Multiple Hypothesis Testing

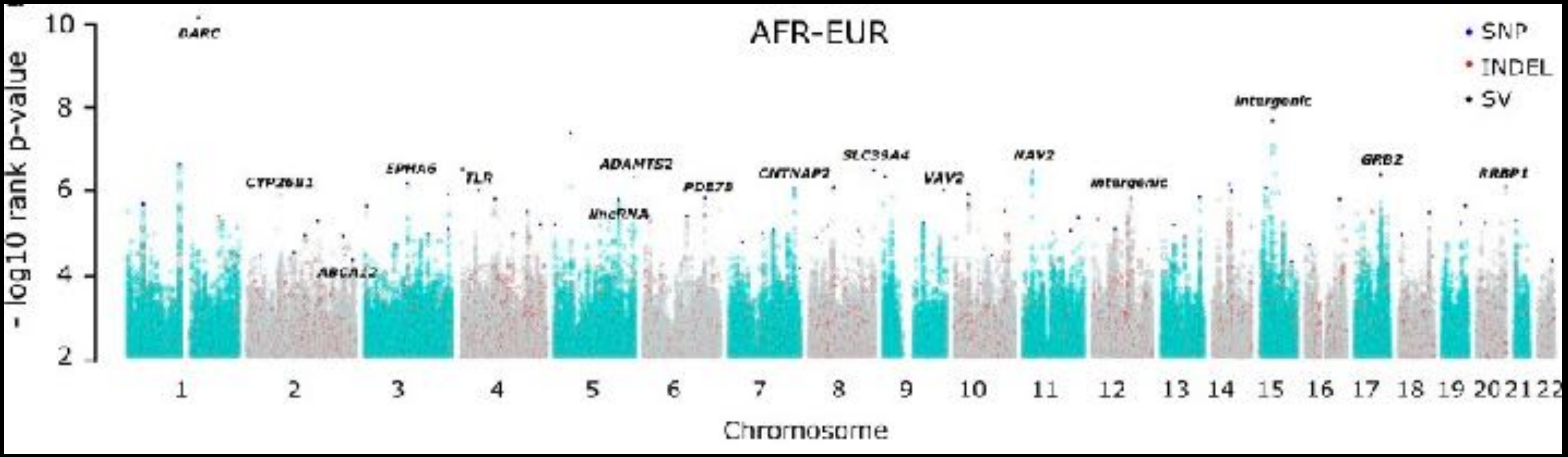
Katie Pollard

BMI 206

In this unit we will learn …

- Why p<0.05 is not sufficient in bioinformatics
- Several ways to quantify error in studies with many statistical tests
- Pros and cons of using these different error rates
- Methods for controlling error rates that differ in their computational complexity
- How to implement multiple testing corrections in R code

Multiple testing in population genetics



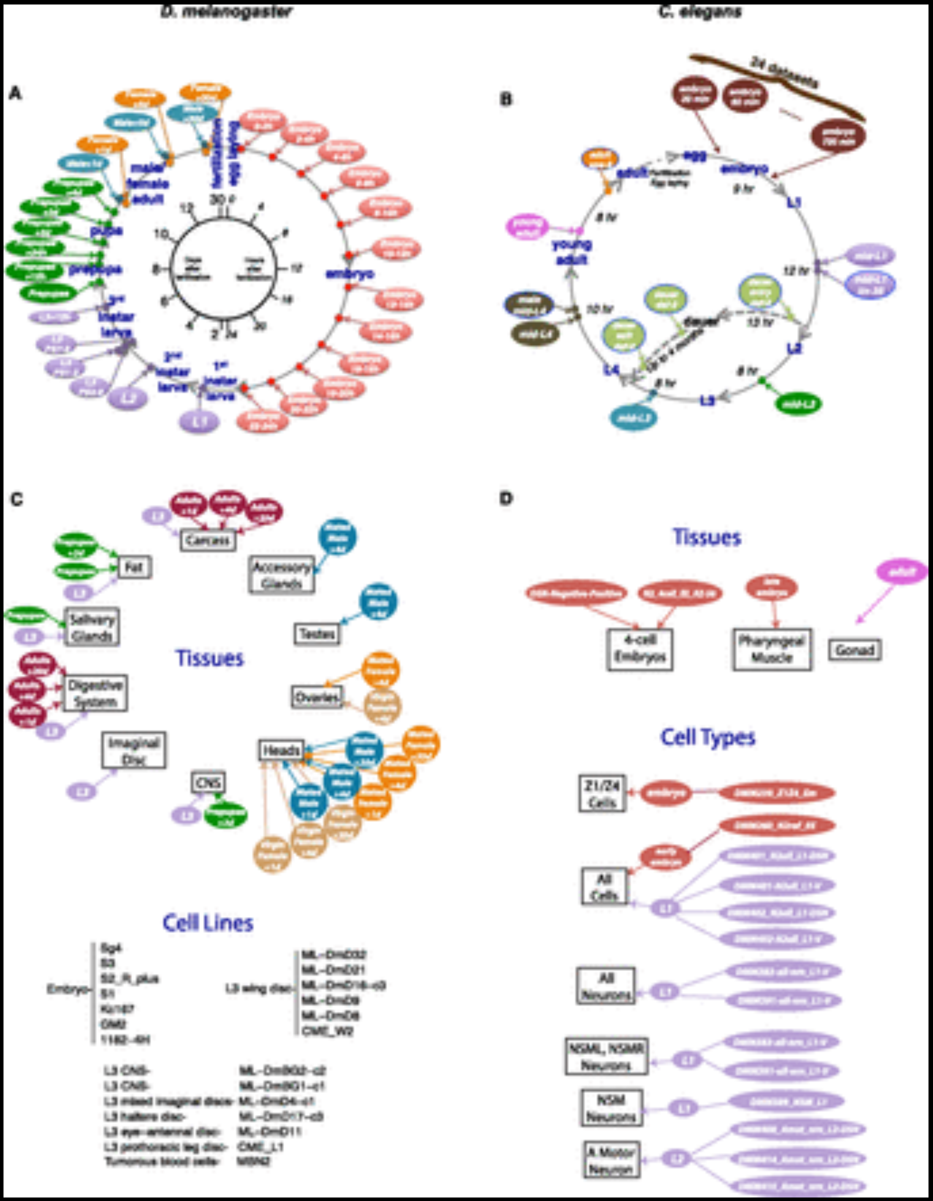
Genomic regions with exceptionally high population differentiation identified in 911 whole genomes

Multiplicity on many levels:

- Genome-wide
- SNPs, indels, SVs
- Several pairs of populations

Colonna et al. (2014) Genome Biology

Multiple testing in RNA-seq



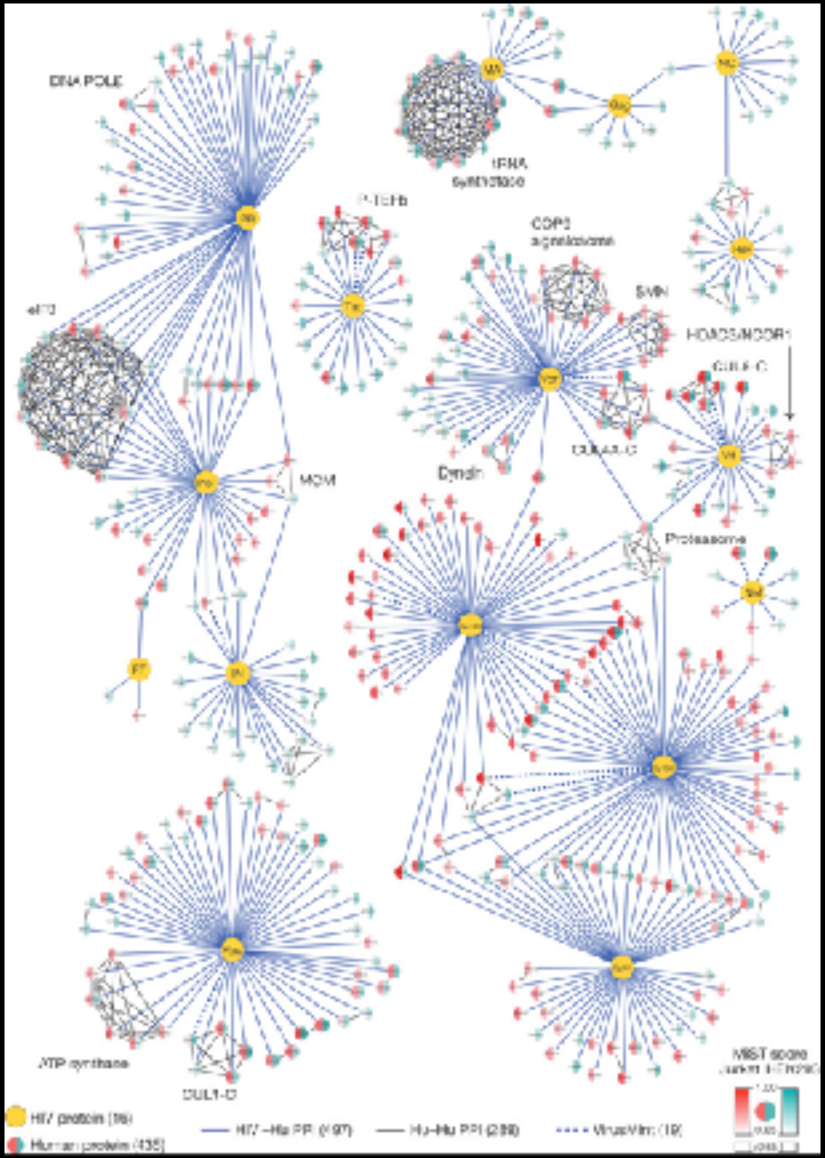
Comparison of fly and worm gene expression across developmental stages

Multiplicity on many levels:

- Two species
- Many stages
- Tissues vs. cell lines

Li et al. (2014) Genome Research

Multiple testing in mass spec



Identifying human proteins that interact with each protein in the HIV genome

Interactions mean many tests:

- Tens of HIV proteins
- Thousands of human proteins
- Many thousands of potential protein-protein interactions

Jager *et al.* (2011) Nature, 481: 365-370

Components of a Multiple Hypothesis Test

1. **Parameters:** quantity of interest
2. **Null and alternative hypotheses:** family of tests; statements about parameter values
3. **Test statistics:** quantify evidence
4. **Error rate:** control mistakes
5. **Null distribution:** assess significance
6. **Procedure:** decision rule for all tests jointly

Errors when performing one test

		Reject	Fail to Reject
H_0	True	Type I error	●
	False	●	Type II error

- P(Type I error) = α = level of significance
- P(Type II error) = β
- P(reject H_0) = power

If H_0 false, power = 1-P(Type II error) = 1- β

Errors in multiple testing

H0 is the null hypothesis

		Reject H0?		
		YES	NO	
H0 actually true?	YES	FP	TN	M0 = # true nulls
	NO	TP	FN	M1 = # false nulls
		R = # rejected nulls	M-R	M = # tests

FP = # of false positives (Type I errors)
TP = # of true positives
TN = # of true negatives
FN = # of false negatives (Type II errors)

Type I error rates

- **Per family error rate (PFER):** Expected number of false positives.

  $$PFER = E(FP)$$

- **Per comparison error rate (PCER):** Expected rate of false positives.

  $$PCER = E(FP)/m$$

Type I error rates

- **Family-wise error rate (FWER):** Probability of at least one false positive.

  $$\text{FWER} = P(\text{FP}>0)$$

- **Generalized FWER (gFWER):** Probability of at least k+1 false positives.

  $$\text{gFWER}(k) = P(\text{FP}>k)$$

Type I error rates

- **False discovery rate (FDR):** Expected proportion of false positives.

$$FDR = E(FP/R)$$

- **False discovery proportion (FDP):** Probability that the proportion of false positives is at least q.

$$FDP(q) = P(FP/R>q)$$

Null distributions for multiple testing

Distribution of the vector of test statistics if the null hypotheses were all true.

Used to convert test statistics to p-values.

Multiple testing p-values can be compared across tests, whereas statistics may be in different scales.

Different types:

same for all tests?

marginal vs. joint

parametric vs. non-parametric

Multiple Testing Procedures

Goal: Given test statistics, an error rate, significance level & a high-dimensional null distribution, make a rejection decision for every test.

- Produces a set of rejected hypotheses
- Equivalently, compute adjusted p-values
  - Related to tail probabilities of the null distribution, but must account for all the other tests so that error rate is controlled
  - Value of multiple testing error rate if reject for all statistics at least this significant

How to get adjusted p-values?

Two different approaches to control multiple testing error rate (e.g., FWER or FDR):

1. Marginal methods

   - Get usual p-values, i.e., tail probabilities under each test's null distribution

   - Adjust these probabilities based on the p-values of all other tests

Types of marginal methods

- **Single-step:** Same p-value adjustment for all hypotheses.
- **Step-wise:** Adjustments depend on observed data (test statistics).
  - Step-down = start with most significant, reduce adjustment at each step, stop at first null hypothesis not rejected
  - Step-up = start with least significant, increase adjustment at each step, stop at first rejected null hypothesis

How to get adjusted p-values?

Two different approaches to control multiple testing error rate (e.g., FWER or FDR):

1. Marginal methods
   - Get usual p-values, i.e., tail probabilities under each test's null distribution
   - Adjust these probabilities based on the p-values of all other tests
2. Joint methods directly compute adjusted p-values from a joint null distribution

Joint methods

Adjusted p-values can be computed directly from a multivariate null distribution

- Parametric (a.k.a. tabled distributions)

  Multivariate Normal distributions

  Multivariate distribution of F-statistics

- Non-parametric (i.e., resampling based)

  Permutation (2+ groups or continuous)

  Bootstrap (various types)

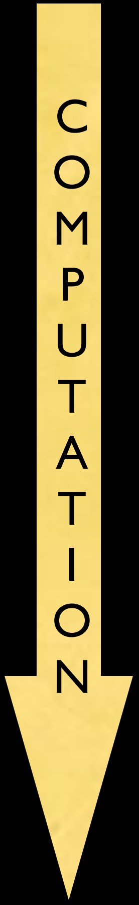
`multtest` package

`MTP` function

Resampling observations jointly

- Permutations
  - Think about the sampling unit
  - Permute label, position, location for vector of observed variables for each sampling unit
  - Scrambling the variables is a common mistake
- Bootstrap
  - Resample vectors of variables with replacement
  - Adjust the joint bootstrap distribution so that the null hypothesis holds

Multiple testing summary



COMPUTATION

- Completely marginal test

  Marginal p-values from tabled distribution or resampling one gene at a time

  Adjust with a marginal method

- Essentially marginal test

  Marginal p-values from joint distribution

  Adjust with marginal method

- Completely joint test

  Marginal and adjusted p-values from joint distribution (also test statistic cut-offs)

Testing many hypotheses at once

Large multiplicity problem: thousands of hypotheses are tested simultaneously!

Increased chance of false positives.

Chance of at least one p-value < α for N independent tests is $\boxed{1-(1-\alpha)^N}$

➔ converges to one as N increases.

e.g., For N=1,000 and α = 0.01, this chance is 0.9999568!

Individual p-values of 0.01 no longer correspond to significant findings.

Need to adjust for multiple testing when assessing the statistical significance of the observed associations.

Marginal methods: FWER controlling p-value adjustment

Name	Type	Adjustment
Bonferroni	Single-step	α/m
Sidak (ss)	Single-step	$1-(1-\alpha)^{1/m}$
Holm	Step-down	$\alpha/(m-r_j+1)$
Sidak (sd)	Step-down	$1-(1-\alpha)^{1/(m-r_j+1)}$
Hochberg	Step-up	$\alpha/(m-r_j+1)$

r_j = order statistics (ranks of test statistics)

Marginal methods: FDR controlling p-value adjustment

Name	Type	Adjustment
Benjamini & Hochberg	Step-up	$r_j\alpha/m$
Benjamini & Yekutieli	Step-up	$r_j\alpha/(m\Sigma_i i^{-1})$
Storey	Step-up	Estimates pFDR and q-value

`qvalue` package

`multtest` package

`mt.rawp2adjp` function

Joint methods for adjusted p-values

Name	Error Rate	Type	Details
ss.maxT	FWER	Single-step	Common cut-off: based on quantiles of max statistics
ss.minP	FWER	Single-step	Common quantile: based on quantiles of min p-values
sd.maxT	FWER	Step-down	Gene-specific cut-offs: based on max over subsets of T
sd.minP	FWER	Step-down	Gene-specific qtiles: based on min over subsets of P
ss.T(k+1)	gFWER	Single-step	Common cut-off: based on k+1st largest T
ss.P(k+1)	gFWER	Single-step	Common qtile: based on k+1st smallest P

`multtest` package

Implementing multivariate resampling

- Simulate two vectors of numbers (n=10 random normal variables per group) 50 times independently. Store as a 50 x 20 matrix.
- Generate b=100 permutation and bootstrap samples (50 rows). For the bootstrap, remember to standardize the original data to have mean zero in each group.
- Compute a t-statistic for each row, 100 times.
- Calculate parametric, permutation and bootstrap p-values. Compare results.
- Repeat for different means in the two groups and with correlation between the rows.

Dependence Assumptions

Independence of test statistics

Bonferroni

Benjamini & Hochberg (or PRD)

Storey

Positive orthant dependent statistics

Sidak (both versions)

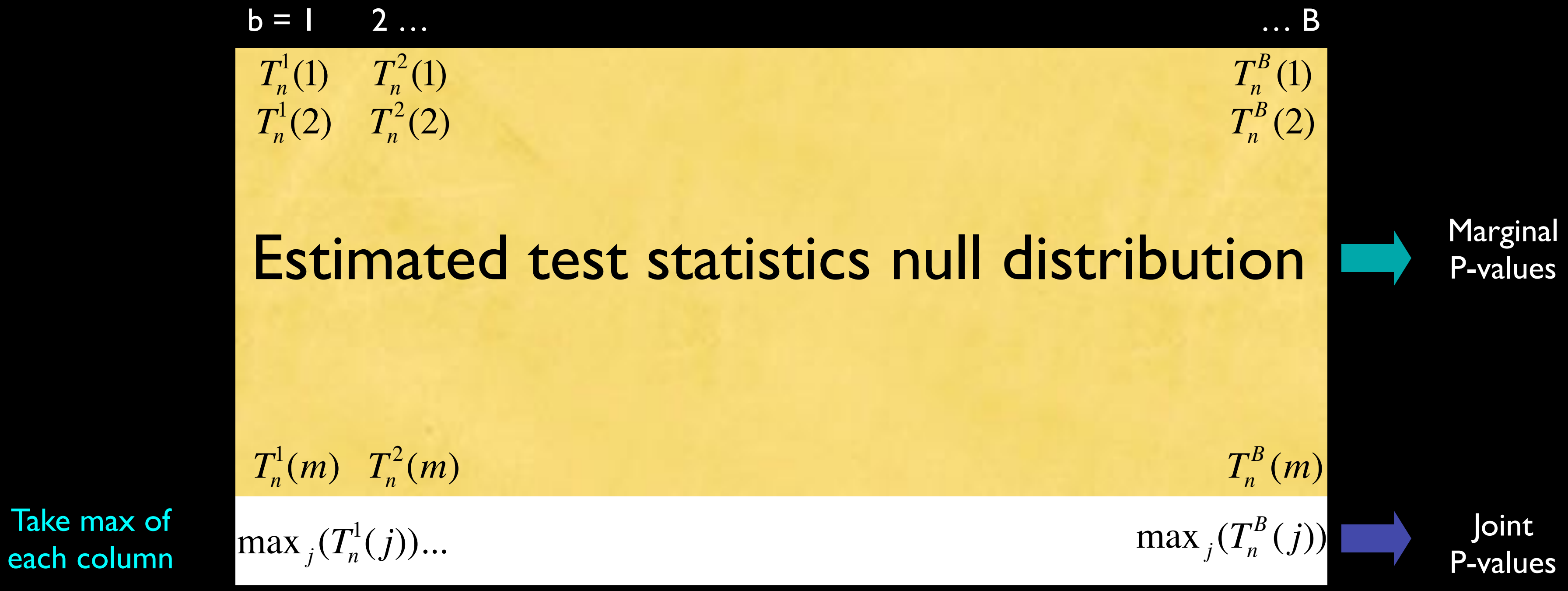
P-values satisfy Simes inequality

$$\boxed{P(p_{r_j} > \alpha r_j / m) \geq 1 - \alpha}$$

Hochberg (also assumes independence)

Joint methods for adjusted p-values

With the joint null distribution of the test statistics, direct control of Type I error rates is possible.



$$\begin{array}{llll}
b = 1 & 2 \ldots & & \ldots B \\
T_n^1(1) & T_n^2(1) & & T_n^B(1) \\
T_n^1(2) & T_n^2(2) & & T_n^B(2) \\
T_n^1(m) & T_n^2(m) & & T_n^B(m) \\
\max_j(T_n^1(j)) \ldots & & & \max_j(T_n^B(j))
\end{array}$$

Estimated test statistics null distribution

Marginal P-values

Take max of each column

Joint P-values