Summer Institutes of Statistical Genetics, 2022

Module 6: GENE EXPRESSION PROFILING

Greg Gibson and Peng Qiu

Georgia Institute of Technology

Lecture 2: HYPOTHESIS TESTING

greg.gibson@biology.gatech.edu

http://www.cig.gatech.edu

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

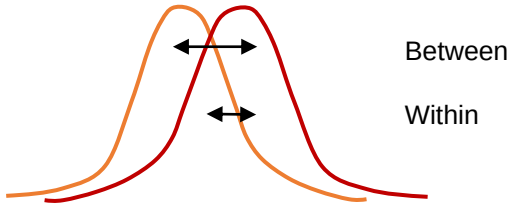
6. Given a list of p-values and DE estimates, we need to evaluate a significance threshold, which is usually done using False Discovery Rate (FDR) criteria, either Benjamini-Hochberg or a qvalue

Fundamentals of Hypothesis Testing: t-tests and ANOVA

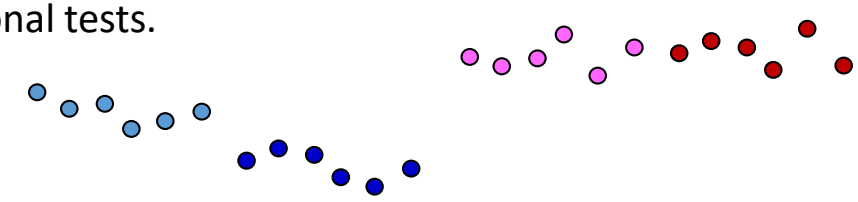
A two-sample t-test evaluates whether the two population means differ one another, given the pooled standard deviations of the sample and the number of observations:

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

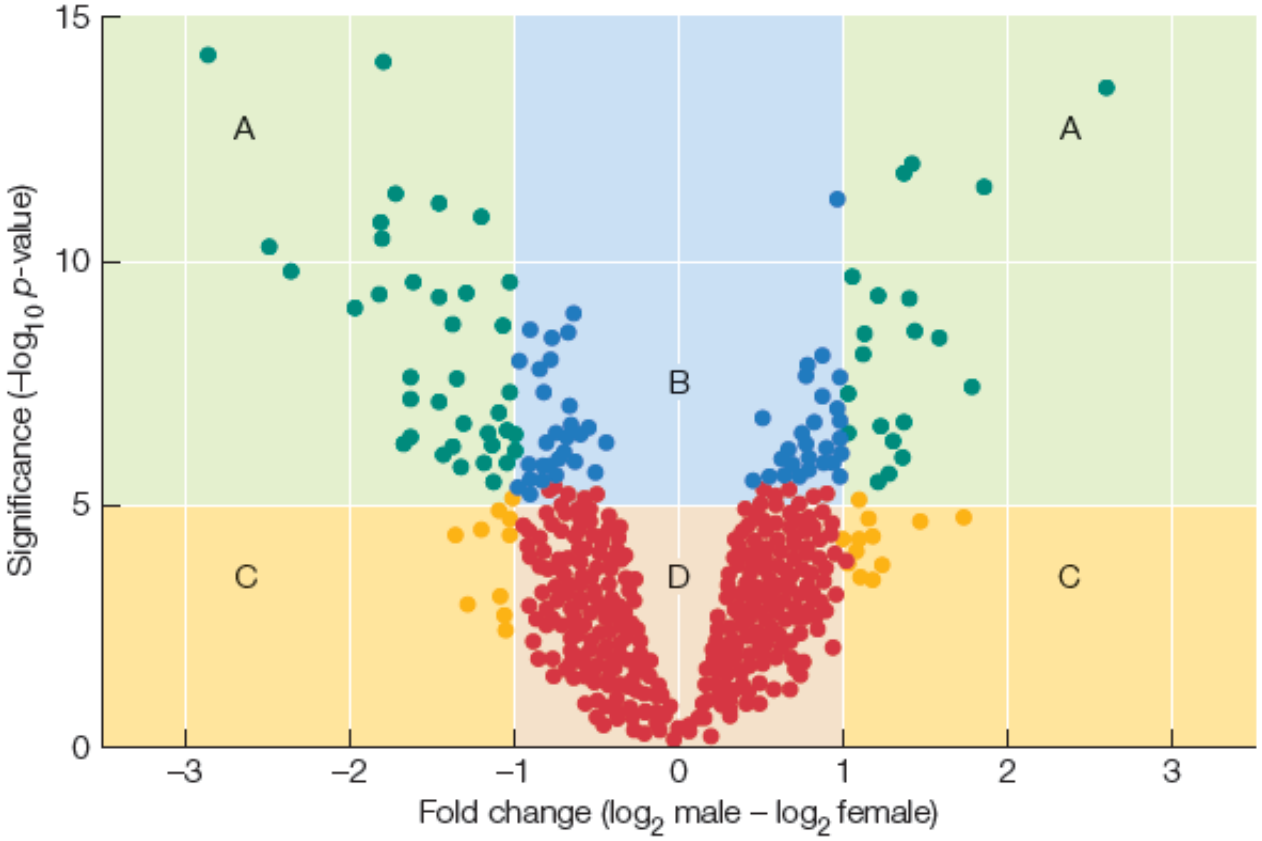
F-statistics generalize this approach by asking whether the deviations between samples samples are large relative to the deviations within samples. The larger the difference between the means, the larger the F-ratio, hence significance is evaluated by contrasting variances.

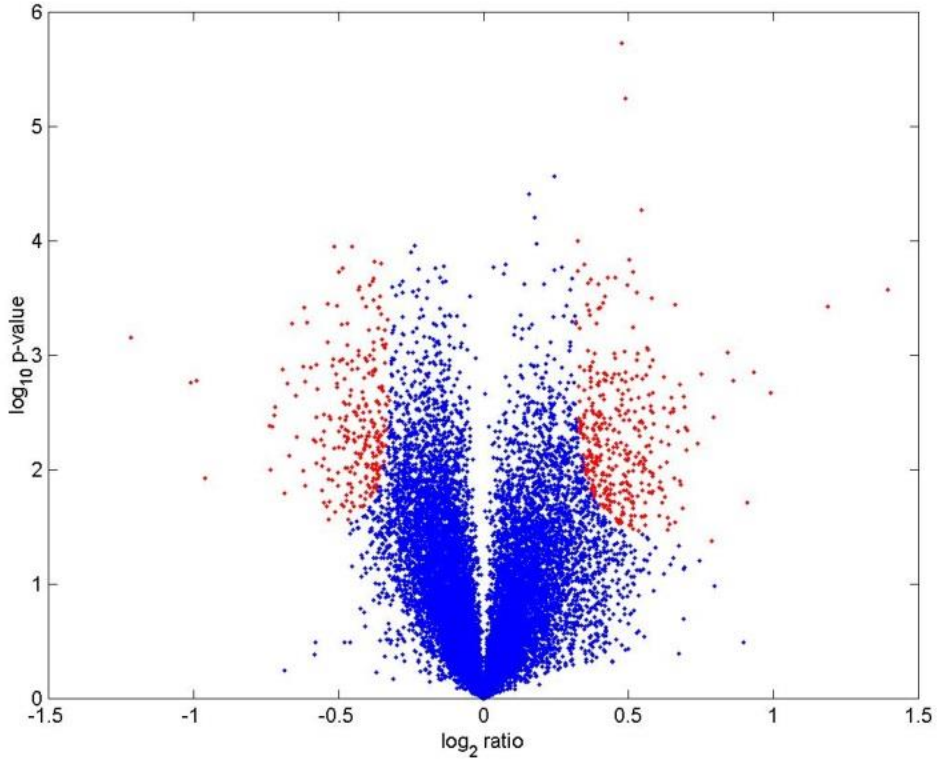


Generally we start by evaluating whether there is variation among all the groups, and then test specific post-hoc contrasts. With multifactorial designs you also should be aware of the difference between Type I and Type III sums of squares, namely univariate and conditional tests.



Volcano Plots: Significance against Fold-Change





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

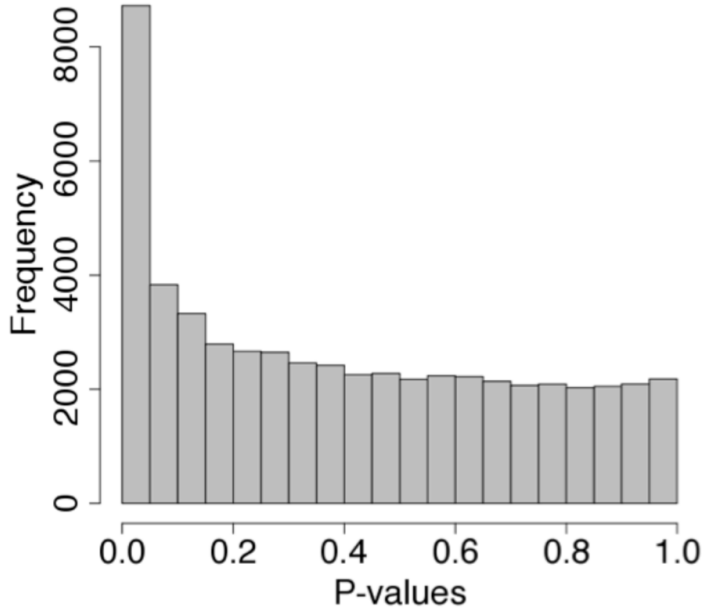
While significance testing allows for inclusion of more genes that have small fold-changes (sector B), you need to be aware that there is also variation in the estimation of the denominator (variance) which can inflate NLP values.

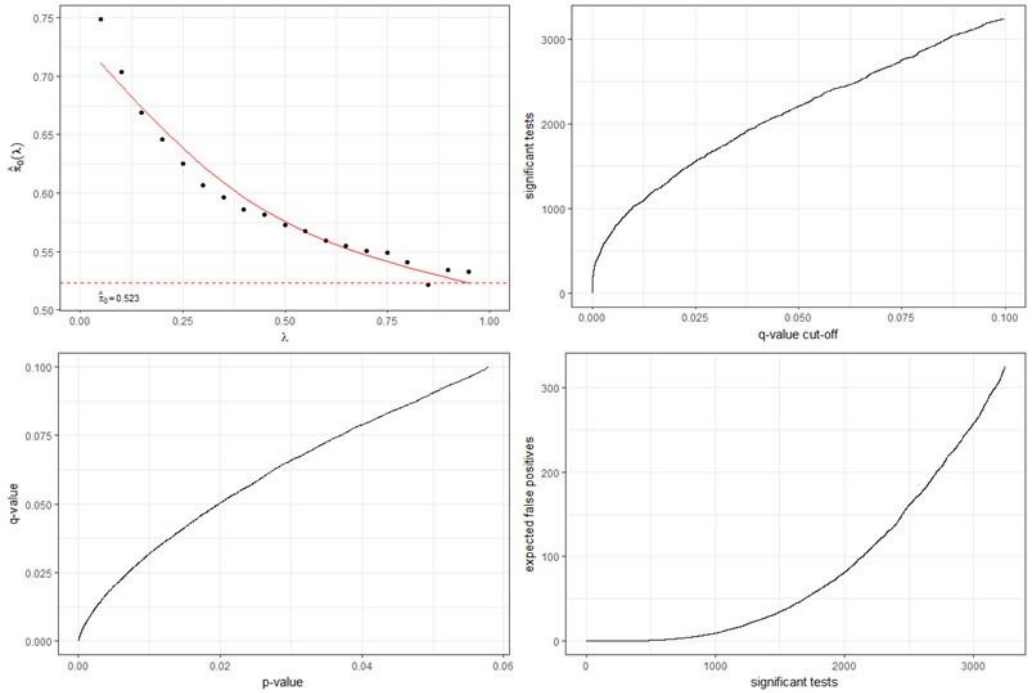
FDR and qValues

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

edgeR User Guide

1.4 Quick start

edgeR offers many variants on analyses. The glm approach is more popular than the classic approach as it offers great flexibilities. There are two testing methods under the glm framework: likelihood ratio tests and quasi-likelihood F-tests. The quasi-likelihood method is highly recommended for differential expression analyses of bulk RNA-seq data as it gives stricter error rate control by accounting for the uncertainty in dispersion estimation. The likelihood ratio test can be useful in some special cases such as single cell RNA-seq and datasets with no replicates. The details of these methods are described in Chapter 2.

A typical edgeR analysis might look like the following. Here we assume there are four RNA-Seq libraries in two groups, and the counts are stored in a tab-delimited text file, with gene symbols in a column called `Symbol`.

```
> x <- read.delim("TableOfCounts.txt",row.names="Symbol")
> group <- factor(c(1,1,2,2))
> y <- DGEList(counts=x,group=group)
> keep <- filterByExpr(y)
> y <- y[keep,,keep.lib.sizes=FALSE]
> y <- calcNormFactors(y)
> design <- model.matrix(~group)
> y <- estimateDisp(y,design)
```

To perform quasi-likelihood F-tests:

```
> fit <- glmQLFit(y,design)
> qlf <- glmQLFTest(fit,coef=2)
> topTags(qlf)
```

To perform likelihood ratio tests:

```
> fit <- glmFit(y,design)
> lrt <- glmLRT(fit,coef=2)
> topTags(lrt)
```

Hypothesis Testing in edgeR

If you compare the output of Likelihood Ratio tests in edgeR to those of a t-test or F-test, particularly for small n, you will often get *very* different results. There are at least 4 reasons for this.

1. edgeR performs a TMM normalization. Since RNASeq generates counts, we adjust for library size by computing cpm (counts per million). If a few high abundance transcripts vary by several percent, they throw off all the other estimates. TMM fixes this.

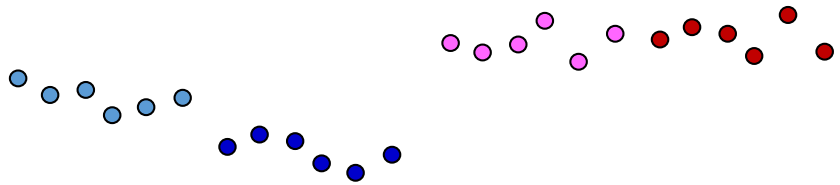
2. edgeR shrinks the variance of low abundance transcripts by fitting the distribution to the "negative binomial" expectation. Basically it adds values to account for sampling error at the low end so that comparing 0, 1 and 2 is more like comparing 10, 11 and 12.

3. edgeR also employs a powerful Bayesian within-sample variance adjustment in its GLM fitting, with the result that it puts much more weight on fold-change than standard F-tests.

4. For a one-way ANOVA the approaches are similar (though you need to be careful about whether you fit an intercept, which means you compare multiple samples to a reference rather than to one another). For more complicated analyses involving two or more factors, nesting, and random effects, the modeling frameworks are really quite different. The Quasi-Likelihood F-test now used in EdgeR adjusts for error in estimating dispersion of each gene.

I prefer to run ProcGLM or ProcMIXED in SAS or JMP-Genomics; the equivalent in R is lme4, but needs looping of the code, eg using glmmSeq.

Making Specific Contrasts



In this 2x2 design, there are two cell types and two treatments, each with three duplicate samples.

Likelihood ratio tests allow you to model specific contrasts under the assumption that an effect of interest is present, relative to the case of no effect. If the effect is significant, more of the variance is explained by the model than you'd expect given the dispersion in the total dataset.

Your conclusion is a function of what you specify as the reference and contrast.

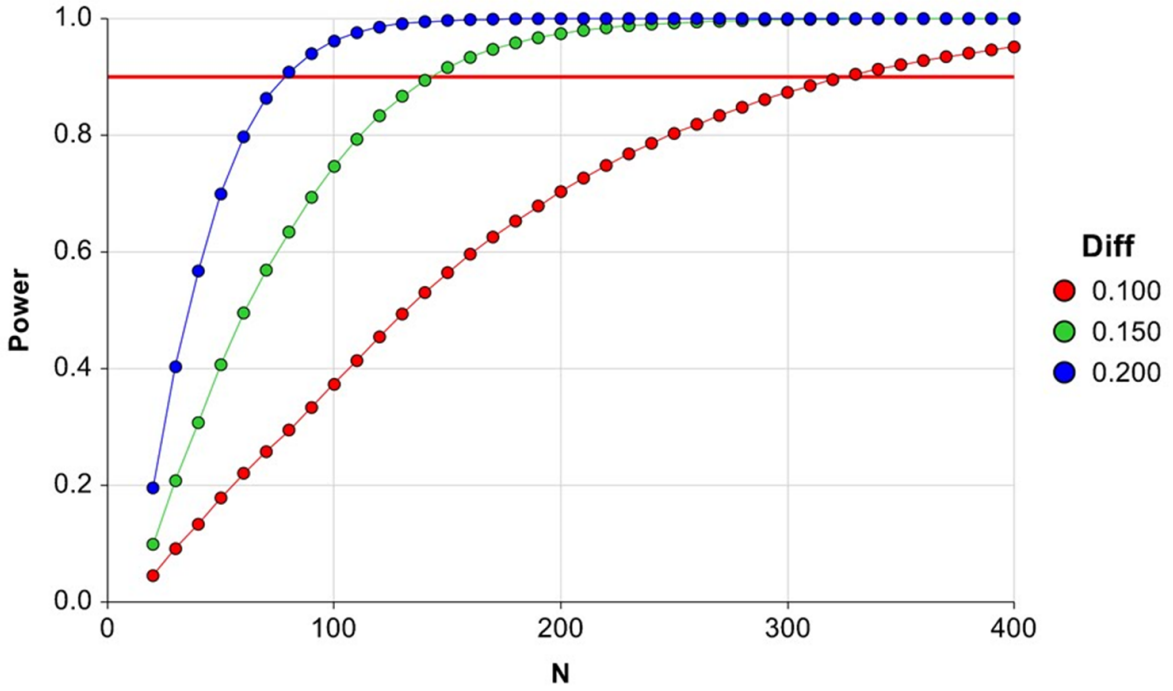
I can ask, are pink and red different from light and dark blue (is there a COVID effect)
Or, are dark blue and red different from light blue and pink (is there a B vs T cell effect)
Or, are dark red different from all others (are COVID B-cells affected), which gives a different answer than are red different from dark blue (are COVID B-cells unlike control B-cells). You use the Model matrix to assign weights (0, 1, -1 etc) to specify whatever contrast you wish.

Fitting more complex models with random effects of individual, and nesting, and interaction requires special skill.

Some notes on Power

Typically, grant reviewers want to know how likely you are to see a fold-change of X for a given sample size at a defined probability level.
Since gene expression levels are so variable, we are usually talking about power to see a proportion of the variance explained.
In my view, it is somewhat pointless, since there is enormous variability in what fold changes are biologically meaningful:

N>4 generally OK, but larger sample sizes are always better, and provide better estimates of the relative variance components.



Nb: this is a generic plot, not representative of RNAseq

Yu L, Fernandez S, Brock G. (2017) *BMC Bioinformatics* **18**: 234
"Power analysis for RNAseq differential expression studies"

Ching T, Huang S, Garmire LX. (2014) *RNA* **20**: 1684-1696
"Power analysis and sample size estimation for RNAseq differential expression"