HYPOTHESIS TESTING

Joseph Powell

SISG- 2018

CONTENTS

- Experimental design

EXPERIMENTAL DESIGN

Typical stages of an RNA-seq study

1. Determine the question/hypothesis you want to address
2. Collect the tissue samples
3. Sequence the RNA
4. Process and clean the data
5. Fit statistical models and perform analysis of the data
6. Interpret and communicate the results

EXPERIMENTAL DESIGN

Typical stages of an RNA-seq study

1. Determine the question/hypothesis you want to address
2. Collect the tissue samples
3. Sequence the RNA
4. Process and clean the data
5. Fit statistical models and perform analysis of the data
6. Interpret and communicate the results

EXPERIMENTAL DESIGN

- Begin with the question or hypothesis and work ‘backwards’

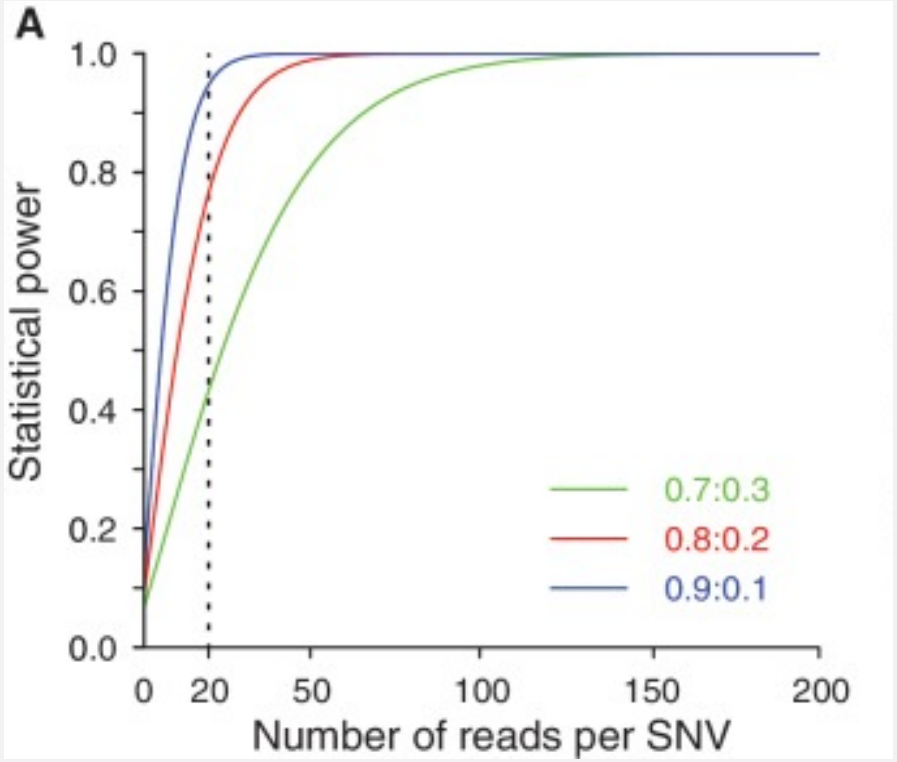
For example, suppose you want to address the question of whether or not gene expression levels are significantly different between disease vs. healthy individuals

You could test that with a Different Expression (DE) analysis.

- The final step of a DE analysis is the application of a statistical model to each gene in your dataset.
- Traditional statistical considerations and basic principals of statistical design of experiments apply.
- **Control** for effects of outside variables, avoid/consider possible biases, avoid confounding variables in sample preparation.
- **Randomization** of samples across batches
- **Replication** is important
- You should know your final (DE) model and comparison contrasts before beginning your experiment.

POWER

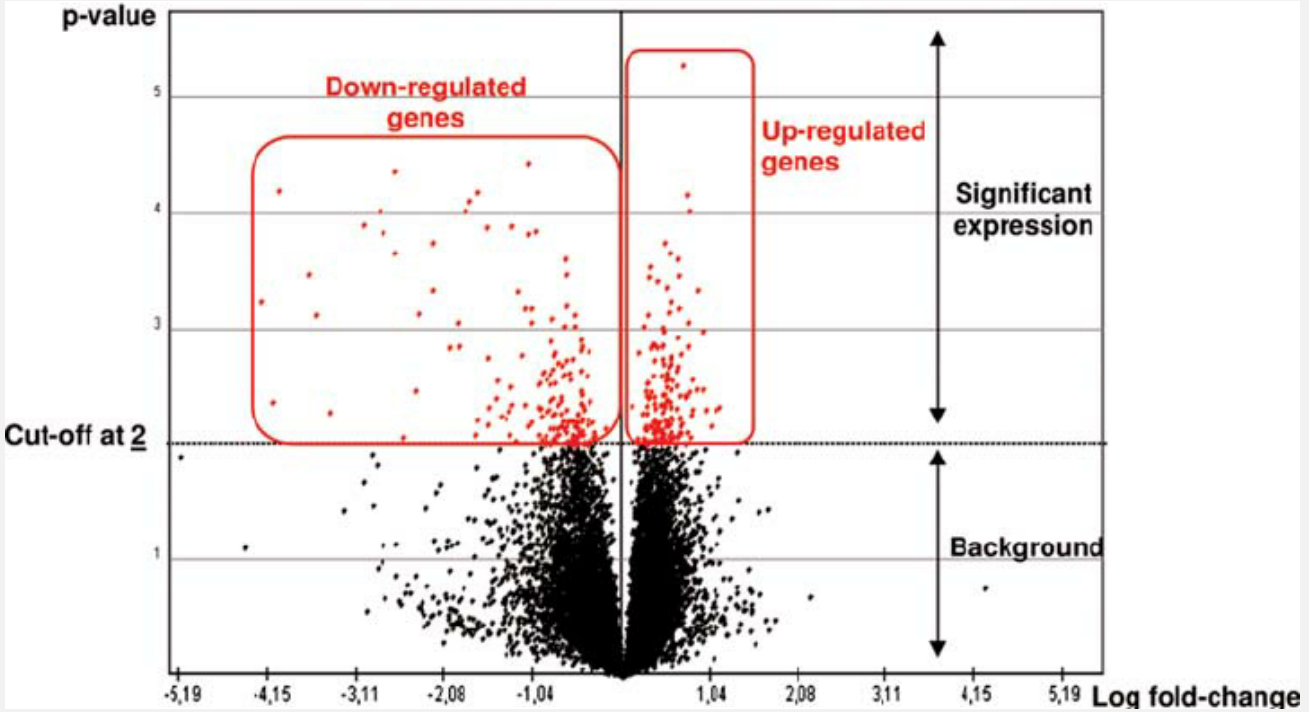
- Classical power calculation that deals with a single hypothesis takes a few simple assumptions. These include:
  - The effect size, representing the minimum difference that is scientifically meaningful between groups in comparison;
  - Within-group variation, representing natural variation in observations regardless of between-group difference;
  - An acceptable type I error rate, usually in the form of p-value; and the sample size.



Finding genes that are differentially expressed between conditions is an integral part of understanding the molecular ba
Joseph Powell
Joseph Powell

DIFFERENTIAL EXPRESSION

Finding genes that are differentially expressed between conditions is an integral part of understanding the molecular basis of phenotypic variation



DIFFERENTIAL EXPRESSION

A gene is declared differentially expressed if an observed difference or change in read counts between two experimental conditions is statistically significant

- Stats for microarrays are based on numerical intensity values
- Stats for RNA-Seq instead analyze read-count distributions

RNA-seq offers several advantages over microarrays, such as an increased dynamic range and a lower background level, and the ability to detect and quantify the expression of previously unknown transcripts and isoforms

MICROARRAYS

And that is all you are going to get!

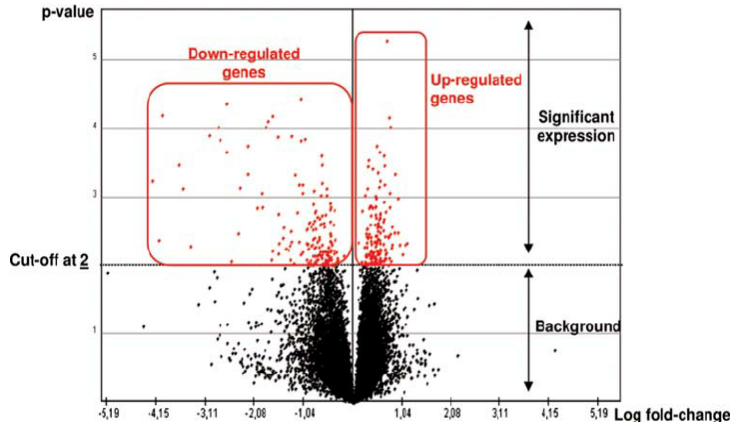
MICROARRAYS

And that is all you are going to get!

Microarrays have been used routinely for differential expression analysis for over a decade, and there are well-established methods available for this purpose (such as limma). These methods are not immediately transferable to analysis of RNA-seq data.
Ritchie *et al.* Nucleic Acids Research, 2015

CONTROLLING TYPE I ERROR RATE

Which genes are significantly differentially expressed?



What is a p-value?

What is the literal meaning of $p < 0.05$?

CONTROLLING TYPE I ERROR RATE

What is a p-value?

The p-value is the probability of obtaining a test statistic at least as extreme as the one that was observed, assuming that the null hypothesis is true.

CONTROLLING TYPE I ERROR RATE

What is a p-value?

The p-value is the probability of obtaining a test statistic at least as extreme as the one that was observed, assuming that the null hypothesis is true.

What is the literal meaning of $p < 0.05$?

CONTROLLING TYPE I ERROR RATE

What is a p-value?

The p-value is the probability of obtaining a test statistic at least as extreme as the one that was observed, assuming that the null hypothesis is true.

What is the literal meaning of $p < 0.05$?

This means that if we performed 100 random tests where we knew the null hypothesis was true, we'd see a test statistic at least this extreme five times.

WHAT IF WE PERFORMED 50,000 TESTS?

- If we set our threshold at $p < 0.05$ and we perform 50,000 tests, we would expect to get 2,500 'significant' results
- To be sure that there is only a 5% chance of a false positive we must adjust our threshold
- Bonferroni correction for multiple testing: set the threshold to:
  - $p < 0.05/50000$
  - $p < 1 \times 10^{-6}$

CONTROLLING FOR FDR

	number declared non-significant	number declared significant	total
true null hypotheses	U	V	m_0
false null hypotheses	T	S	$m - m_0$
	m - R	R	m

$$FDR = E[V/R]$$

(Benjamini and Hochberg, 1995)

THANK YOU

- Email me: j.powell@garvan.org.au
- Twitter: @JP_Garvan