Summer Institutes of Statistical Genetics, 2023

Module 6: GENE EXPRESSION PROFILING

Greg Gibson and Peng Qiu

Georgia Institute of Technology

Lecture 1: EXPERIMENTAL DESIGN

greg.gibson@biology.gatech.edu

http://www.cig.gatech.edu

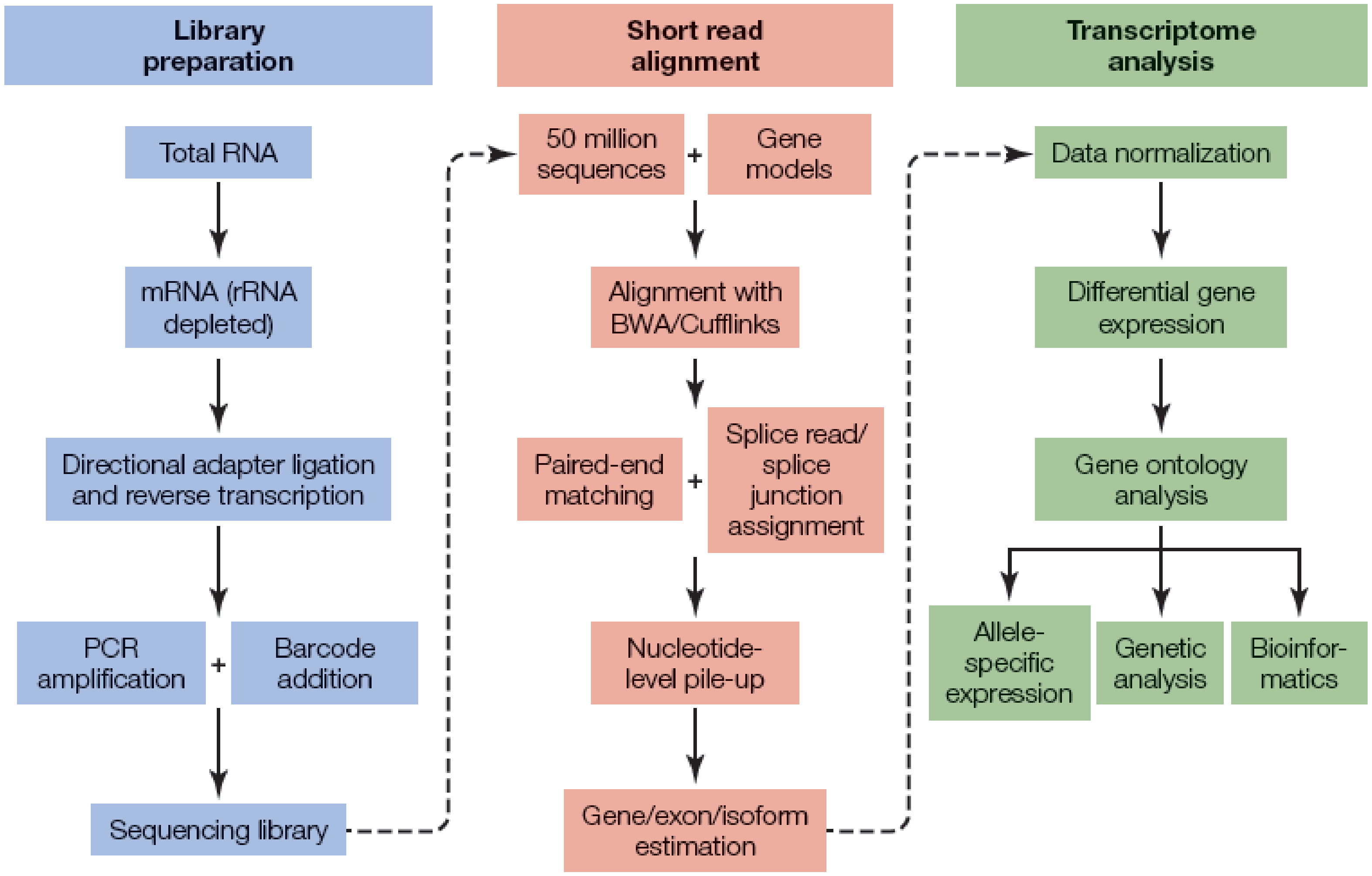
SISG Module 6 Schedule

Date	Time (PST)	Topic	Instructor
Wednesday, July 12	1:30 – 3:00	Experimental Design for Gene Expression Profiling	GG
	3:00 – 3:30	Break	
	3:30 – 5:00	Hypothesis Testing, Significance and Power	GG
Thursday, July 13	8:30 – 10:00	Foundations of Clustering and Dimension Reduction	PQ
	10:00 – 10:30	Break	
	10:30 – 12:00	Practical session with EdgeR	GG
Thursday, July 13	1:30 – 3:00	Computational Challenges in scRNAseq Analysis	PQ
	3:00 – 3:30	Break	
	3:30 – 5:00	Epigenomics	GG
Friday, July 14	8:30 – 10:00	Clustering, Trajectory Inference and Geometric Properties in scRNAseq	PQ
	10:00 – 10:30	Break	
	10:30 – 12:00	Practical session with Seurat	PQ
Friday, July 14	1:30 – 3:00	Automated Cell Type Annotation and Interpretation	PQ
	3:00 – 3:30	Break	
	3:30 – 5:00	eQTL and Genetics of Gene Expression	GG

Steps in a Gene Expression Profiling Study

1. Experimental Design	(this session)
2. RNA Sequencing	(next)
3. Short read alignment	
4. Normalization	(later this session)
5. Hypothesis testing	(after the break today)
6. Downstream analyses	(Module 16)
7. Genetic analysis	(Friday afternoon)

RNAseq Workflow



Library preparation

Total RNA

mRNA (rRNA depleted)

Directional adapter ligation and reverse transcription

PCR amplification + Barcode addition

Sequencing library

Short read alignment

50 million sequences + Gene models

Alignment with BWA/Cufflinks

Paired-end matching + Splice read/ splice junction assignment

Nucleotide-level pile-up

Gene/exon/isoform estimation

Transcriptome analysis

Data normalization

Differential gene expression

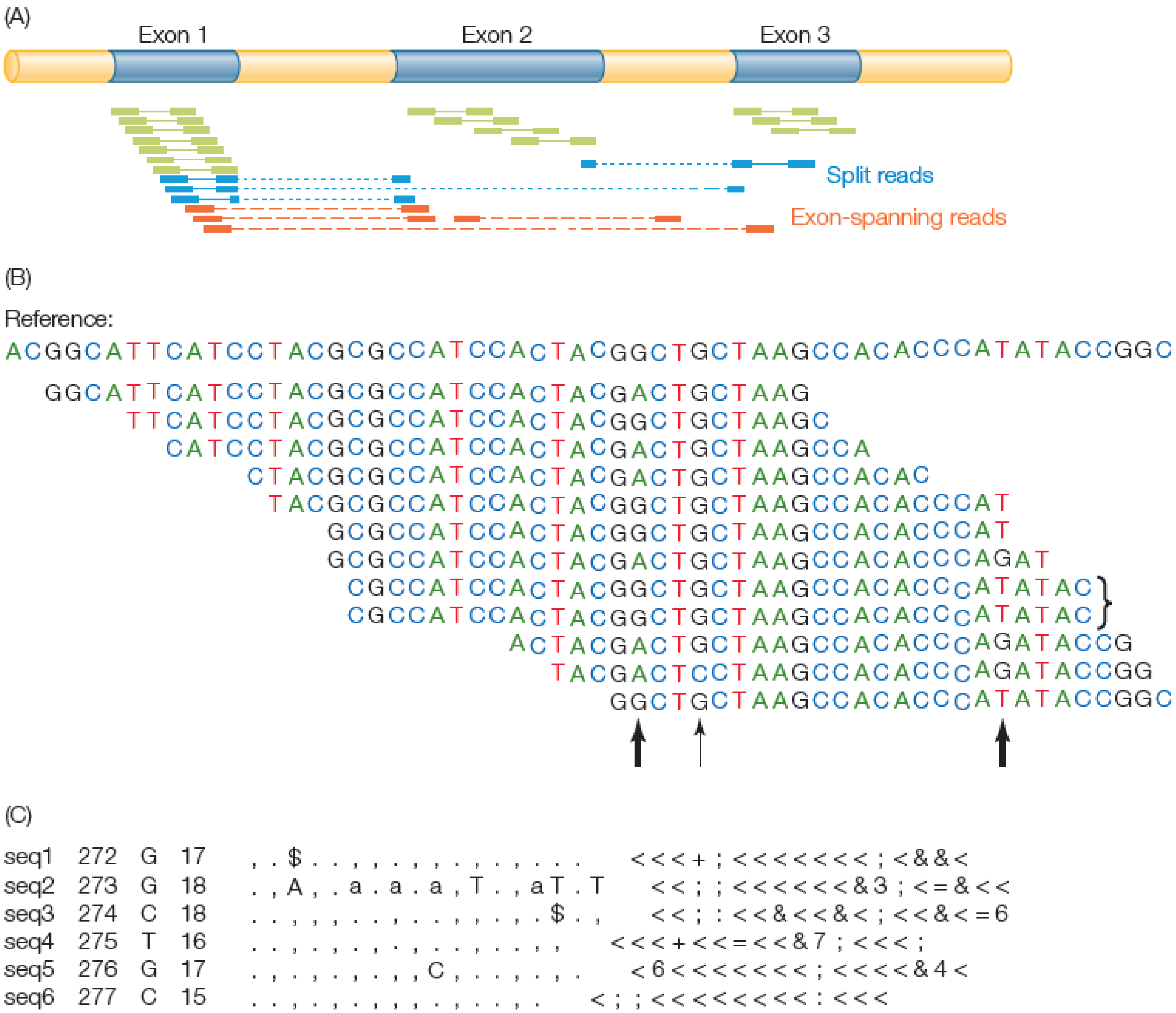
Gene ontology analysis

Allele-specific expression

Genetic analysis

Bioinfor-matics

Read Alignment



Reference:

```
ACGGCATTCATCCTACGCGCCATCCACTACGGCTGCTAAGCCACACCCATATACCGGC

  GGCATTCATCCTACGCGCCATCCACTACGACTGCTAAG
      TTCATCCTACGCGCCATCCACTACGGCTGCTAAGC
        CATCCTACGCGCCATCCACTACGACTGCTAAGCCA
            CTACGCGCCATCCACTACGACTGCTAAGCCACAC
             TACGCGCCATCCACTACGGCTGCTAAGCCACACCCAT
                GCGCCATCCACTACGGCTGCTAAGCCACACCCAT
                GCGCCATCCACTACGACTGCTAAGCCACACCCAGAT
                 CGCCATCCACTACGGCTGCTAAGCCACACCCATATAC }
                 CGCCATCCACTACGGCTGCTAAGCCACACCCATATAC
                         ACTACGACTGCTAAGCCACACCCAGATACCG
                           TACGACTCCTAAGCCACACCCAGATACCGG
                              GGCTGCTAAGCCACACCCATATACCGGC
                               ↑  ↑              ↑
```

```
seq1  272  G  17   , . $ . . , , . , , . , , . , . . .      < < < + ; < < < < < < < ; < & & <
seq2  273  G  18   . , A , . a . a . a , T . , a T . T      < < ; ; < < < < < < & 3 ; < = & < <
seq3  274  C  18   . . , , , , . . , . . . , , . , . $ . ,    < < ; : < < & < < & < ; < < & < = 6
seq4  275  T  16   . . , . . , . , . , . . , . . , ,      < < < + < < = < < & 7 ; < < < ;
seq5  276  G  17   . , . . , . , . . , , C , . . , . , .      < 6 < < < < < < < ; < < < < & 4 <
seq6  277  C  15   . . , . . , . , , , , . , . . . , .      < ; ; < < < < < < < < : < < <
```

Modes of Bulk RNA sequencing

RNA is prepared, mRNA is captured on polyT beads, fragmented, and converted to cDNA using either a stranded or unstranded protocol, usually with 12-24X multiplexing

1. Single-end reads
   - ➢ Maximizes the total number of independent reads (50M optimal)
   - ➢ When RNA is degraded, eg FFPE specimens

2. Paired-end reads
   - ➢ Slightly more accurate alignment
   - ➢ But typically lower coverage (25M reads)
   - ➢ Better for estimation of alternate splicing and ASE

3. 3’ targeted
   - ➢ Lexogen protocol is one fifth the cost ($70 vs $350 per sample)
   - ➢ Ideal for large sample studies when funds are a concern
   - ➢ Single Cell drop digital dd-scRNASeq is also 3’ targeted

RNAseq Software

1. Short Read Alignment

STAR	https://github.com/alexdobin/STAR/releases
HISAT2	https://ccb.jhu.edu/software/hisat2/index.shtml

2. Read counting

HTseq	http://www-huber.embl.de/HTSeq/doc/overview.html
SAMtools	http://www.htslib.org/

3. Differential Expression

DESeq	https://bioconductor.org/packages/release/bioc/html/DESeq2.html
DExSeq	https://www.bioconductor.org/packages/release/bioc/html/DEXSeq.html
edgeR	https://bioconductor.org/packages/release/bioc/html/edgeR.html
Voom	http://web.mit.edu/~r/current/arch/i386_linux26/lib/R/library/limma/html/voom.html

4. Data Normalization

SVASeq	https://www.bioconductor.org/packages/release/bioc/html/sva.html
Combat	https://www.rdocumentation.org/packages/sva/versions/3.20.0/topics/ComBat
PEER	http://www.sanger.ac.uk/science/tools/peer
SNM	https://www.bioconductor.org/packages/release/bioc/html/snm.html

Another option is the Tuxedo protocol (Bowtie, Tophat, Cufflinks, Cuffdiff,
https://ugene.net/wiki/display/WDD31/RNA-seq+Analysis+with+Tuxedo+Tools

Notes on Alignment Biases

1. Adapter trimming can alter read alignment and counts – pay attention to all steps in workflow
2. Very short reads make it more difficult to align across introns: software biases exist in old datasets
3. Some regions of the genome are known not to align well, often due to repeats or odd GC content
4. There is a global tendency for reads to align preferentially to the reference allele
    - Consequently, average allelic expression % is closer to 60-40 (or 55-45) rather than 50-50
    - This may affect pile-ups of homologous genes that differ by just a SNP or two
5. Alignment algorithms treat things like whether the library is stranded differently

    (stranded libraries allow you to evaluate whether the Watson or Crick strand was transcribed)

Basics of Experimental Design: Levels of Replication

Often you will have a fixed budget that constrains how many arrays can be processed. So your first task is to determine what levels of replication you can afford, and how they will impact statistical power.

Technical Replication:

- RNA preparation (eg. from adjacent biopsies)
- cDNA synthesis (pooling minimizes outlier effects)
- library preparation
- sequencing lane or array hybridization (usually a minimal effect)

Biological Replication:

Fixed effects:

- sex
- treatment (drug, growth regimen, tissue)
- time of sampling (repeated measures in some cases)
- genotype (IF specifically chosen and resampled)

Random effects

- individual from a population
- field plot

Basics of Experimental Design: Specifying Contrasts of Interest

At the same time, you need to be aware of the contrasts you wish to make since by tweaking the design you may gain a lot in terms of what you can infer.

Suppose you want to compare B cells and T cells from Healthy controls and COVID-19 patients, and you have the funds to generate 24 RNASeq profiles

What is the best design?

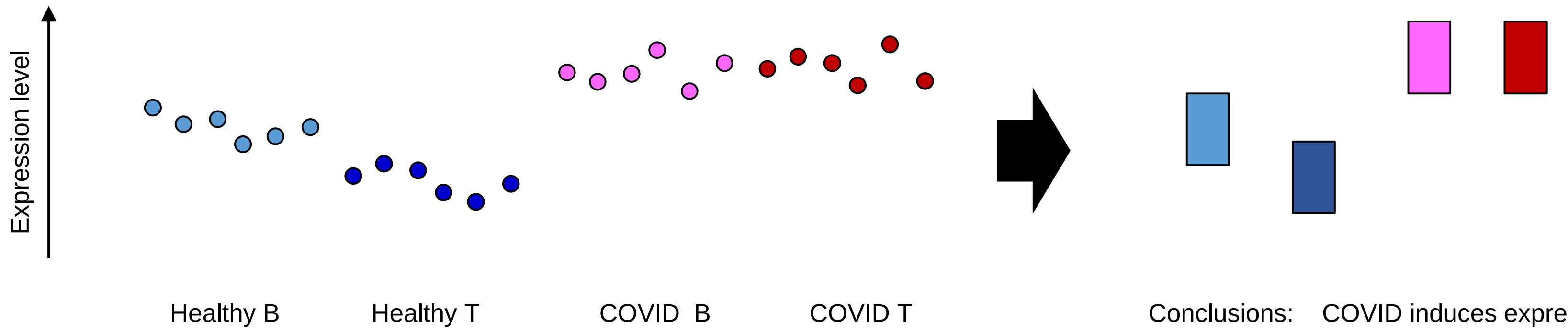
- 6 controls and 6 patients, each donating both a B and a T cell sample
- 12 controls and 12 patients, each donating either a B or a T cell sample
- 3 controls and 3 patients, each donating a B and a T cell sample, processed twice
- 3 controls and 3 patients, each donating 2 B and 2 T cell samples, on separate days
- same as above, but only men or only women
- 12 controls and 12 patients, each donating either a B or a T cell sample, but pooling two visits

Main effects can only be contrasted if you have biological replicates:

reducing the number of individuals may allow you to address intra-individual variability

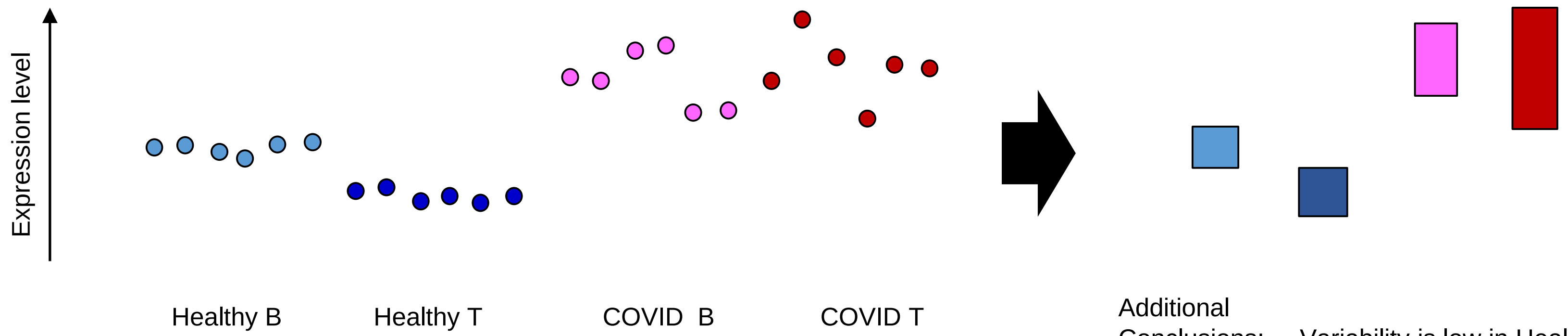
Interaction effects allow you to ask questions like whether B cells and T cells differ more between healthy volunteers or patients

Two Hypothetical Sets of Results Illustrating Design Principles



Conclusions: COVID induces expression

T < B only in healthy people



Additional Conclusions: Variability is low in Healthy controls

COVID B cell response is individualized

COVID T cell response is hypervariable

Basics of Experimental Design: Confounding

At the design step, avoid confounding biological factors:

- don’t contrast bloods from young males and old females
- don’t contrast hearts from normal mice and livers from obese ones

as far as possible, balance all biological factors

Be aware of the potential for technical confounding:

- date of RNA extraction or sequencing run
- batch of samples (particularly for microarray studies)
- person who prepared the libraries
- SE or PE, read length and quality of reads
- quality of RNA (RIN = Bioanalyzer RNA Integrity Number)

Fundamentals of Hypothesis Testing: Linear Modeling

1. Normalize the samples:

   log(fluorescence) = μ + Array + Residual

   log(CPM) = (read counts + 1) × 10^6 / total read counts

   OR variance transforms, OR supervised methods

2. For each gene, assess significance of treatment effects on the Residual or the log(CPM) (ie. expression level) with a linear model:

Expression = μ + Sex + Genotype + Treatment + Interaction + Error

OR likelihood ratio test or Wilcoxon Rank-Sum test, etc

Fundamentals of Hypothesis Testing: Gene-level Contrasts

1. Generally we are interested in asking whether there is a significant difference between two or more treatment group(s) on a gene-by-gene basis

2. For a simple contrast, we can use a t-test to test the hypothesis. Significance is always a function of:
    1. The difference between the two groups: [5,6,4] vs [7,5,6] has a diff of 1
    2. The variance within the groups: [2,5,8] vs [3,6,9] does as well, but is less obvious
    3. The sample size: [5,6,4,4,6,5] vs [7,5,6,5,6,7] is better

3. For contrasts involving multiple effects, we usually use General Linear Models in the ANOVA framework (analysis of variance: significance is assessed as the F ratio of the between sample to residual sample variance.

4. edgeR uses limma to perform One-Way ANOVAs. This likelihood framework is very powerful, but constrains you to contrast treatments with a reference

5. Robust statistics (eg using lme4, glmmSeq) also allow you to evaluate INTERACTION EFFECTS, namely not just whether two treatments are individual significant, but also whether one depends on the other

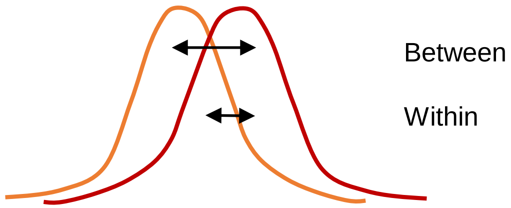
6. Given a list of p-values and DE estimates, we need to evaluate a significance threshold, which is usually done using False Discovery Rate (FDR) criteria, either Benjamini-Hochberg or a qvalue

Fundamentals of Hypothesis Testing: t-tests and ANOVA

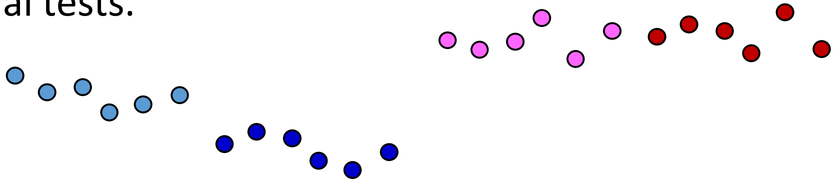
A two-sample t-test evaluates whether the two population means differ one another, given the pooled standard deviations of the sample and the number of observations:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_p \sqrt{2/n}} \qquad s_p = \sqrt{\frac{s^2_{X_1} + s^2_{X_2}}{2}}$$

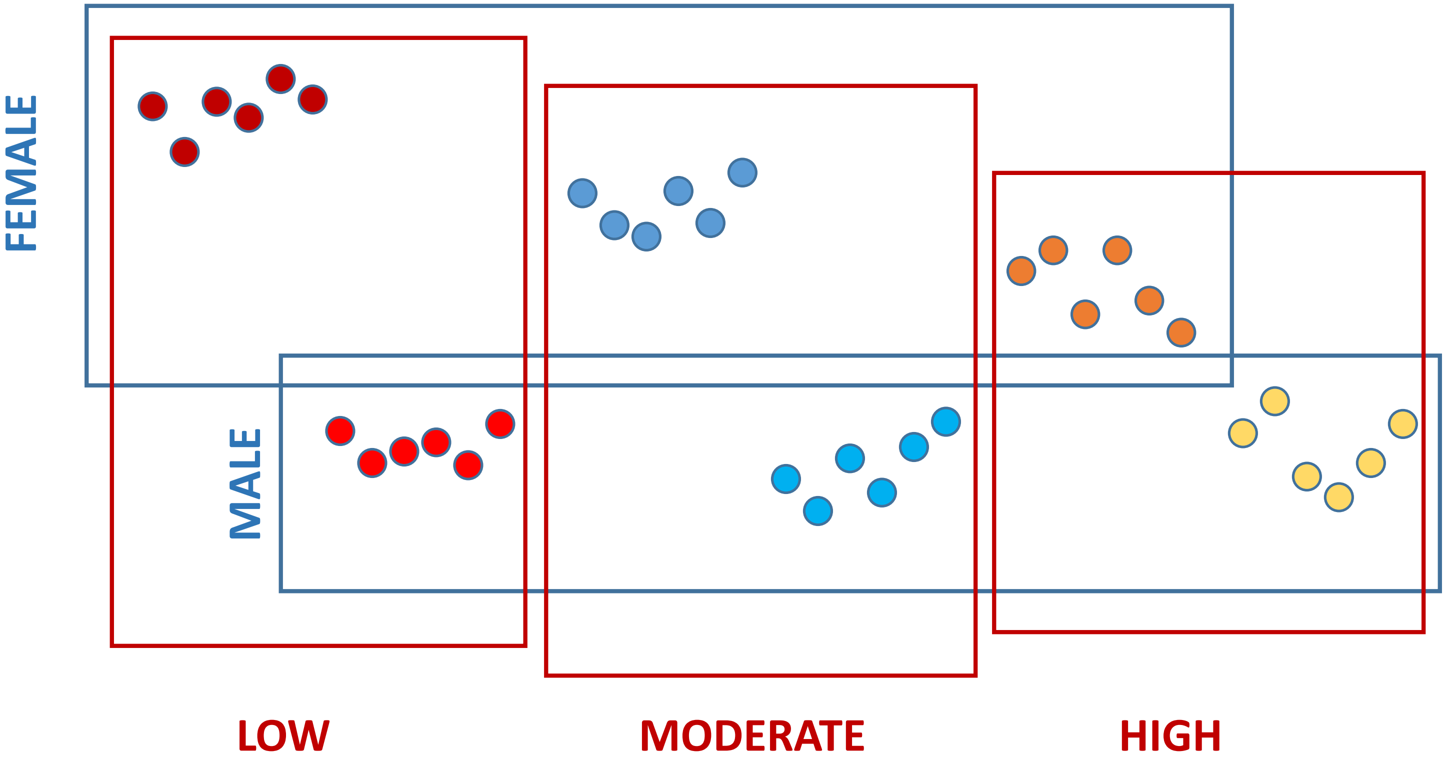
F-statistics generalize this approach by asking whether the deviations between samples samples are large relative to the deviations within samples. The larger the difference between the means, the larger the F-ratio, hence significance is evaluated by contrasting variances.



Generally we start by evaluating whether there is variation among all the groups, and then test specific post-hoc contrasts. With multifactorial designs you also should be aware of the difference between Type I and Type III sums of squares, namely univariate and conditional tests.



Type of Modeling I: Pairwise Contrasts



Sex: $p<0.00001$

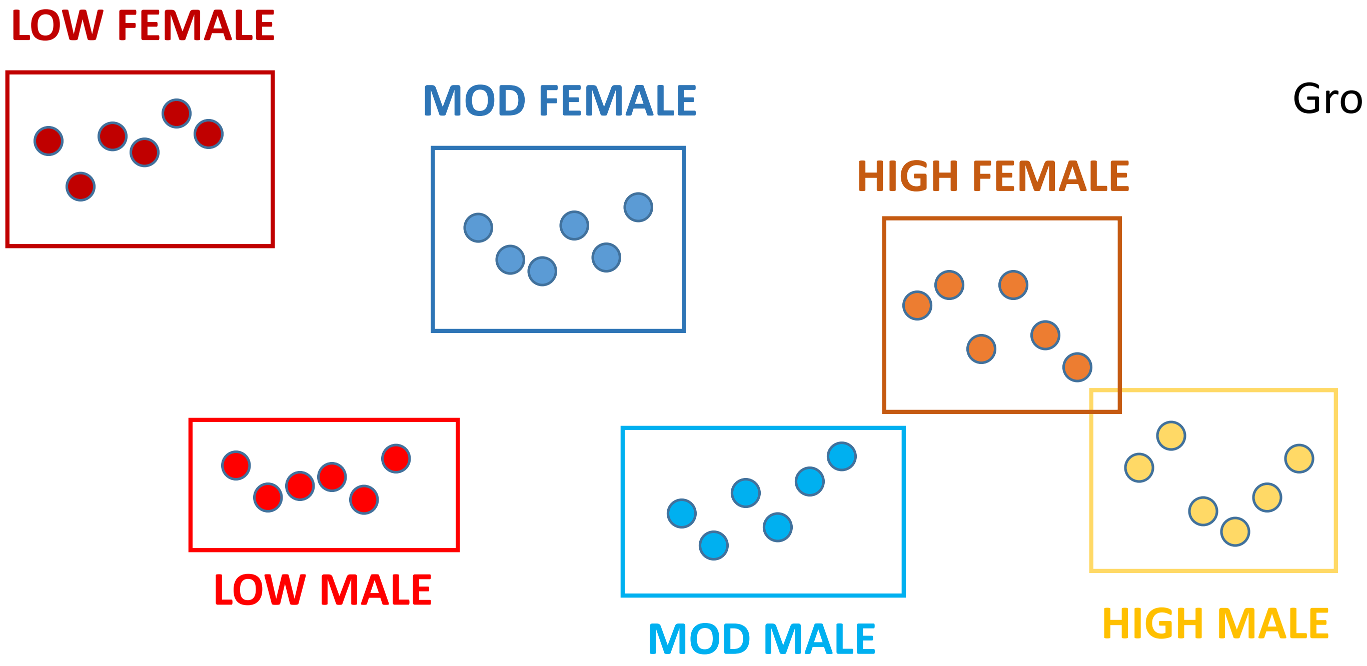
Condition: $p = 0.135$

Joint Model:

Sex: $p<0.00001$

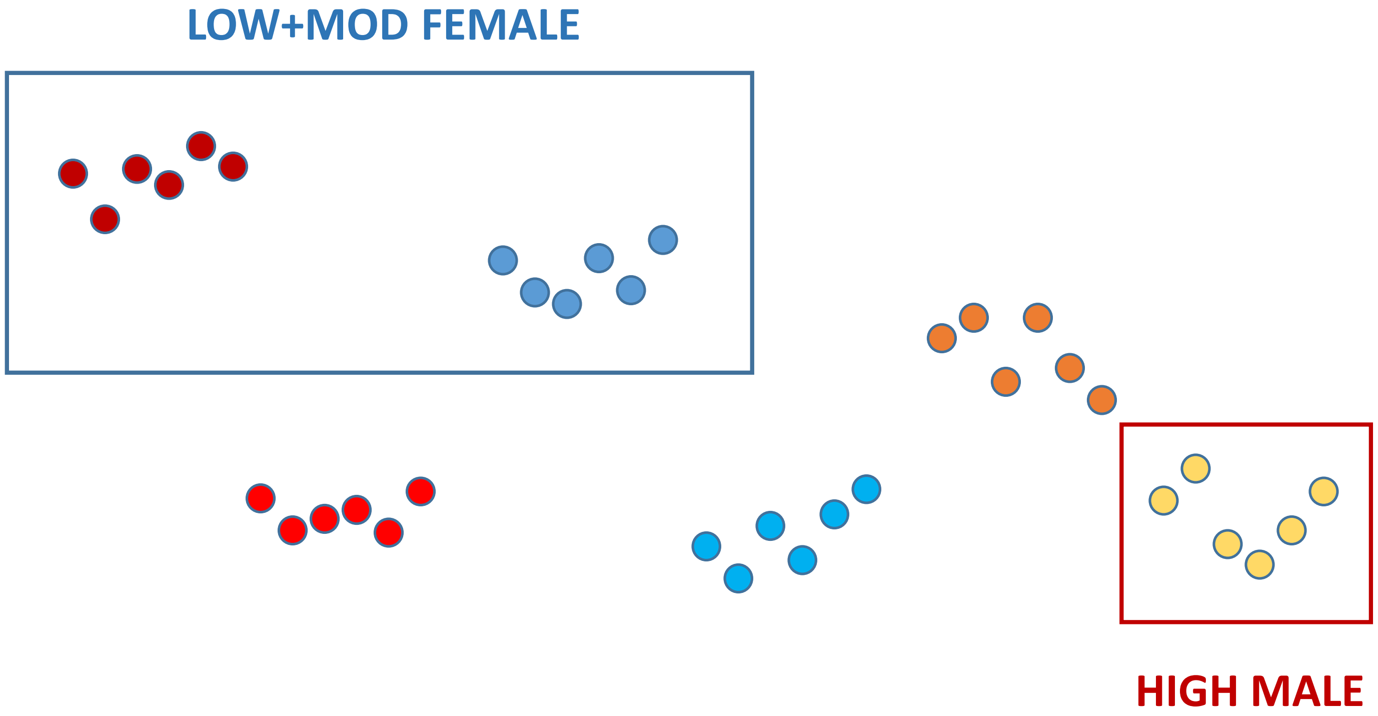
Condition: $p = 0.0002$

Type of Modeling II: One-Way ANOVA

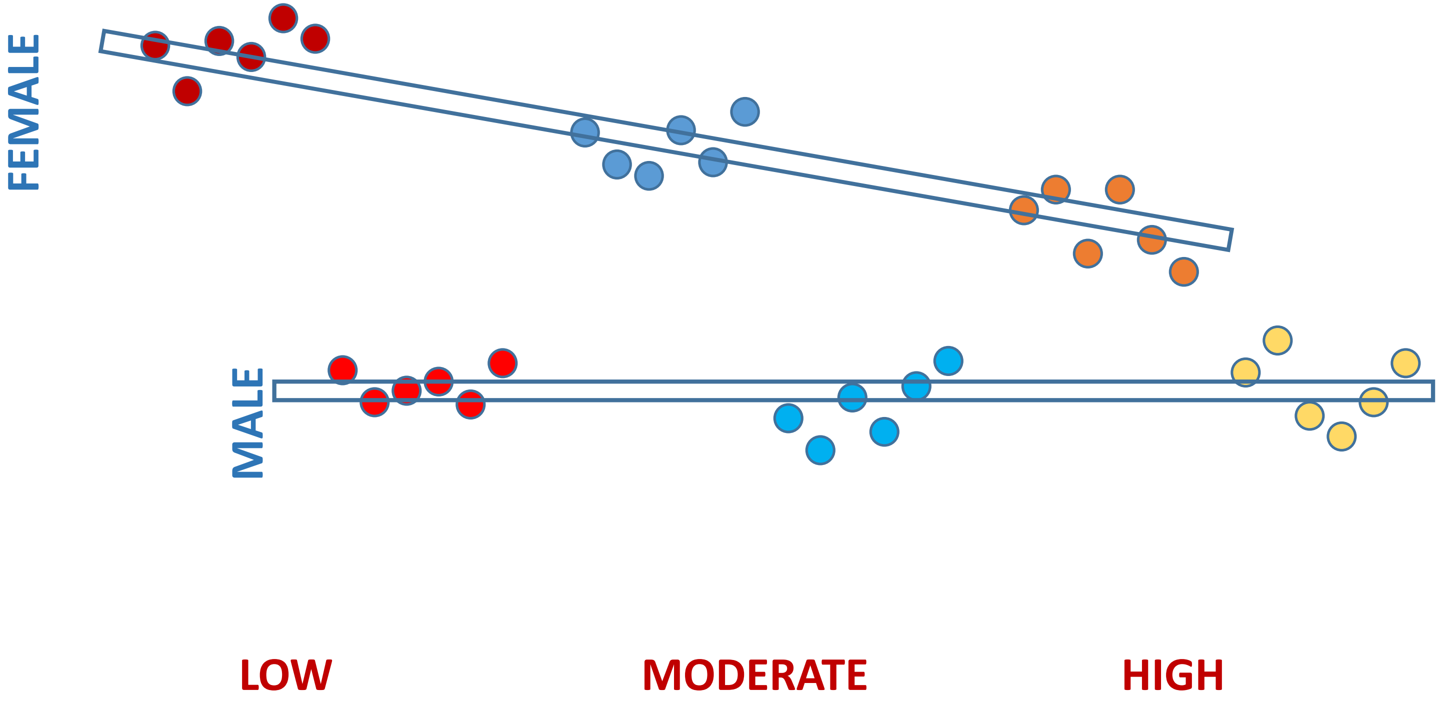


Group: p<0.00001

Type of Modeling III: Post-Hoc Contrasts



Type of Modeling IV: 2-Way ANOVA with Interactions



Fixed Effect Model:

Sex: $p<0.00001$

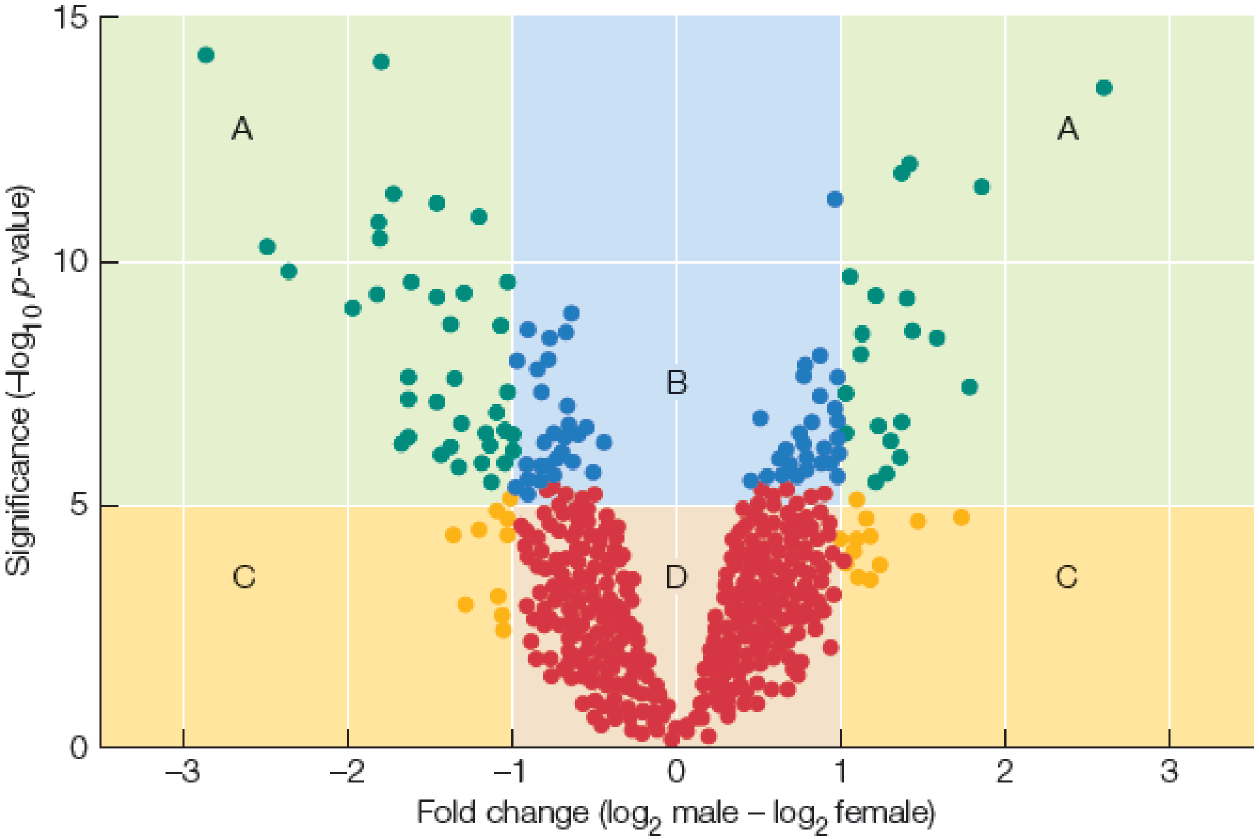
Condition: $p < 0.00001$

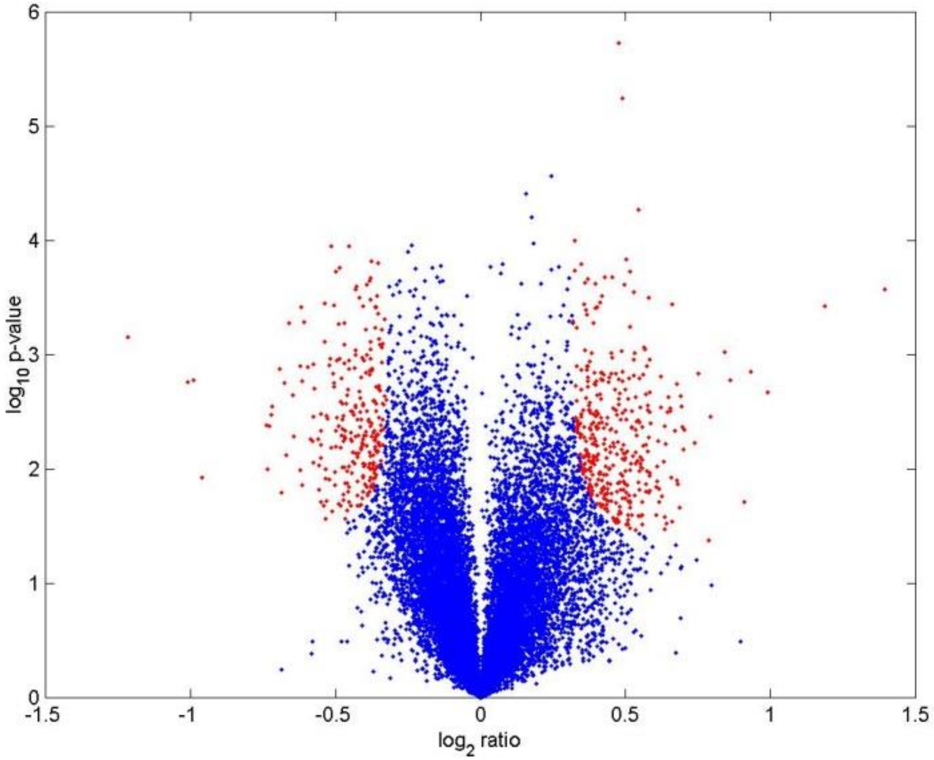
Interaction: $p = 0.00001$

Random Condition Model:

Sex: $p=0.06$

Volcano Plots: Significance against Fold-Change





The Significance Threshold at NLP = 5 (p<10-5) is conservatively Bonferroni corrected for 5,000 genes.

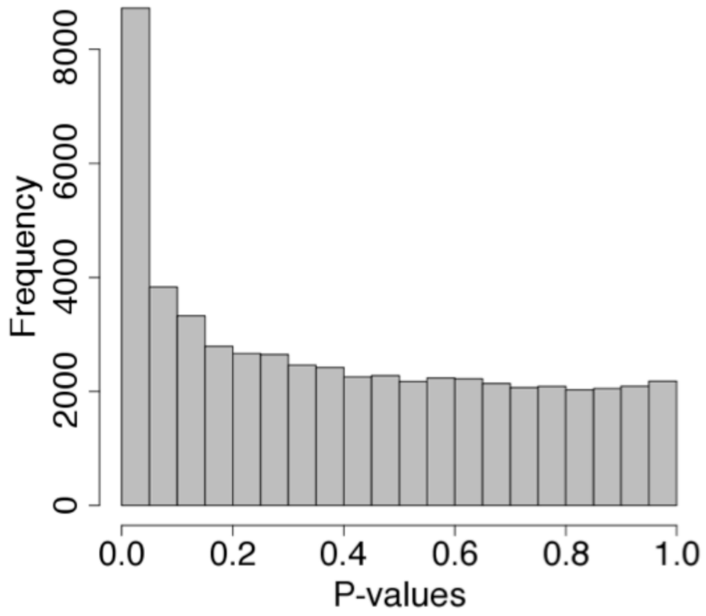
While significance testing allows for inclusion of more genes that have small fold-changes (sector B), you need to be aware that there is also variation in the estimation of the denominator (variance) which can inflate NLP values.

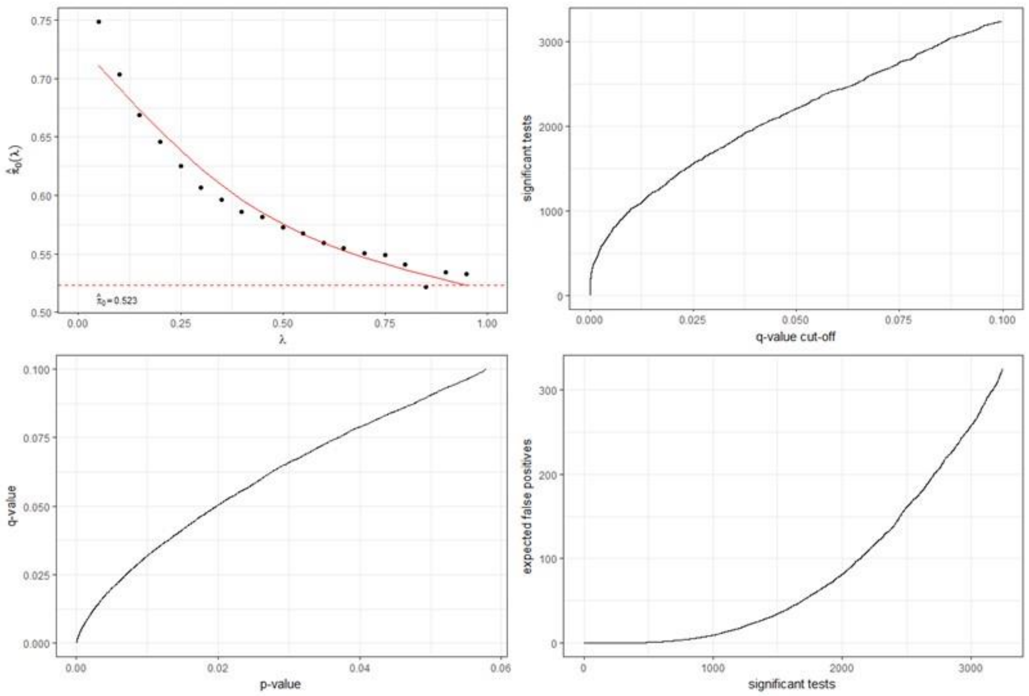
FDR and qValues

Traditional B-H (Benjamin-Hochberg) False Discovery Rates simply evaluate the difference between the observed and expected number of positives at a given threshold, in order to estimate the proportion of false positives.

Eg. In 10,000 tests, you expect 100 p-values < 0.01. If you observe 1000, then the FDR is 100/1000 = 10%.

The q-value procedure estimates the proportion of true negatives from the distribution of p-values.





https://www.bioconductor.org/packages/release/bioc/html/qvalue.html