Overview

Differential Expression

Kasper D. Hansen

Fall 2023

Differential expression

Setting: We have measured the expression of a large number of genes for a few samples.

Goal: Identify genes which change between groups of samples

This is one of the most common types of statistical analysis in genomics.

The ideas behind differential expression translates to other types of genomics: proteomics, ChIP, ATAC, etc.

Components of a successful analysis

(sketch)

1. Experimental design.
2. Summarize data: align, create gene by samples matrix
3. Quality control
4. Scale / normalize data
5. Assess data for unwanted variation / technical confounders / batch effects
6. Differential expression analysis
7. Interpretation
8. (Sometimes) Additional analyses like GO or gene set enrichment (GSEA)

Here, our focus is on steps 6 and 7 (and a bit of 4).

Tools for the trade

For bulk and single-cell RNA sequencing, there are 3 popular and well-performing tools in R/Bioconductor.

1. edgeR (RNA-seq)
2. DESeq2 (RNA-seq)
3. limma-voom (RNA-seq)

In python we have an implementation of DESeq2 called PyDESeq2.

The 3 methods perform reasonably similar. Pick 1 based on your personal preferences and stick with it.

(For single-cell sequencing, there are some reasons to prefer 1 and 2 for some experimental designs).

(Much time has been wasted on comparing these 3 tools on the same data. Don't waste time going down this road in your own work.)

Statistical keywords

Key statistical concepts for a high-quality analysis

1. Test statistic
2. Filtering
3. Multiple testing
4. Variance shrinkage / empirical Bayes
5. Model formula / design matrices / linear models
6. Unwanted variation / batch effects
7. Mean-variance relationship (specific to count data)

We will start by focusing on 1-4 in the context of a t-test.

Two group comparison

The simplest experimental design is a two group comparison, and this is what we will focus on here.

We have two groups of samples (sometimes referred to as cases and controls or treatment or constrols); could be mutant vs wild-type, vehicle vs drug etc.

Outline

- Testing 1 gene with a t-test
- Testing many genes: multiple testing, filtering, variance shrinkage
- More complicated experimental designs: linear models
- Mean-variance relationship and models for RNA-seq

Numbers

In humans, we have a bit more than $20,000$ protein-coding genes. In humans, each gene has multiple transcripts (isoforms) and we will ignore this and work with a hypothetical "gene' '.

We will (for now) focus on a two group comparison, where each group has a small number of samples. I tend to think of these categories

1. 3-5 samples per group (small, but standard)
2. 10's of samples per group (medium)
3. 100's or 1,000's of samples per group (large, unusual)

So our data matrix is something like

$$20,000 \times 6 - 10$$

Having this small of a sample size imposes limitations on the statistics / modeling we can do.

Our goal is to do well, while recognizing the limitations of the experiment.

Not all genes are expressed; filtering

Determining which genes are expressed in a given sample is a hard question. But usually we find that ~50% of genes are unexpressed (say 40%-60%), leaving us with 8,000-12,000 genes for analysis.

This is a kind of **filtering**; we will return to this later.