Genetics and variation

Recitation 7b

MIT - 6.802 / 6.874 / 20.390 / 20.490 / HST.506 - Spring 2021

Jackie Valeri

Slides adapted from Corban Swain and previous course materials

Onto recitation R07!

A. Dimensionality reduction review (ft. Zheng)

B. Genetics review (ft. Jackie)
   I. GWAS
   II. Quantile-quantile plots
   III. Multiple hypothesis test corrections

Genetics review: GWAS

- **The primary purpose of GWAS is to identify associations of phenotype with single nucleotide polymorphisms across the genome**
  - SNP = one base pair change in the DNA (ex: A → T)

Genetics review: GWAS

- **The primary purpose of GWAS is to identify associations of phenotype with single nucleotide polymorphisms across the genome**
  - SNP = one base pair change in the DNA (ex: A → T)
  - Most base pairs will be invariant across different populations
  - Some SNPs are statistically over- or under-enriched in a specific population compared to a "normal" population.

Genetics review: GWAS

- **The primary purpose of GWAS is to identify associations of phenotype with single nucleotide polymorphisms across the genome**
  - SNP = one base pair change in the DNA (ex: A → T)
  - Most base pairs will be invariant across different populations
  - Some are statistically over- or under-enriched in a specific population compared to a "normal" population.
  - Specific populations could include persons having a certain disease, people who respond adversely to a specific therapeutic, and other phenotypic classifications.

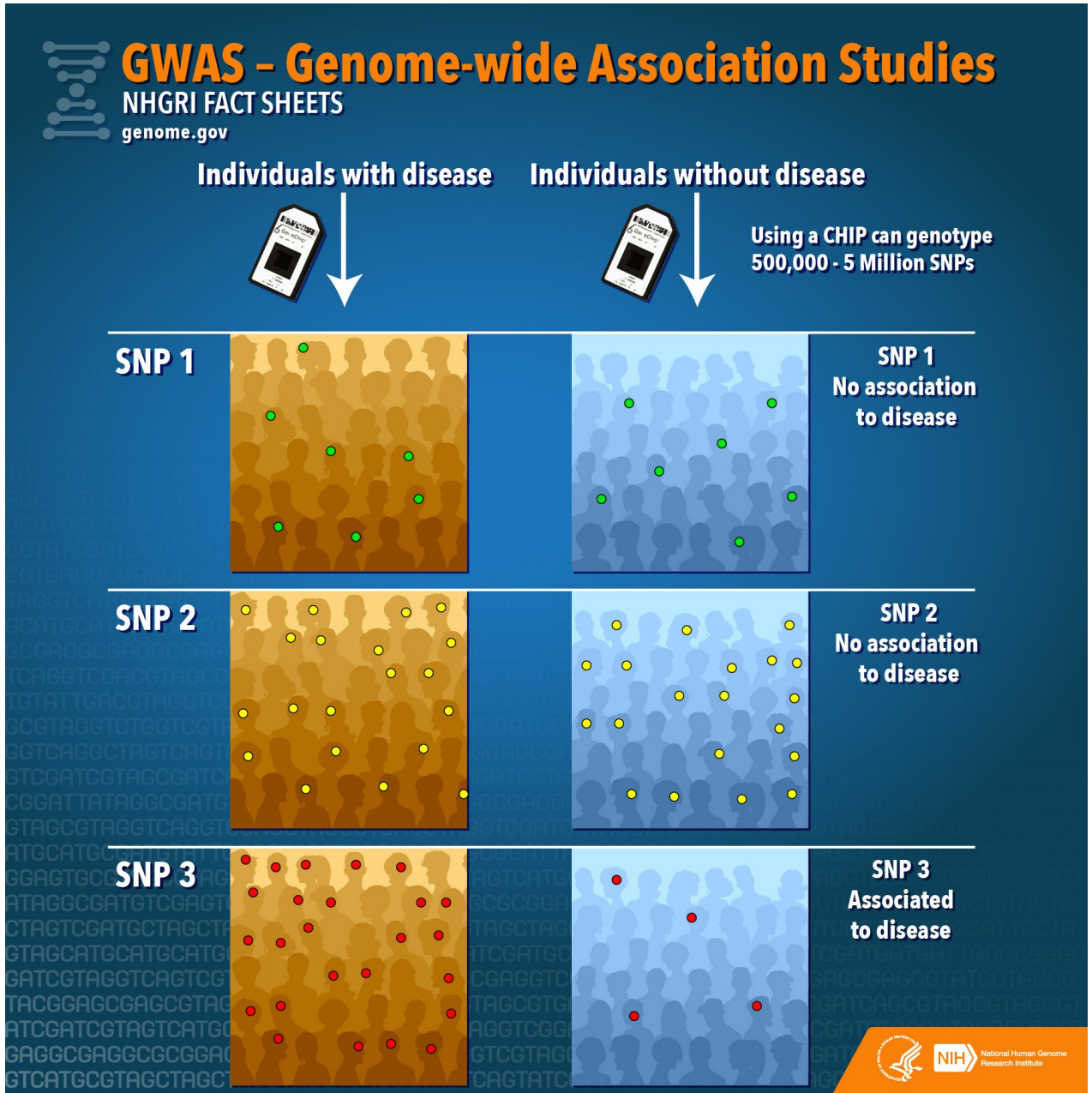
Genetics review: GWAS

- **The primary purpose of GWAS is to identify associations of phenotype with single nucleotide polymorphisms across the genome**
  - SNP = one base pair change in the DNA (ex: A → T)
  - Most base pairs will be invariant across different populations
  - Some are statistically over- or under-enriched in a specific population compared to a "normal" population.
  - Specific populations could include persons having a certain disease, people who respond adversely to a specific therapeutic, and other phenotypic classifications.
- The "library" of SNPs to be tested is often on the order of 100k.

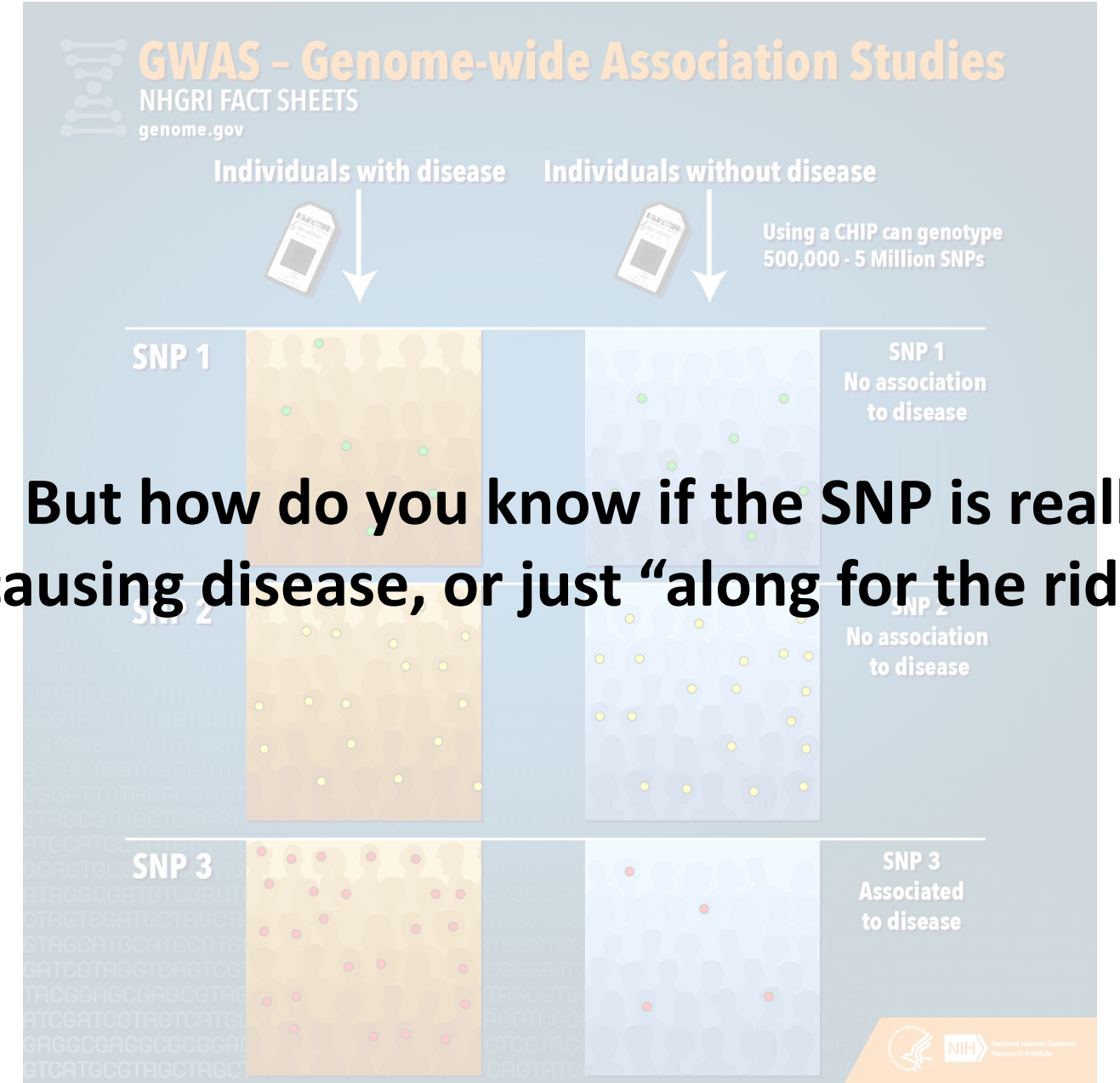
Genetics review: GWAS

- **The primary purpose of GWAS is to identify associations of phenotype with single nucleotide polymorphisms across the genome**
  - SNP = one base pair change in the DNA (ex: A → T)
  - Most base pairs will be invariant across different populations
  - Some are statistically over- or under-enriched in a specific population compared to a "normal" population.
  - Specific populations could include persons having a certain disease, people who respond adversely to a specific therapeutic, and other phenotypic classifications.
- The "library" of SNPs to be tested is often on the order of 100k.
  - GWAS has large cohort sizes (thousands to hundreds of thousands) and high coverage of genomic sites
  - So GWAS is well suited to identify many SNPs related to "polygenic" disorders, where each genomic site has a small, but significant, effect on phenotype perturbation.

Genetics review: GWAS



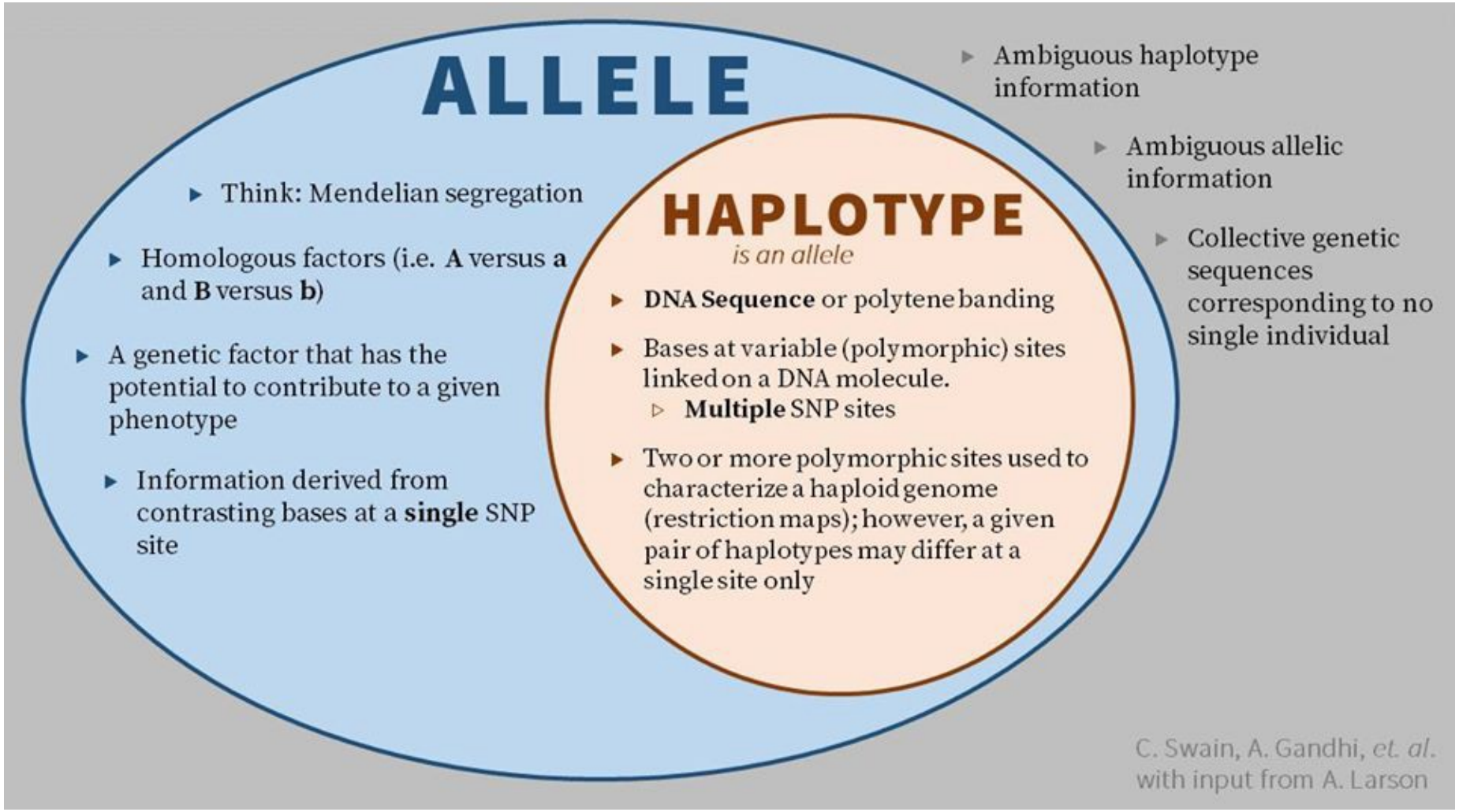
Genetics review: GWAS



But how do you know if the SNP is really causing disease, or just "along for the ride"?

Genetics review: Haplotypes and linkage disequilibrium

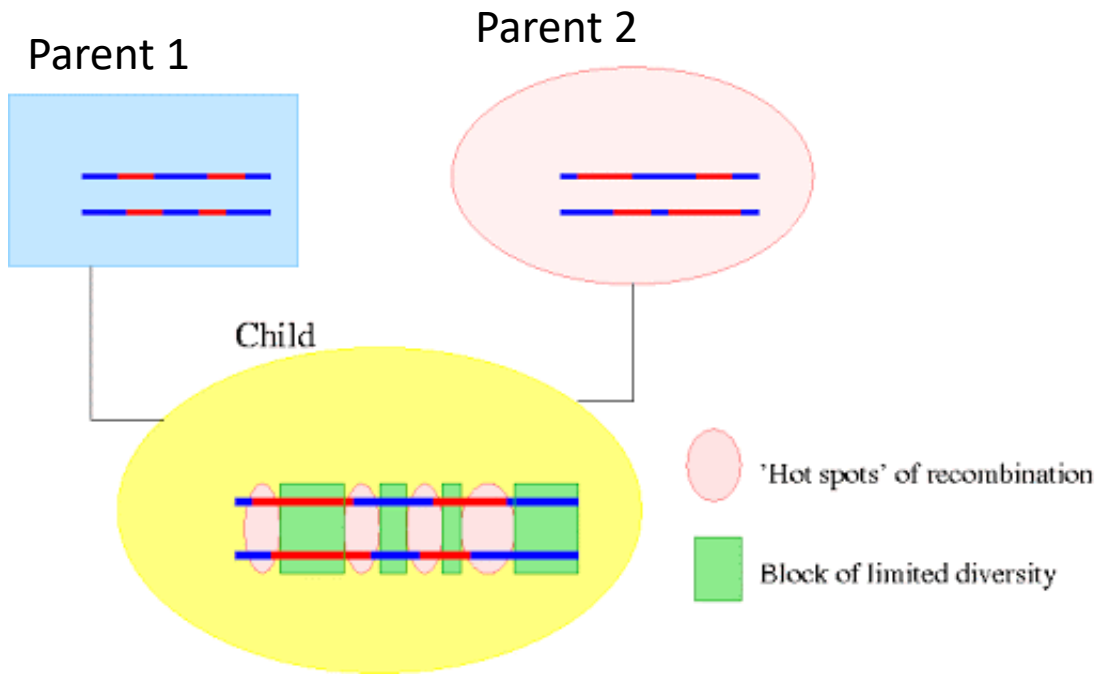
Allelic groupings called “haplotypes” are important confounding factors in GWAS analysis



Genetics review: Haplotypes and linkage disequilibrium

Allelic groupings called “haplotypes” are important confounding factors in GWAS analysis

Haplotypes arise because of recombination



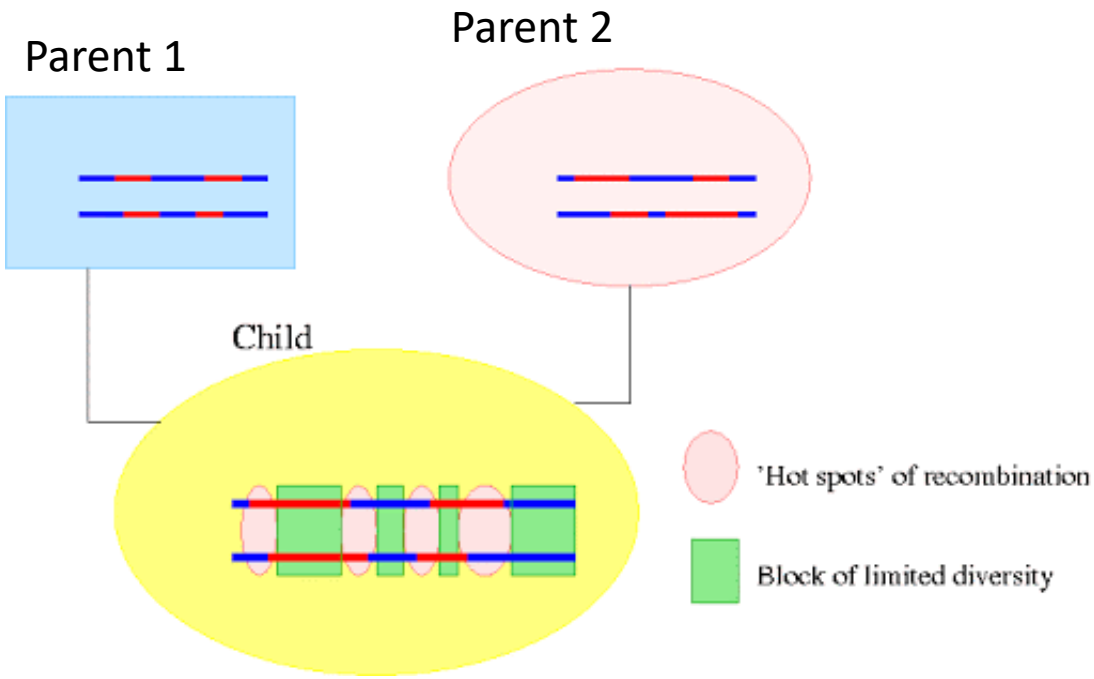
Genetics review: Haplotypes and linkage disequilibrium

Allelic groupings called “haplotypes” are important confounding factors in GWAS analysis

Haplotypes arise because of recombination

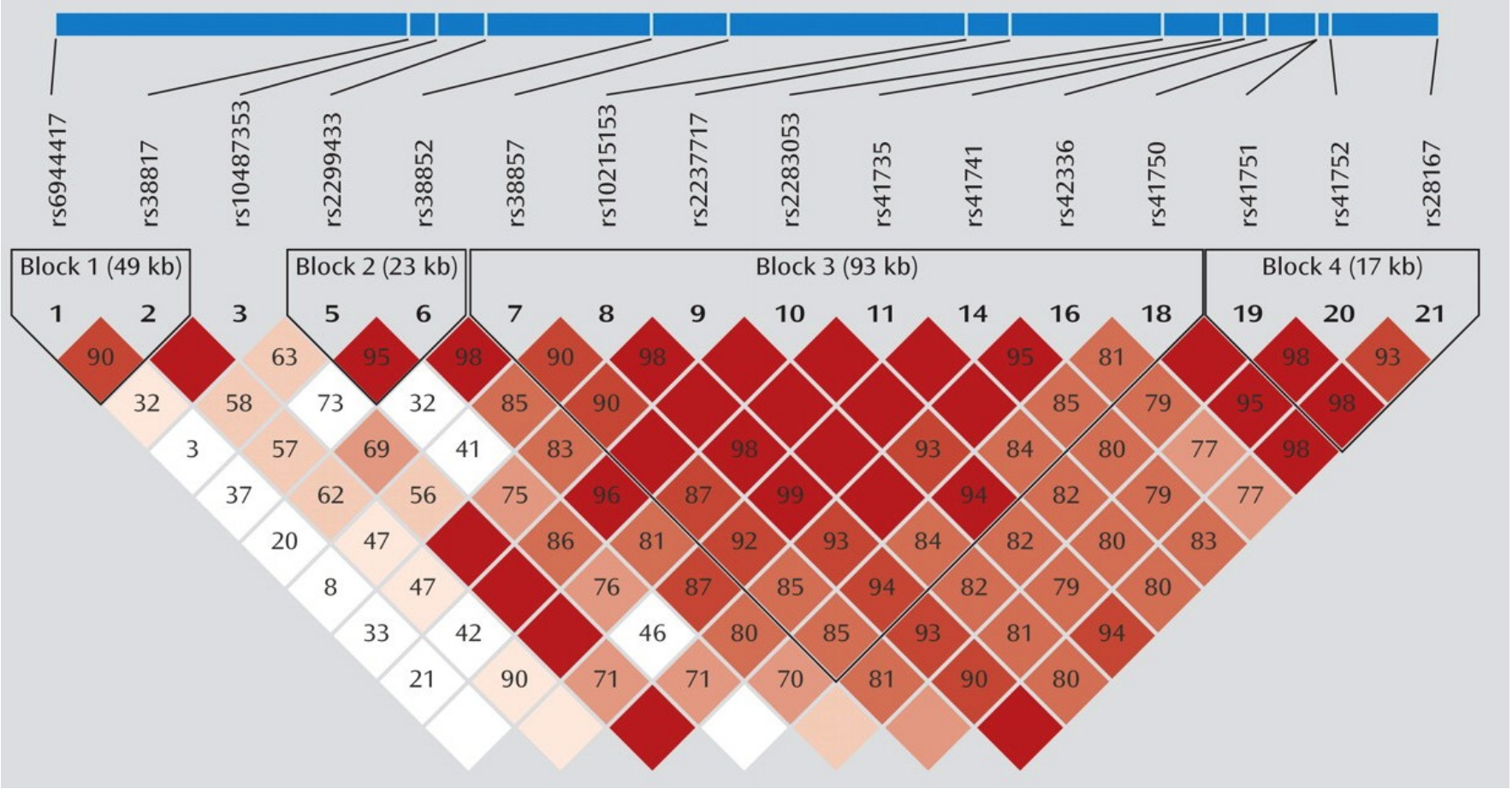
Low haplotype diversity can cause high “linkage disequilibrium” (LD) = a whole set of SNPs which tend to arise in specific patterns

LD makes de-tangling biologically significant SNP sites from those that simply happen to be included in the haplotype difficult!



Genetics review: Haplotypes and linkage disequilibrium

Linkage disequilibrium [Useful description of linkage diagrams at this link.]



Genetics review: Applications of GWAS

- An important goal after performing GWAS is the identification of the **causal link** between a given allele and the perturbation it is known to cause.

- Furthermore, in a disease context, we want to be able to use our knowledge of the causal mechanism to **propose and design therapeutics** to decrease of mitigate the risk presented by a certain genetic predisposition.

- An excellent example of this was described in lecture with respect to obesity and the FTO locus!

Onto recitation R07!

A. Dimensionality reduction review (ft. Zheng)

B. Genetics review (ft. Jackie)
   I. GWAS
   II. Quantile-quantile plots
   III. Multiple hypothesis test corrections

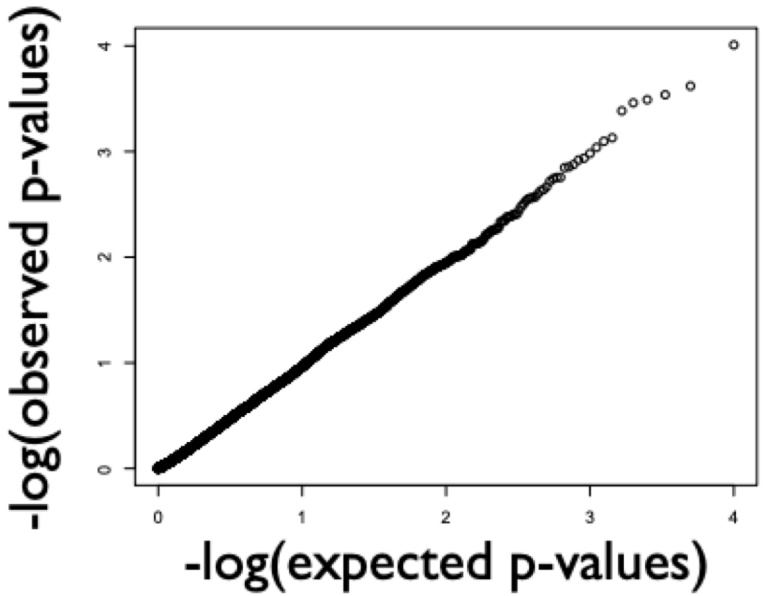
Slides adapted from https://physiology.med.cornell.edu/people/banfelder/qbio/resources_2013/2013_1_Mezey.pdf

Genetics review: Quantile-quantile plots, explained

- An essential tool for detecting the problems in a GWAS is a Quantile-Quantile (QQ) plot
- **quantile** - regular, equally spaced intervals of a random variable that divide the random variable into units of equal distribution
- A Quantile-Quantile (QQ) plot (in general) plots the observed quantiles of one distribution versus another OR plots the observed quantiles of a distribution versus the quantiles of the ideal distribution
- In GWAS we use a QQ plot to plot our the quantile distribution of observed p-values (on the y-axis) versus the quantile distribution of expected p-values

Genetics review: Quantile-quantile plots, explained

- In an ideal GWAS case where there ARE NO causal polymorphisms, your QQ plot will be a line:

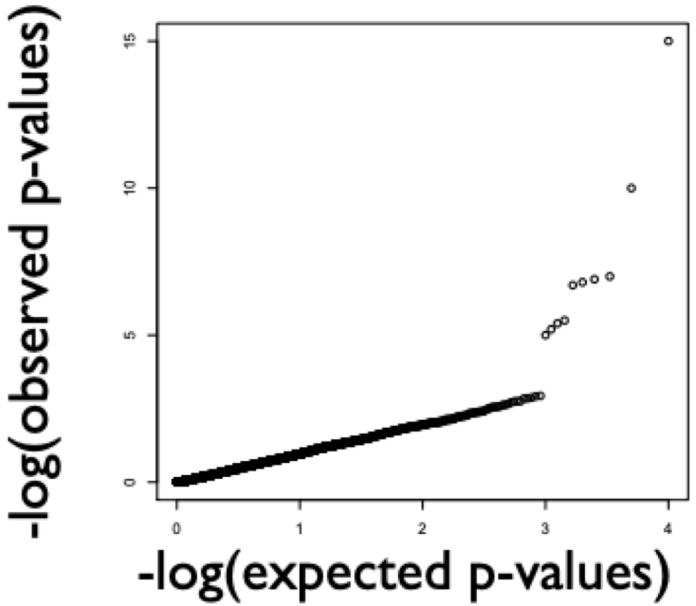


- The reason is that we will observe a uniform distribution of p-values from such a case and in our QQ we are plotting this observed distribution of p-value versus the expected distribution of p-values: a uniform distribution (where both have been -log transformed)

- Note that if you GWAS analysis is correct but you do not have enough power to detect positions of causal polymorphisms, this will also be your result (!!), i.e. it is a way to assess whether you can detect any hits in your GWAS (!!)

Genetics review: Quantile-quantile plots, explained

- In an ideal GWAS case where there ARE causal polymorphisms, your QQ plot will be a line with a tail (!!):

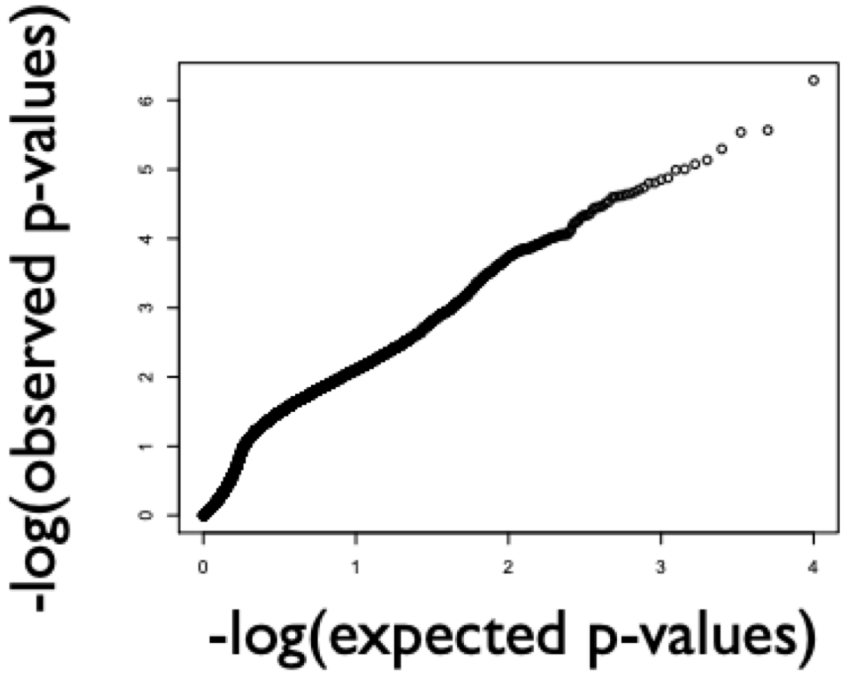


- This happens because most of the p-values observed follow a uniform distribution (i.e. they are not in LD with a causal polymorphism so the null hypothesis is correct!) but the few that are in LD with a causal polymorphism will produce significant p-values (extremely low = extremely high -log(p-values)) and these are in the "tail"

- This is ideally how you want your QQ-plot to look - if it does, you are in good shape!

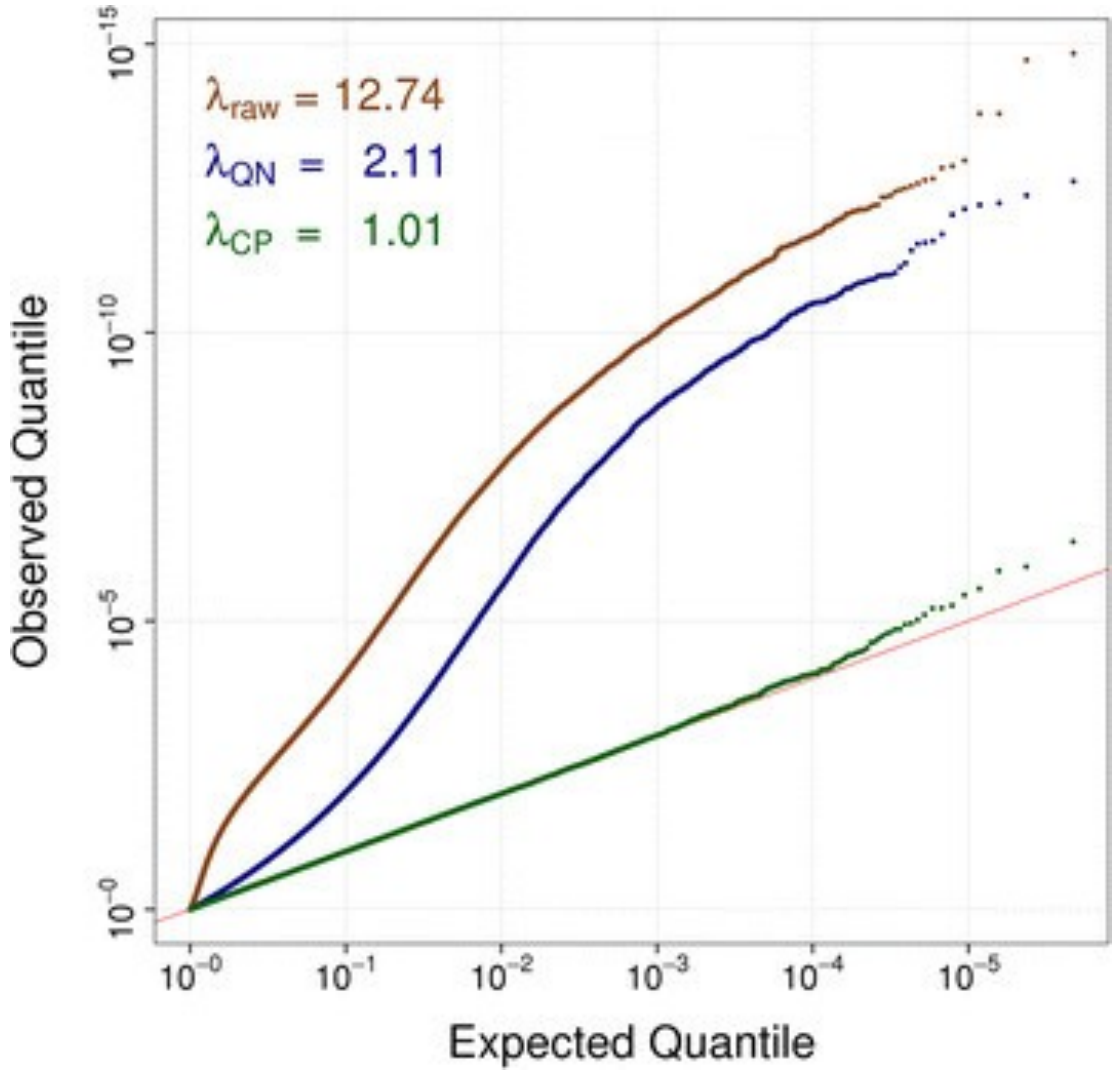
Genetics review: Quantile-quantile plots, explained

- In practice, you can find your QQ plot looks different than either the "null GWAS" case or the "ideal GWAS" case, for example:

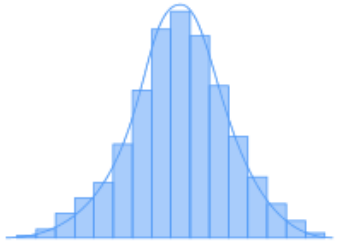


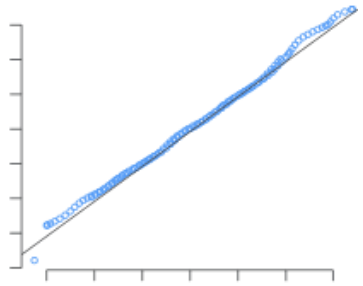
- This indicates that something is wrong (!!!!) and if this is the case, you should not interpret any of your significant p-values as indicating locations of causal polymorphisms (!!!!)
- Note that this means that you need to find an analysis strategy such that the result of your GWAS produces a QQ plot that does NOT look like this (note that this takes experience and many tools to do consistently!)

Genetics review: Quantile-quantile plots, explained

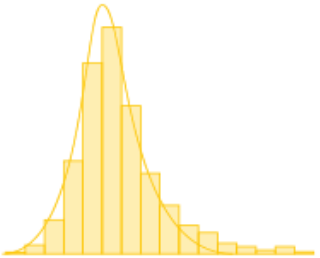


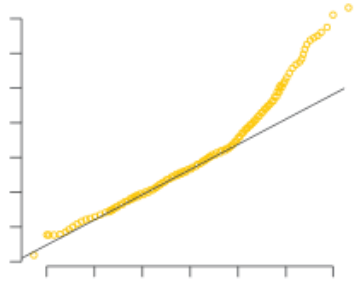
Normally distributed data



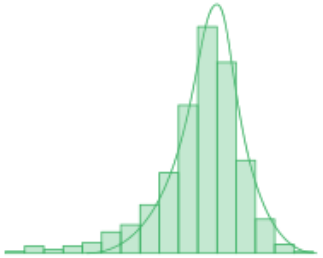


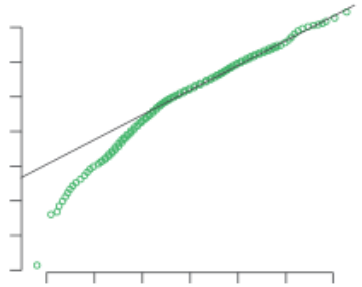
Right-skewed data





Left-skewed data





Genetics review: Quantile-quantile plots, explained

- Let's say that we obtain p-values from 10^5 SNPs computed using the χ^2 distribution between 2 populations.
- If our quality control is performed well, and our SNP distribution is indeed normal (recall that the χ^2 distribution represents the squared distance between normal random variables) then we would expect our p-values to be uniformly distributed.
- So if we rank-order our observed p-values from smallest to largest and plot it against the expected uniform distribution of 10^5 p-values evenly spaced from 0.00001 to 1, we would expect to observe a straight line.
- To evaluate this we can look at the goodness of fit graphically or compute the genomic inflation factor λ which effectively compares the median test statistic to the expected median test statistic.
- All of this is to ensure that our population data is, generally, not biased or prone to producing many false positives.
- We do expect, however, for the SNPs biologically associated with a phenotype of interest to fall off above this line of expectation and be **more significant than would be expected** if all data conformed to the null hypothesis.

Onto recitation R07!

A. Dimensionality reduction review (ft. Zheng)

B. Genetics review (ft. Jackie)
   I. GWAS
   II. Quantile-quantile plots
   III. Multiple hypothesis test corrections

Slides adapted from https://www.gs.washington.edu/academics/courses/akey/56008/lecture/lecture10.pdf

Genetics review: Why multiple hypothesis correction?

Genomics = Lots of Data = Lots of Hypothesis Tests

A typical microarray experiment might result in performing 10000 separate hypothesis tests. If we use a standard p-value cut-off of 0.05, we'd expect ***500*** genes to be deemed "significant" by chance.

Genetics review: Why multiple hypothesis correction?

- In general, if we perform m hypothesis tests, what is the probability of at least 1 false positive?

$$P(\text{Making an error}) = \alpha$$

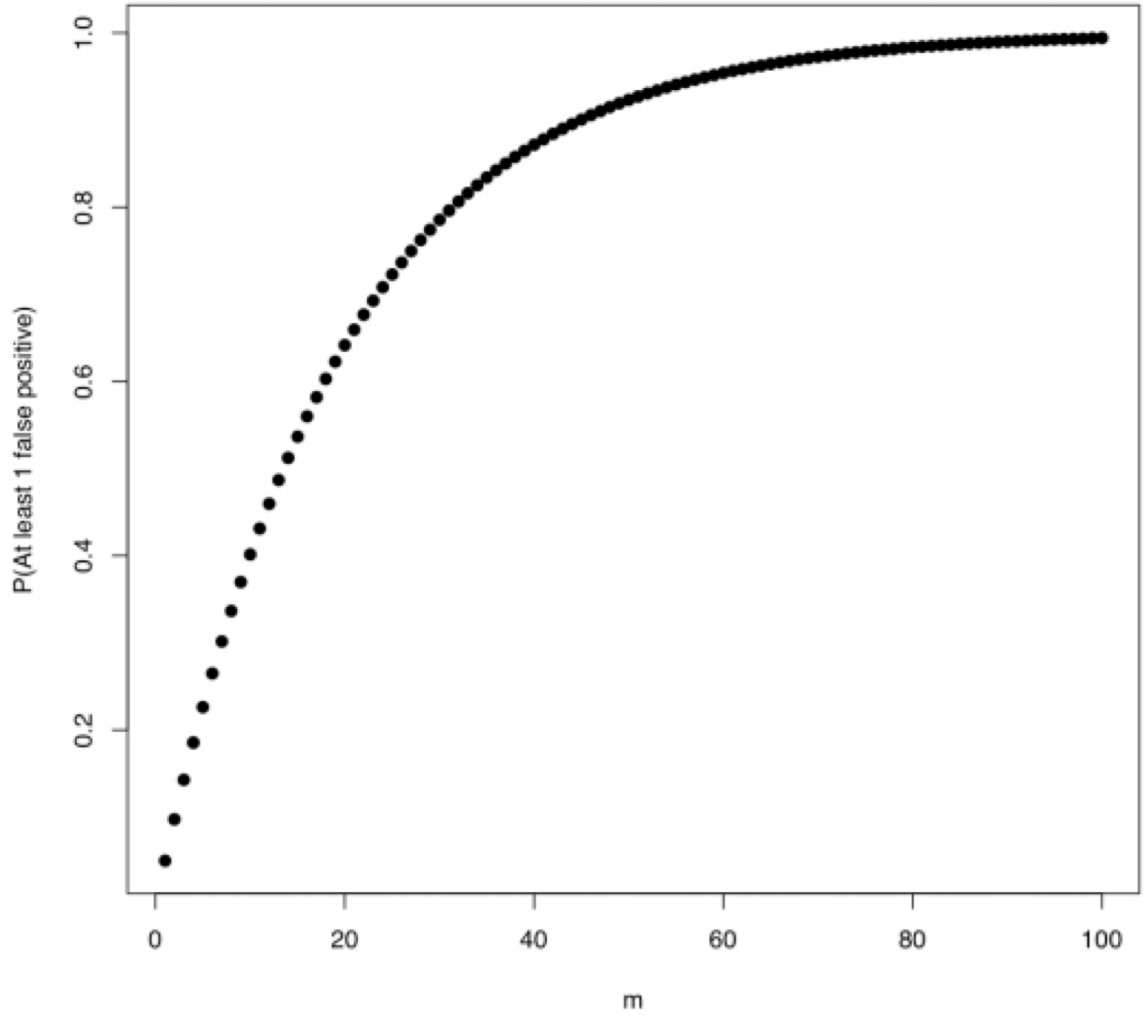
$$P(\text{Not making an error}) = 1 - \alpha$$

$$P(\text{Not making an error in m tests}) = (1 - \alpha)^m$$

$$P(\text{Making at least 1 error in m tests}) = 1 - (1 - \alpha)^m$$

Genetics review: Why multiple hypothesis correction?

Probability of At Least 1 False Positive



Genetics review: How to do MHT correction?

- When people say "adjusting p-values for the number of hypothesis tests performed" what they mean is ***controlling the Type I error rate***

- Very active area of statistics - many different methods have been described

- Although these varied approaches have the same goal, they go about it in fundamentally different ways

Our focus is on two: Bonferroni and Benjamini-Hochberg

Single Step Approach: Bonferroni

- Very simple method for ensuring that the overall Type I error rate of α is maintained when performing m independent hypothesis tests

- Rejects any hypothesis with p-value $\leq \alpha/m$:

$$\tilde{p}_j = \min[mp_j, 1]$$

- For example, if we want to have an experiment wide Type I error rate of 0.05 when we perform 10,000 hypothesis tests, we'd need a p-value of $0.05/10000 = 5 \times 10^{-6}$ to declare significance

Philosophical Objections to Bonferroni Corrections

"Bonferroni adjustments are, at best, unnecessary and, at worst, deleterious to sound statistical inference" Perneger (1998)

- Counter-intuitive: interpretation of finding depends on the number of other tests performed
- The general null hypothesis (that all the null hypotheses are true) is rarely of interest
- High probability of type 2 errors, i.e. of not rejecting the general null hypothesis when important effects exist

Genetics review: How to do MHT correction?

Benjamini and Hochberg FDR

- To control FDR at level δ:

  1. Order the unadjusted p-values: $p_1 \leq p_2 \leq \ldots \leq p_m$

  2. Then find the test with the highest rank, j, for which the p value, p_j, is less than or equal to (j/m) x δ

  3. Declare the tests of rank 1, 2, …, j as significant

$$p(j) \leq \delta \frac{j}{m}$$

False discovery rate (FDR) is the expected proportion of Type I errors among the rejected null hypotheses

FDR = E(V/R | R>0) P(R>0)

Genetics review: How to do MHT correction?

B&H FDR Example

Controlling the FDR at $\delta = 0.05$

Rank (j)	P-value	(j/m)× δ	Reject H_0 ?
1	0.0008	0.005	1
2	0.009	0.010	1
3	0.165	0.015	0
4	0.205	0.020	0
5	0.396	0.025	0
6	0.450	0.030	0
7	0.641	0.035	0
8	0.781	0.040	0
9	0.900	0.045	0
10	0.993	0.050	0

Genetics review: How to do MHT correction?

What's a q-value?

- q-value is defined as the minimum FDR that can be attained when calling that "feature" significant (i.e., expected proportion of false positives incurred when calling that feature significant)

- The estimated q-value is a function of the p-value for that test and the distribution of the entire set of p-values from the family of tests being considered (Storey and Tibshiriani 2003)

- Thus, in an array study testing for differential expression, if gene X has a q-value of 0.013 it means that 1.3% of genes that show p-values at least as small as gene X are false positives

Next week

eQTLs

UK BioBank

Genetics Wrap Up