

Inferring genetic variation and discovering associations with phenotypes

BMI/CS 776

www.biostat.wisc.edu/bmi776/

Spring 2017

Anthony Gitter

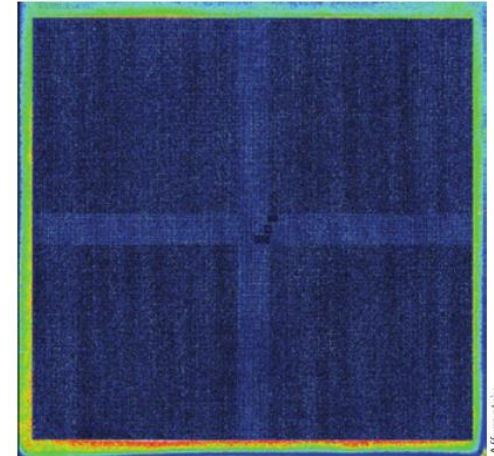
gitter@biostat.wisc.edu

Outline

- Variation detection
 - Array technologies
 - Whole-genome sequencing
- Genome-wide association study (GWAS) 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



Affymetrix SNP chip



Illumina HiSeq sequencer

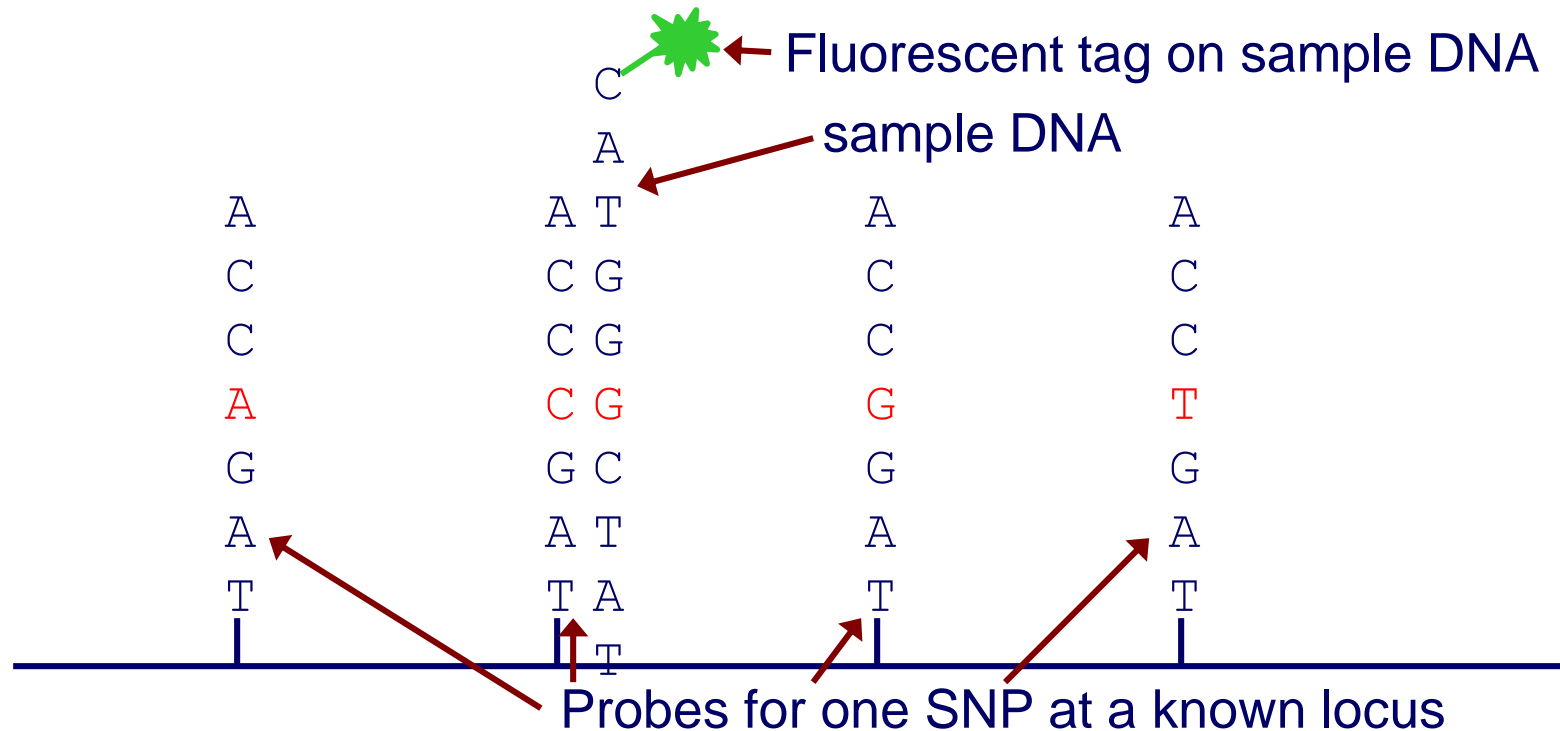
Array-based technologies

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



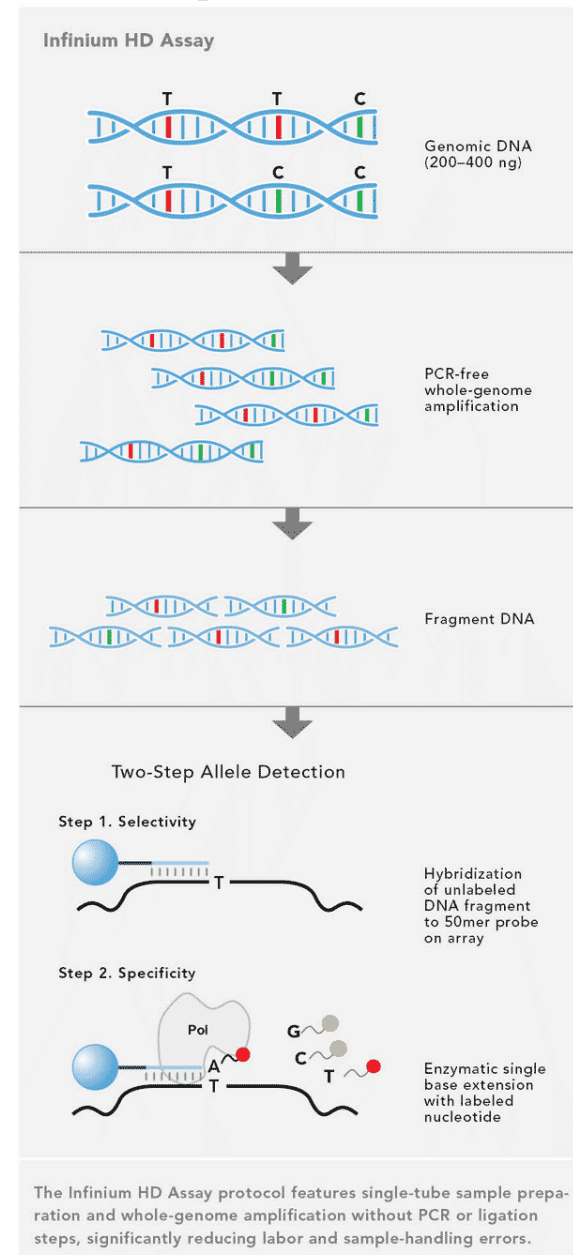
Affymetrix SNP arrays

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



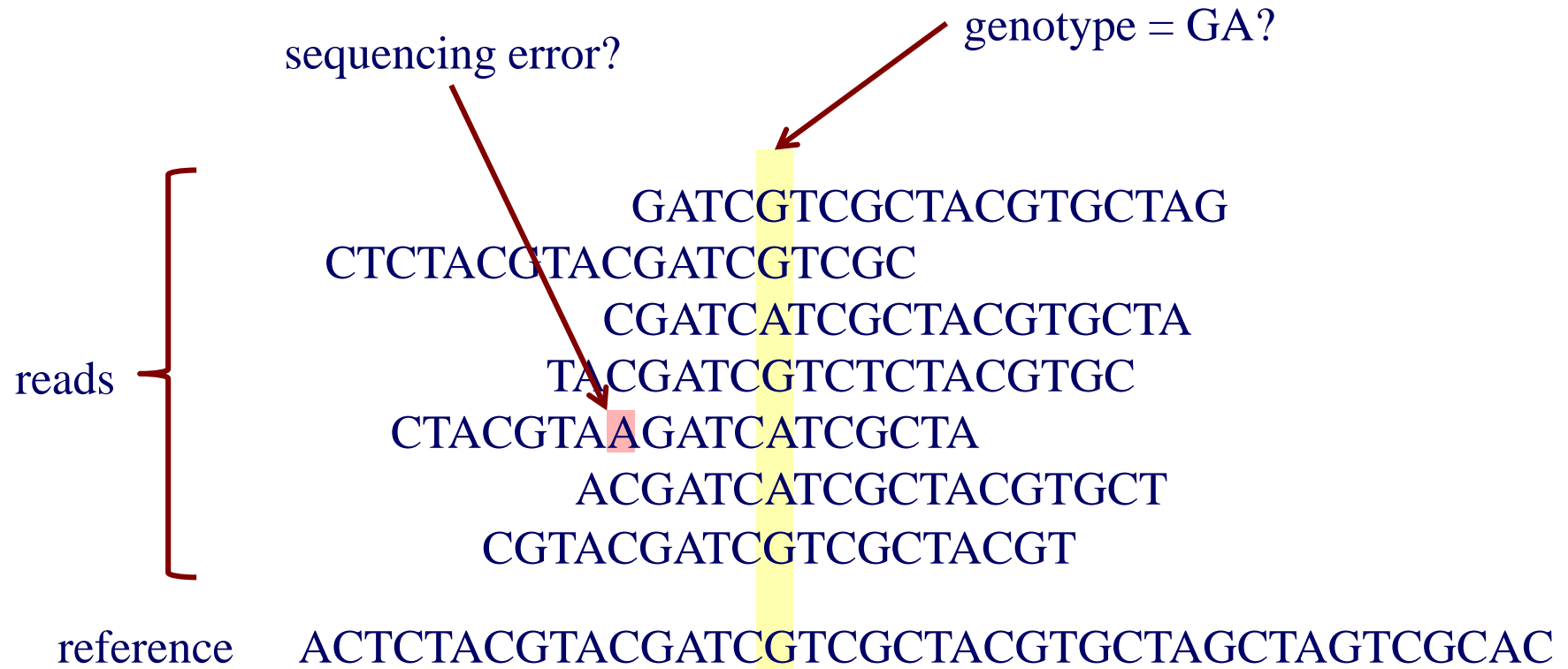
Illumina BeadChips

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



Sequencing-based genotyping

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



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)

Compiled from <http://www.nature.com/scitable/definition/allele-48> <http://www.nature.com/scitable/definition/genotype-234>
<http://www.nature.com/scitable/definition/haplotype-142> <http://www.nature.com/scitable/definition/snp-295> <https://en.wikipedia.org/wiki/Allele>
<http://www.nature.com/nrg/journal/v9/n6/full/nrg2361.html> <https://www.snpedia.com/index.php/Glossary>

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

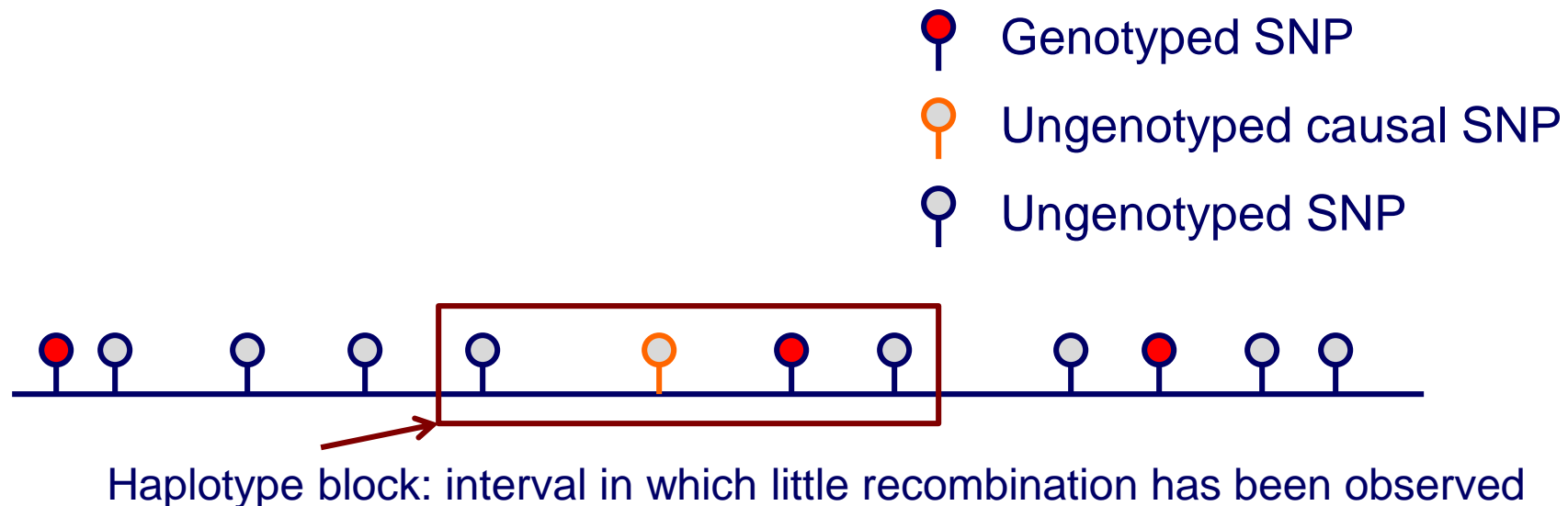
- N individuals genotyped at M positions
- Disease status (or other phenotype) is measured for each individual

GWAS task

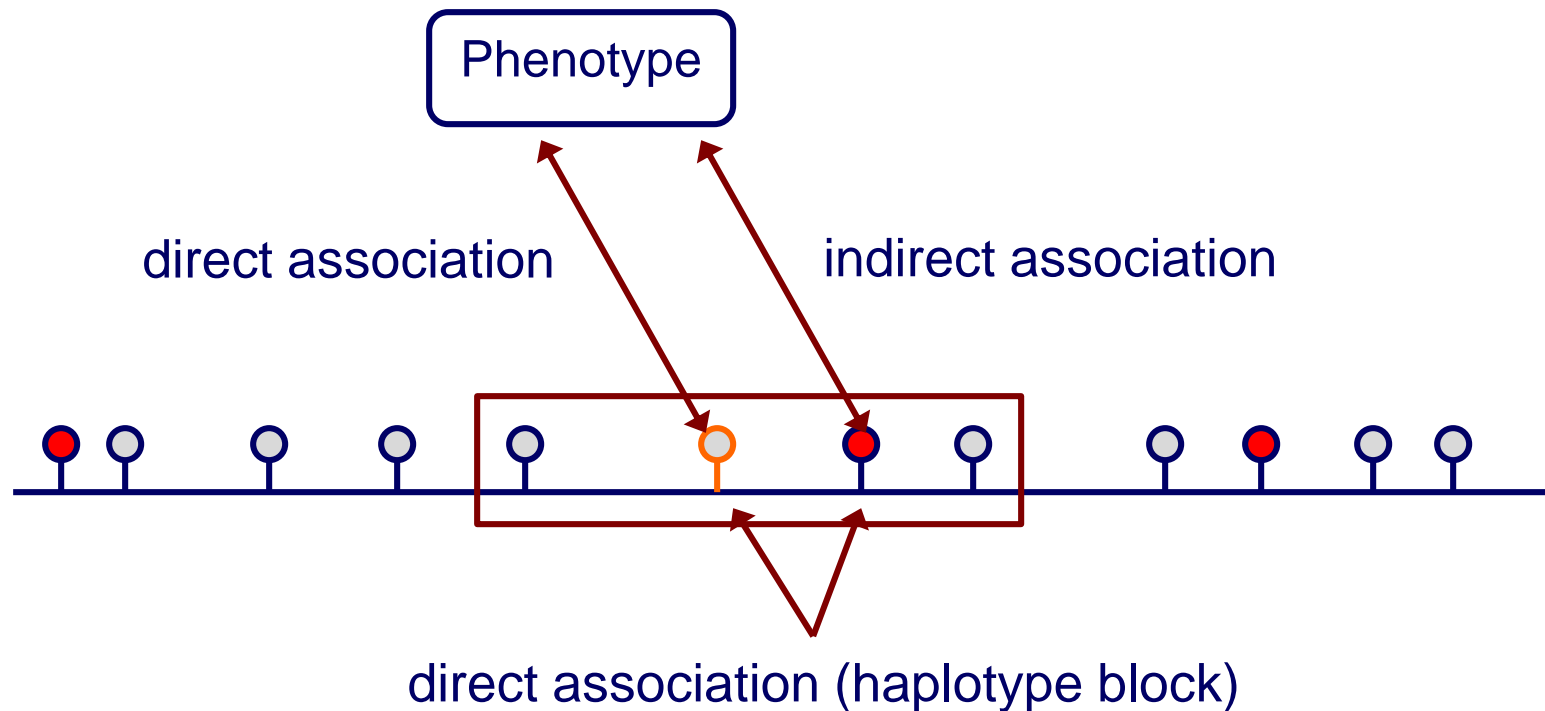
- *Given*: genotypes and phenotypes of individuals in a population
- *Do*: identify which genomic positions are associated with a given phenotype

Can we identify causal SNPs?

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



Direct and indirect associations

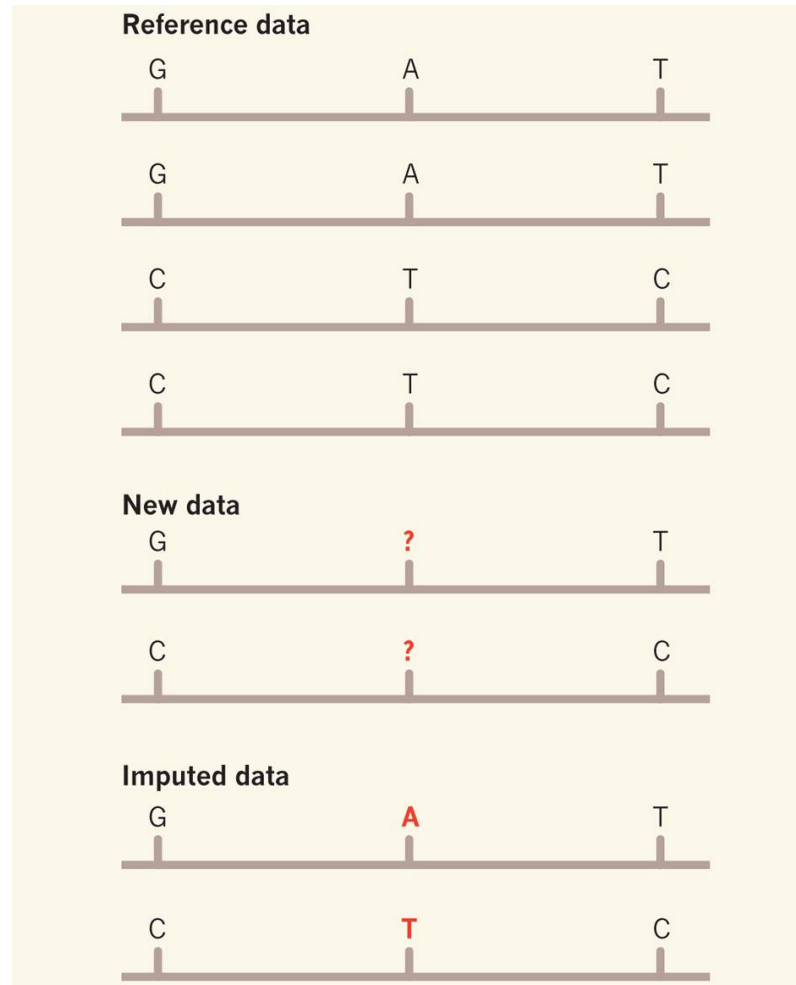


SNP imputation

- Estimate the ungenotyped SNPs using reference haplotypes

1000 Genomes

SNP array



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

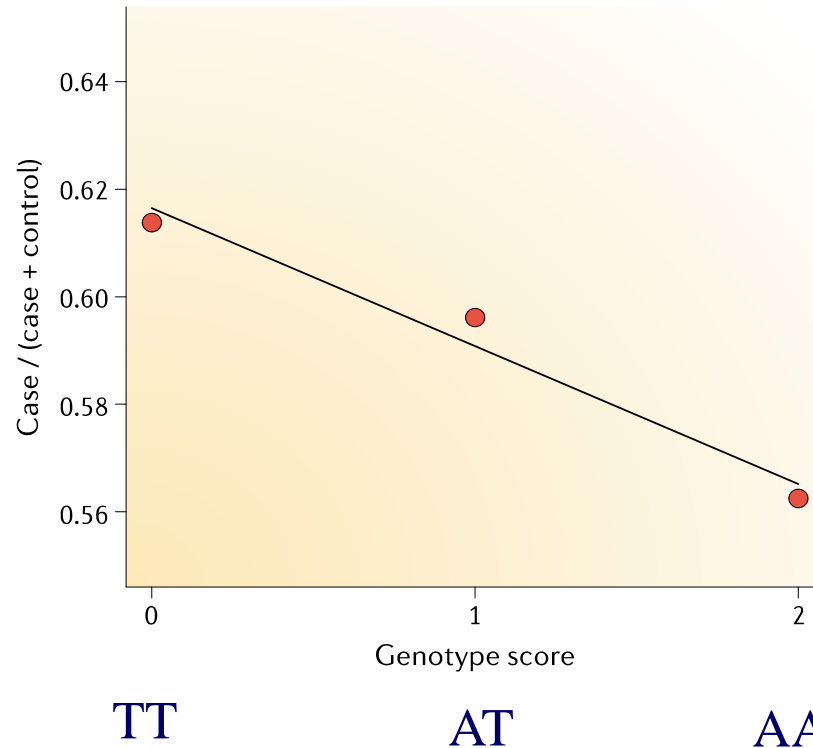
		genotype		
phenotype		AA	AT	TT
	Disease	40	30	30
	No disease	70	20	10

Chi-squared test p -value = $4.1\text{e-}5$ (2 degrees of freedom)

Fisher's Exact Test p -value = $3.4\text{e-}5$

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

		genotype		
phenotype		AA	AT	TT
	Disease	40	30	30
	No disease	70	20	10
Disease proportion		0.36	0.60	0.75

Trend in Proportions test p -value = $8.1\text{e-}6$

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

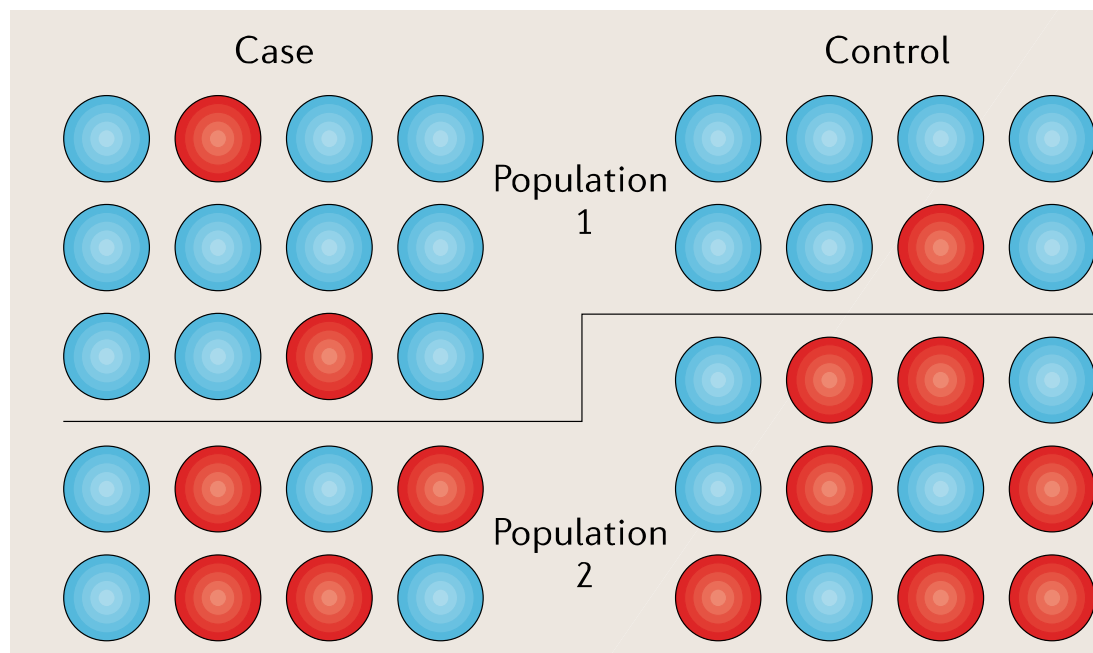
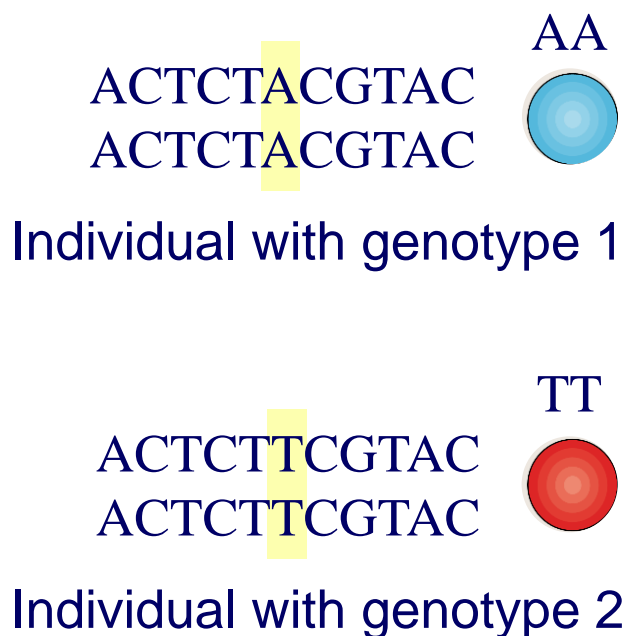
GWAS challenges

- 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

One SNP for $N = 40$ individuals



Balding *Nature Reviews Genetics* 2006

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?

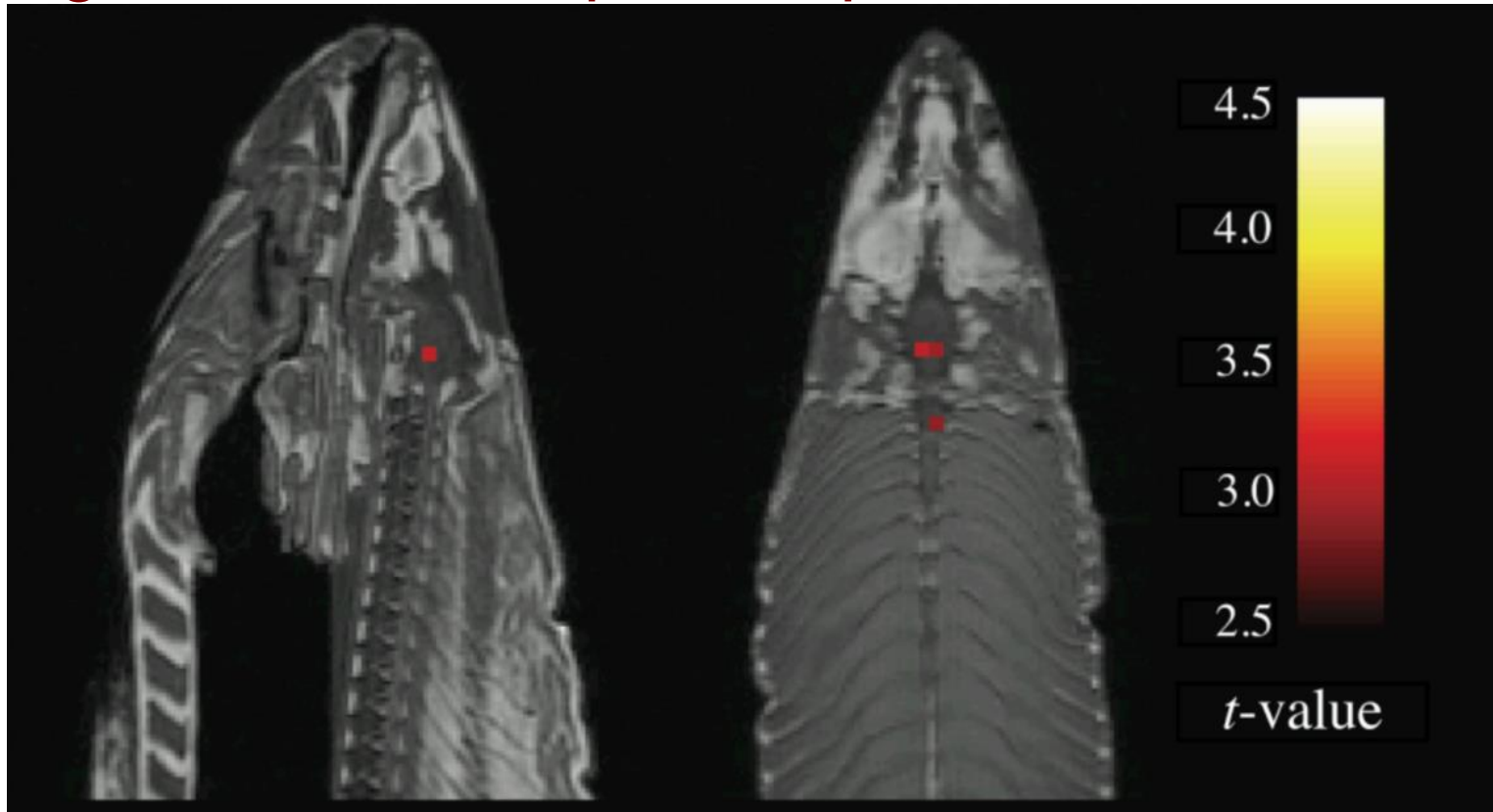
Multiple testing

Bennett et al. “Neural correlates of interspecies perspective taking in the post-mortem Atlantic Salmon: An argument for multiple comparisons correction”

- “One mature Atlantic Salmon (*Salmo salar*) participated in the fMRI study. The salmon was... not alive at the time of scanning.”
- “The salmon was shown a series of photographs depicting human individuals... [and] asked to determine what emotion the individual in the photo must have been experiencing.”
- fMRI to assess changes in brain activity

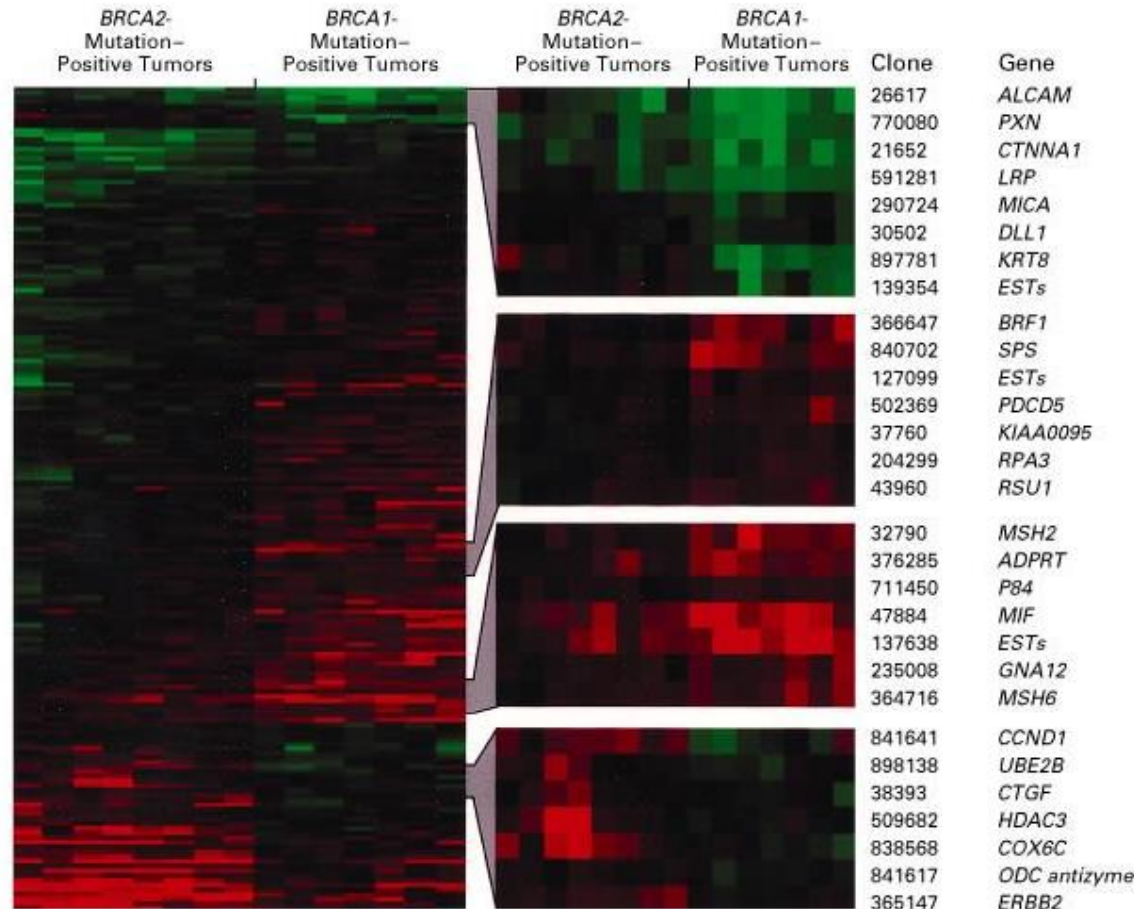
Multiple testing

Bennett et al. “Neural correlates of interspecies perspective taking in the post-mortem Atlantic Salmon: An argument for multiple comparisons correction”



t-test finds 16 significant voxels ($p < 0.001$)

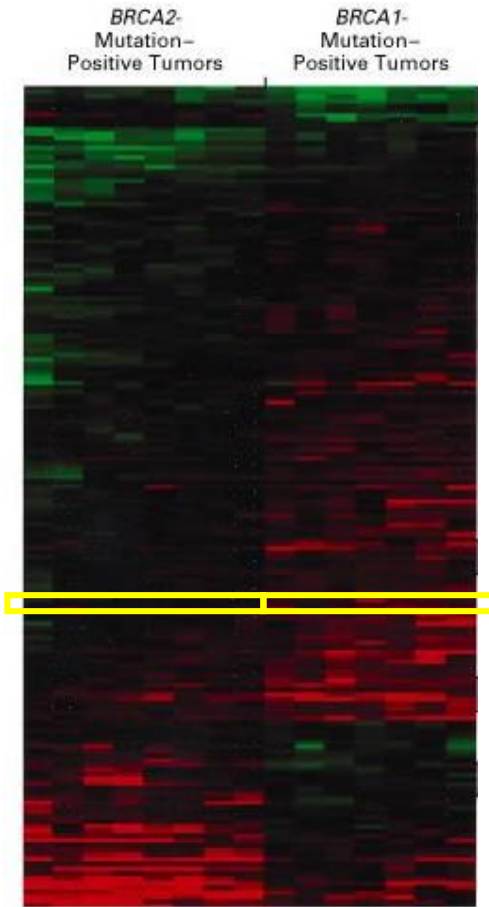
Expression in BRCA1 and BRCA2 Mutation-Positive Tumors



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
- 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)

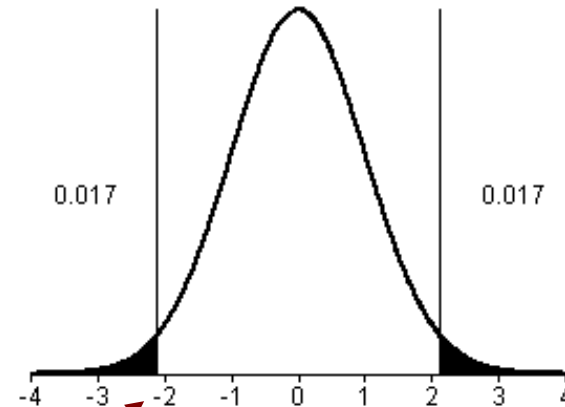
2. See how much mass in null
distribution with value this
extreme or more

BRAC2   BRAC1

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

where $\bar{x}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} x_{ij}$

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



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 ≤ 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 ≤ 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 β 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 $\beta=1\text{e-}5$


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

false positives (false discoveries)

	Called significant	Called not significant	Total
Null true	F	$m_0 - F$	m_0
Alternative true	T	$m_1 - T$	m_1
Total	S	$m - S$	m

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

total significant at threshold

features (genes)

true positives

The False Discovery Rate

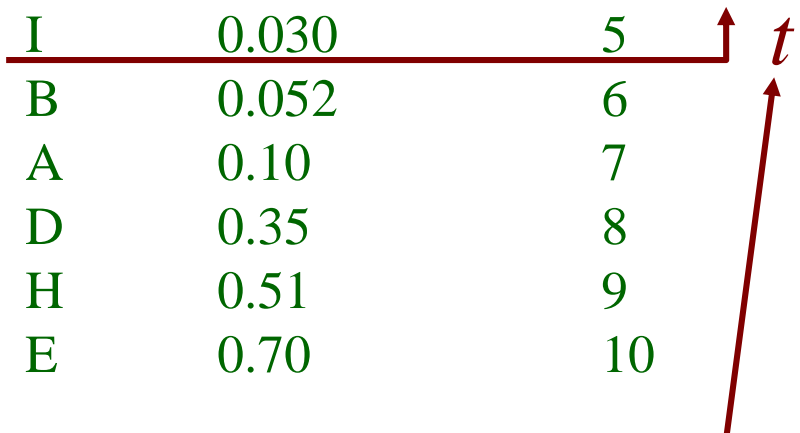
$$F(t) = \# \{ \text{null } p_i \leq t; i = 1 \dots m \}$$

 # genes

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

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

gene	<i>p</i> -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



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 $S(t)$

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

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

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

Estimating $E[F(t)]$

- Two approaches we'll consider
 - Benjamini-Hochberg
 - 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

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

- 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

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

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

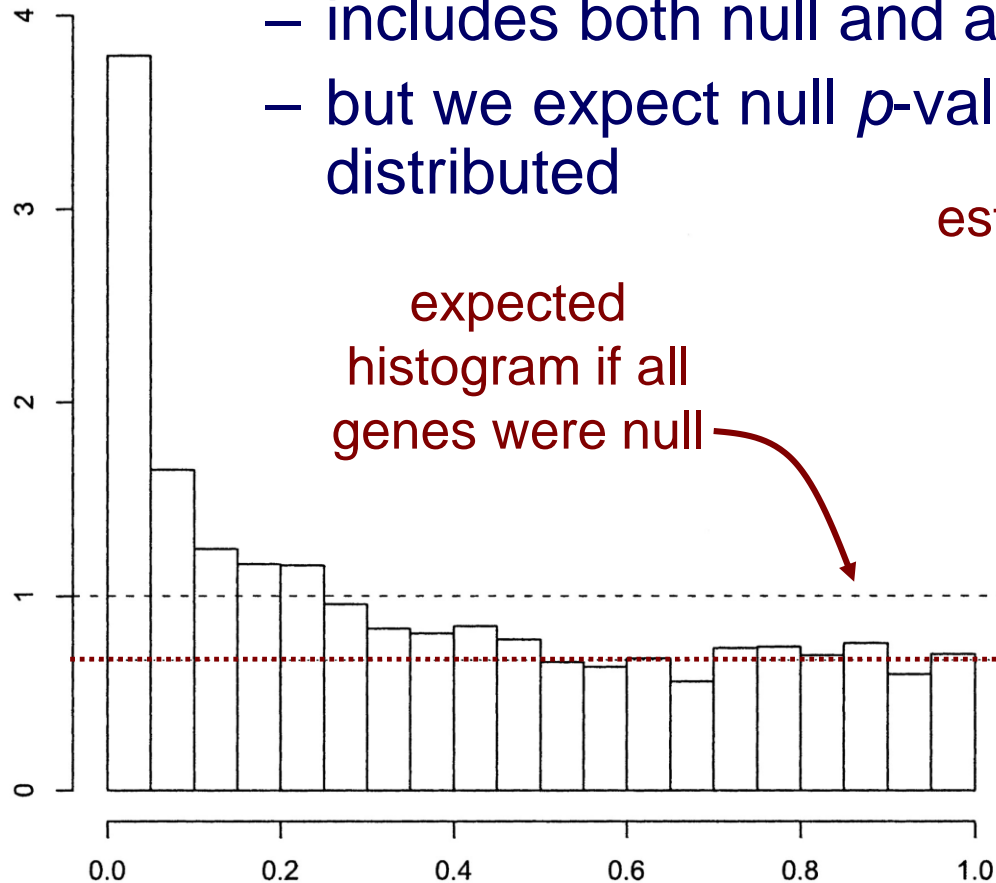
- Select the largest k such that

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

- 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



actual proportion of null p -values

$$\pi_0 = \frac{m_0}{m}$$

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

Storey & Tibshirani approach

estimated proportion of null p -values

genes

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

p -value threshold

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 -value example for gene J

$$m = 20$$

$$t = 0.019$$

$$\hat{\pi}_0 = 0.5$$

$$\#\{p_i \leq t\} = 4$$

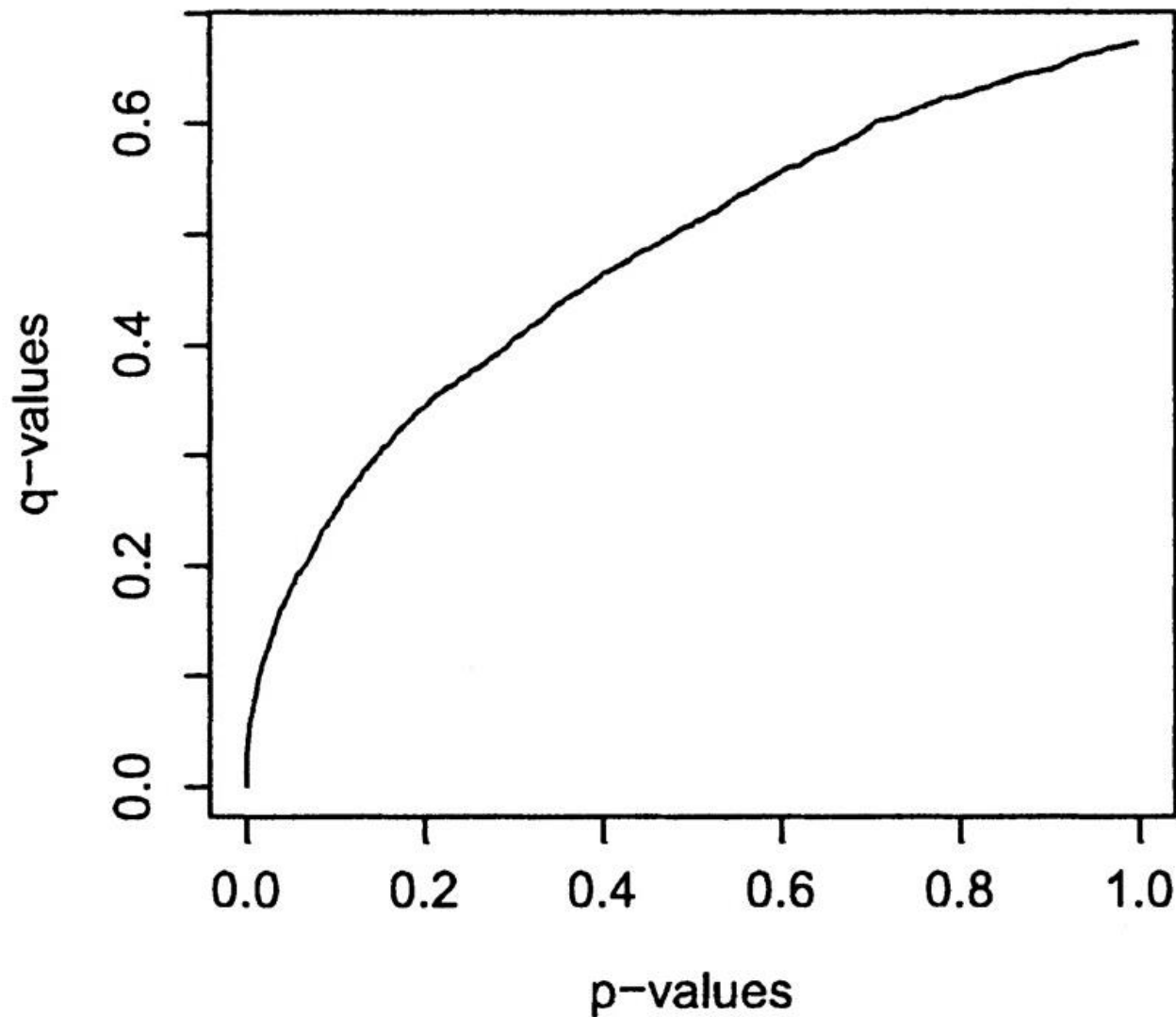
$$\begin{aligned} FDR(t) &\approx \frac{\hat{\pi}_0 \times m \times t}{\#\{p_i \leq t\}} \\ &= \frac{0.5 \times 20 \times 0.019}{4} = 0.0475 \end{aligned}$$

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

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

In this case, already have minimum FDR for all greater thresholds

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



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`)