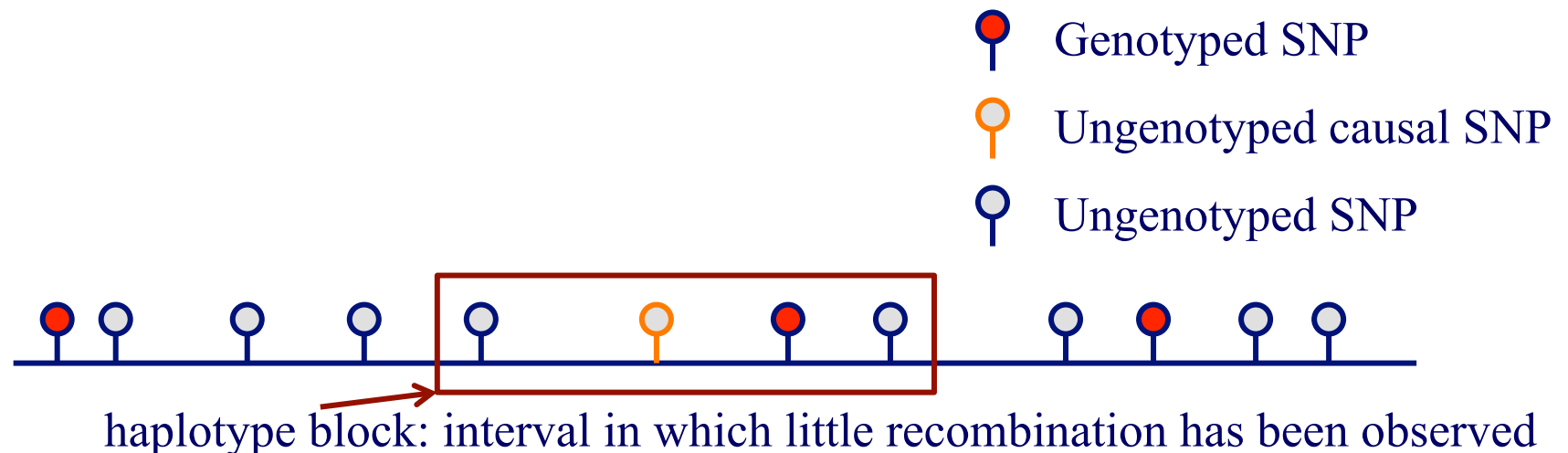


# GWAS task

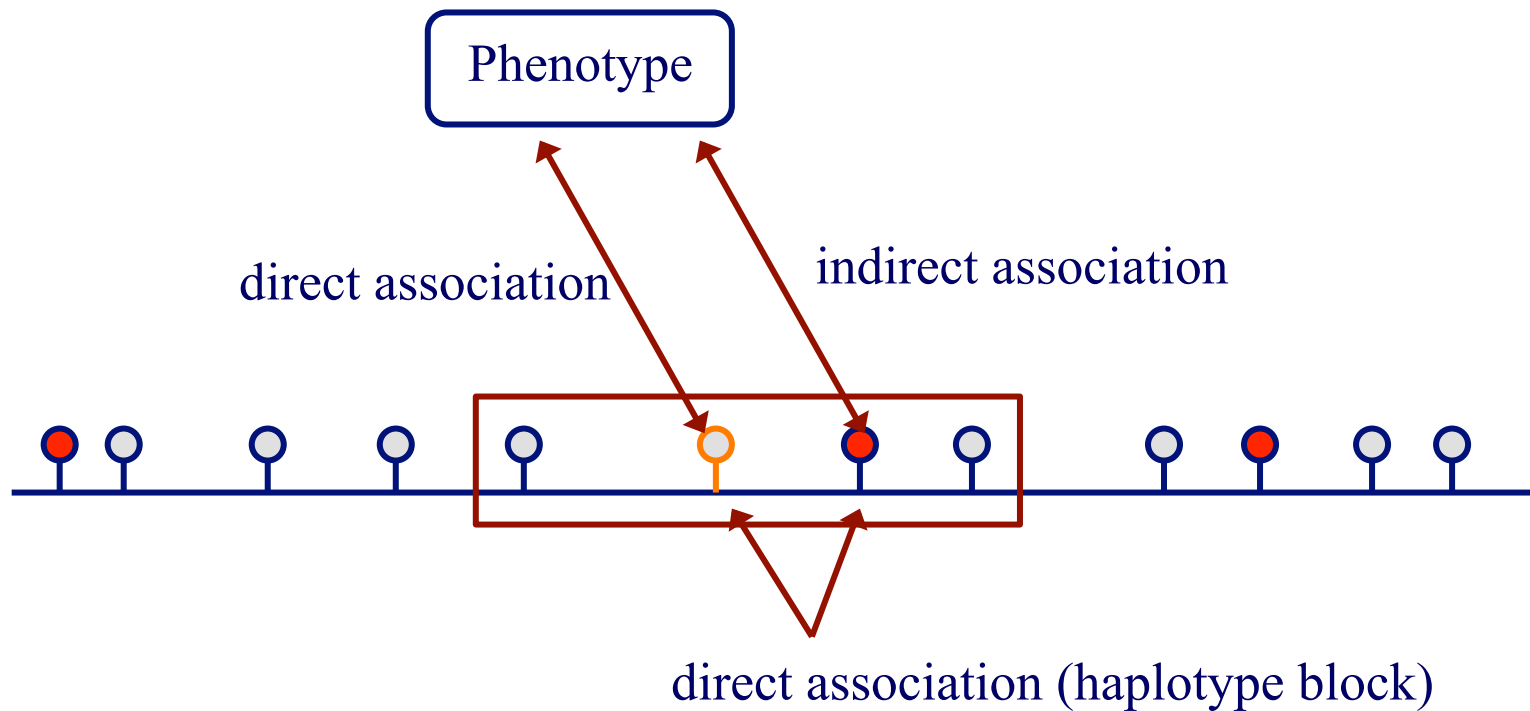
- *Given*: genotypes and phenotypes of individuals in a population
- *Do*: Identify which genomic positions are associated with a given phenotype

# Can we identify causal SNPs?

- Typically only genotype at 1 million sites
- Humans vary at more than 10 million sites
- Unlikely that an associated SNP is causal
- “Tag SNPs”: however, associated SNPs “tag” blocks of the genome that contain the causal variant



# Direct and indirect associations



# 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/no disease)
- Test for significant association with Pearson's Chi-square test or Fisher's Exact Test

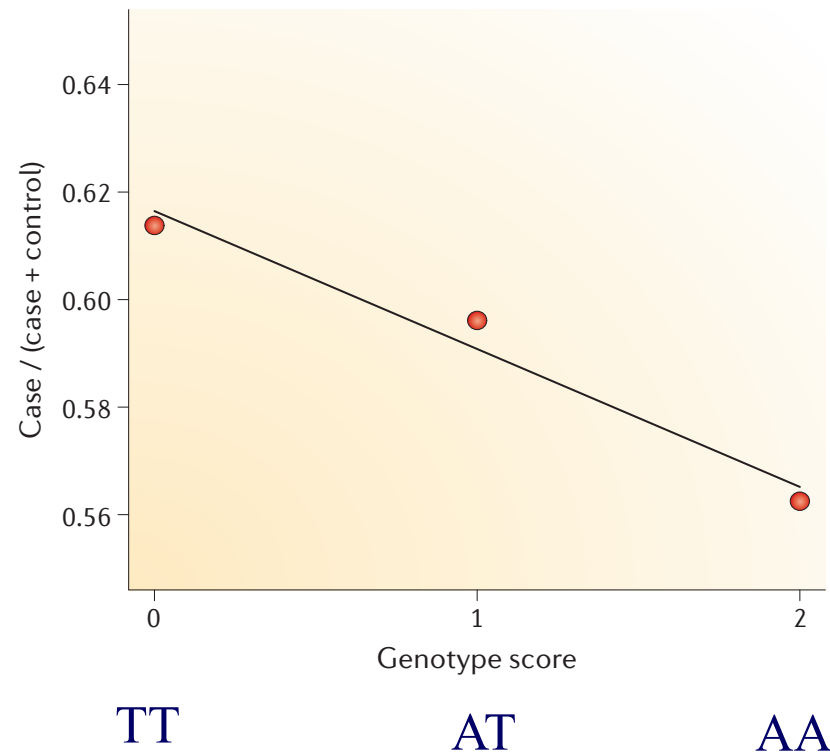
		genotype		
		AA	AT	TT
phenotype	Disease	40	30	30
	No disease	70	20	10

Chi-square test p-value =  $4.1 \times 10^{-5}$  (2 degrees of freedom)

Fisher's exact test p-value =  $3.4 \times 10^{-5}$

# 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



# Trend test example

		genotype		
		AA	AT	TT
phenotype	Disease	40	30	30
	No disease	70	20	10
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)

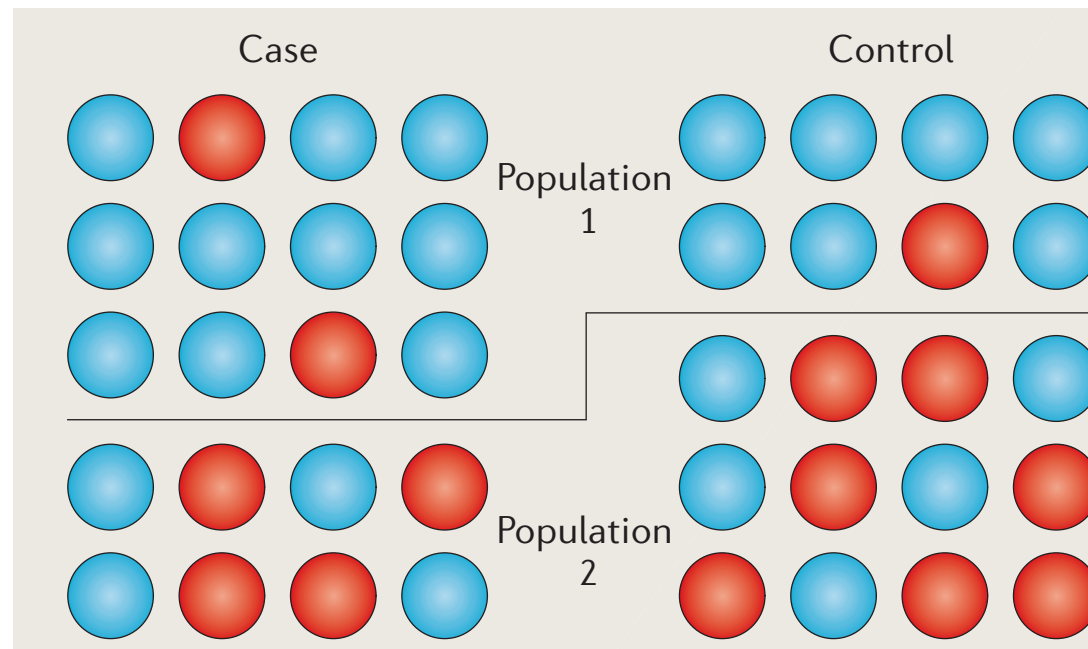
# GWAS Challenges

- Population structure
- Multiple testing
- Interacting variants



# Population structure issues

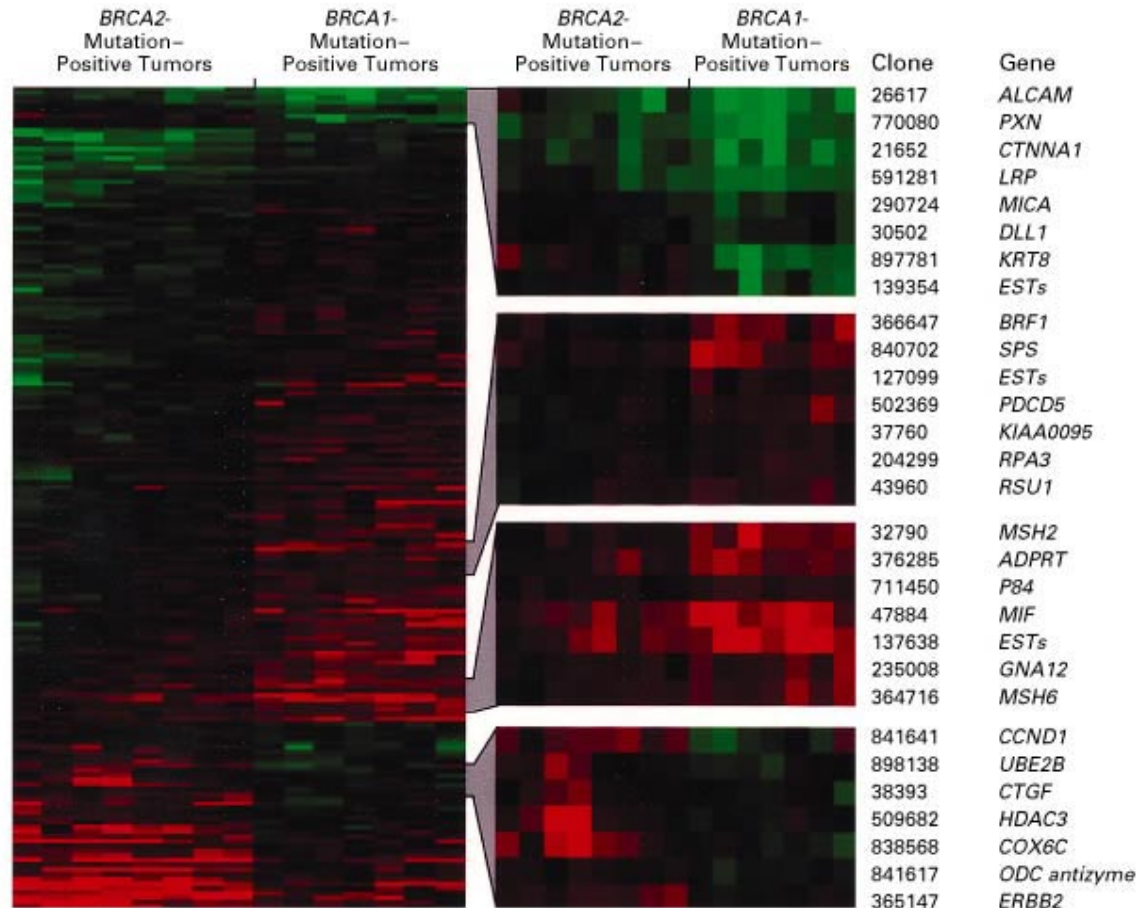
- If certain populations disproportionately represent cases or controls, then spurious associations may be identified



# 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 (~ 50 thousand)
- Do traditional p-value thresholds apply in these cases?

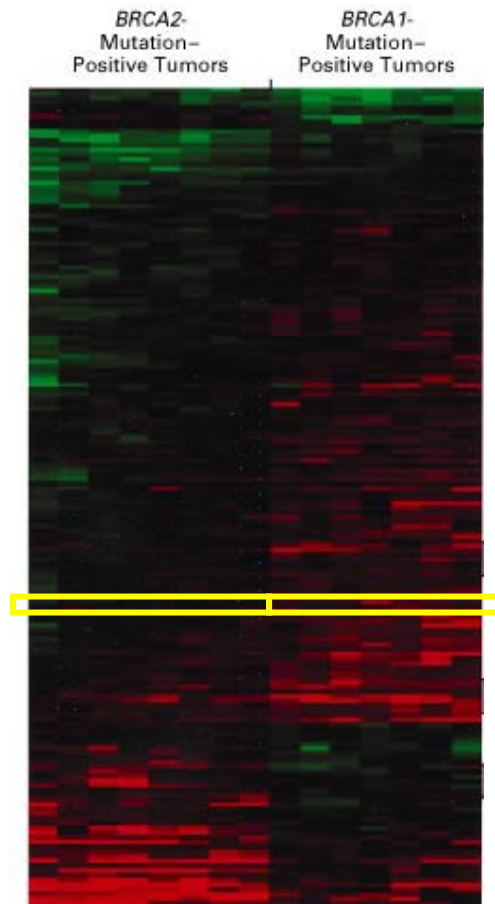
# Expression in BRCA1 and BRCA2 Mutation-Positive Tumors



Hedenfalk et al., *New England Journal of Medicine* 344:539-548, 2001.

- 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
- we 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)

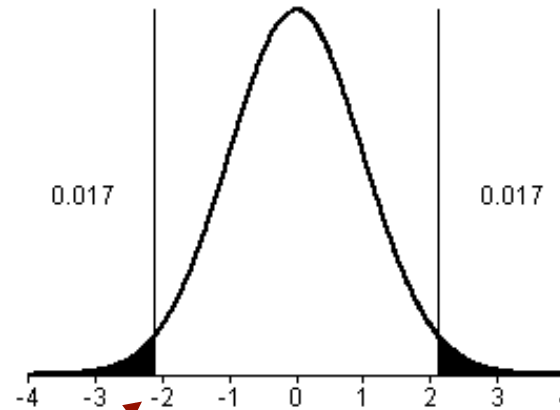


$$T = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2 + s_2^2}{n}}}$$

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,  $p = 0.034$

# The 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 (this is the typical situation)
- 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 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  $1e-5$

# 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]

gene	$p$ -value	rank
C	0.0001	1
F	0.001	2
G	0.016	3
J	0.019	4
I	0.030	5
B	0.052	6
A	0.10	7
D	0.35	8
H	0.51	9
E	0.70	10

- 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

gene	$p$ -value	rank
C	0.0001	1
F	0.001	2
G	0.016	3
J	0.019	4
I	0.030	5
B	0.052	6
A	0.10	7
D	0.35	8
H	0.51	9
E	0.70	10

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

 # genes

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

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

# 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  $S(t)$

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

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

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

# Benjamini-Hochberg

- Suppose the fraction of genes that are truly null is very close to 1
- Then

$$E[F(t)] = E[\#\{\text{null } p_i \leq t; i = 1 \dots m\}] \approx mt$$

- 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 procedure

- Suppose we want the  $\text{FDR} \leq \alpha$
- Sort the  $p$ -values of your genes so that they are in increasing order

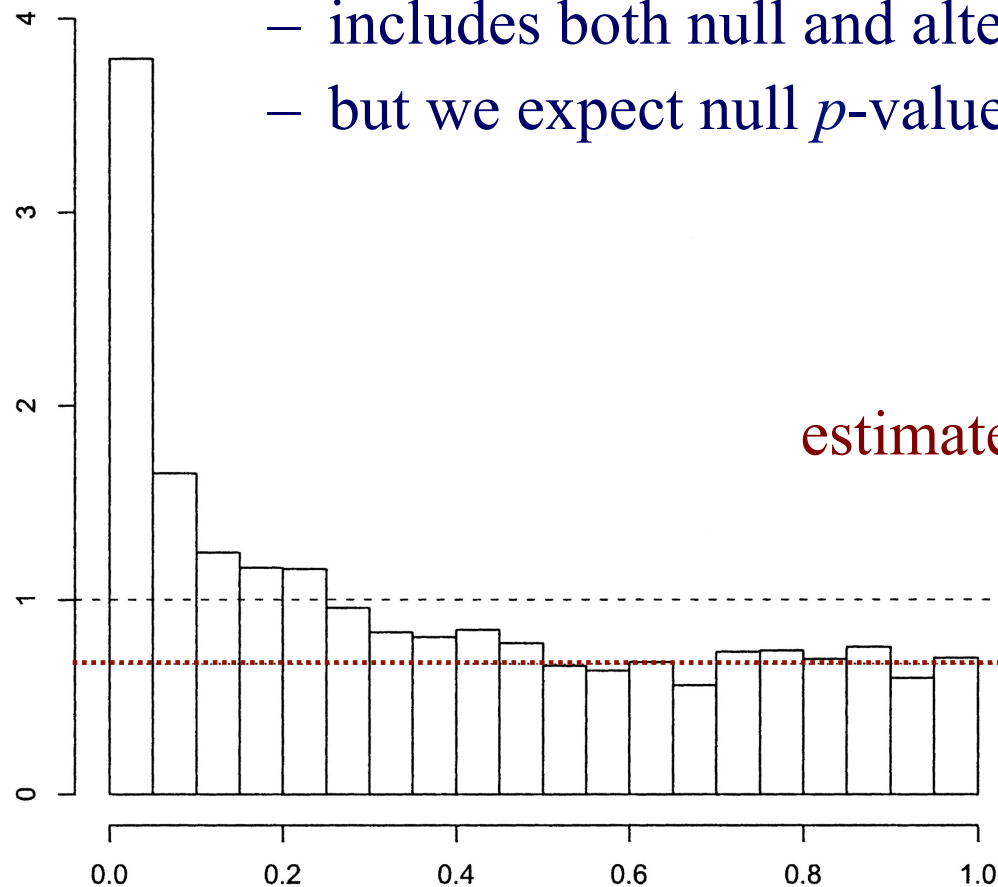
$$P_{(1)} \leq P_{(2)} \leq \dots \leq P_{(m)}$$

- Select the largest  $k$  such that

$$P_{(k)} \leq \frac{k}{m} \alpha$$

# What Fraction of the Genes are Truly Null?

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



estimated proportion of null  $p$ -values

$$\hat{\pi}_0(\lambda) = \frac{\#\{p_i > \lambda; i = 1 \dots m\}}{m(1 - \lambda)}$$

Figure from Storey & Tibshirani *PNAS* 100(16), 2002.



# Storey & Tibshirani approach

estimated proportion of null  $p$ -values  $\hat{\pi}_0$   $\times$   $m$   $\times$   $t$   $\div$   $\# \{p_i \leq t\}$   $\#$  genes

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

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

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

pick minimum FDR for all greater thresholds

# $q$ -values vs. $p$ -values for Hedenfalk et al.

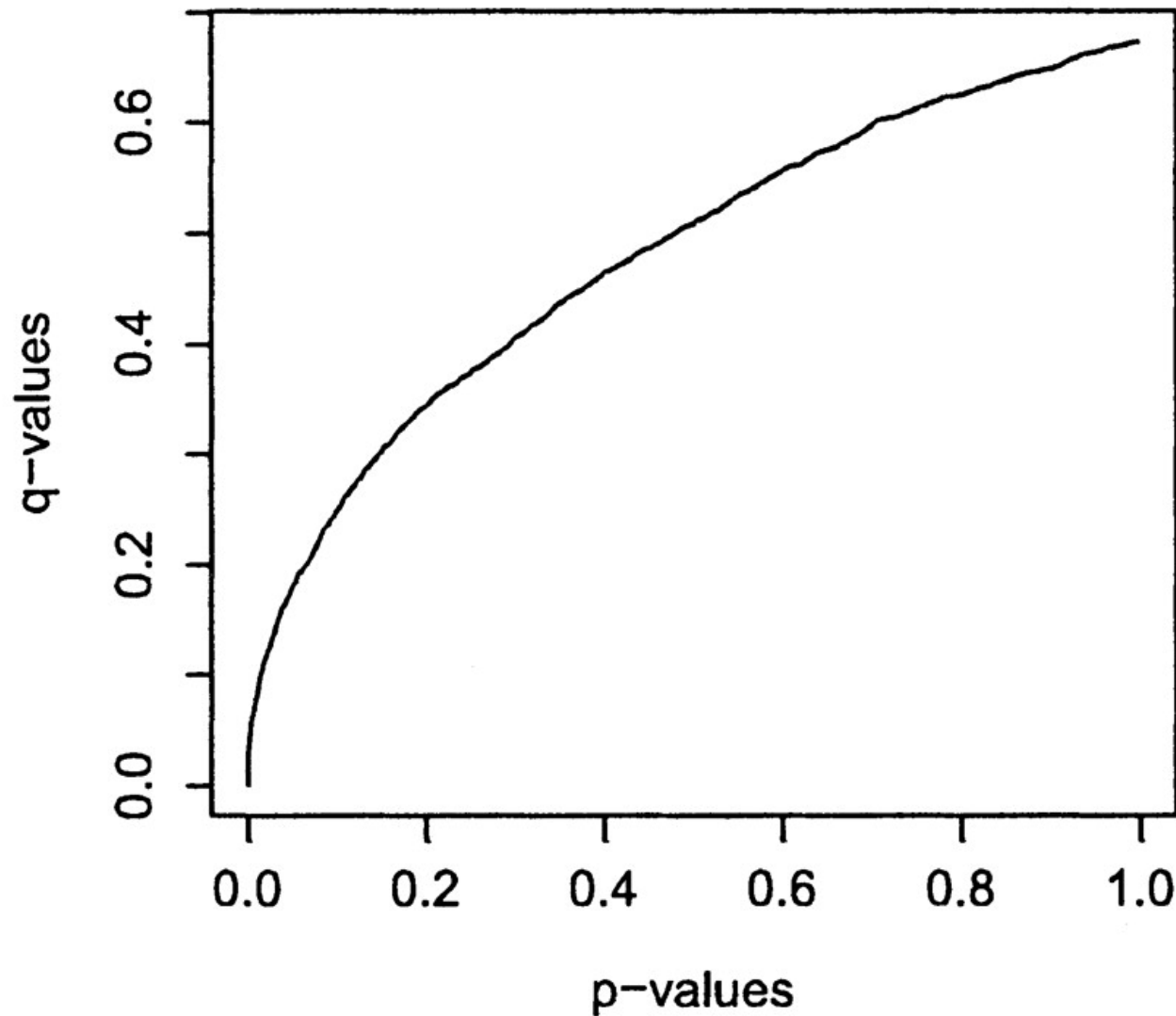


Figure from Storey & Tibshirani *PNAS* 100(16), 2002.

# 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
- the 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 package QVALUE)

# Back to GWAS: 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*

# The era of “BIG Data”