Overview of Differential Expression

Kasper D. Hansen

Fall 2022

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.

Tools

For bulk and single-cell RNA sequencing, there are 3 popular and well-performing tools

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

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

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

(Much time has been wasted on comparing these 3 tools on the same data. Don't do it.)

Keywords

Key statistical concepts for a high-quality analysis

1. Test statistics (traditional statistics)
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.

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

Outline

- Testing 1 gene: t-filtering
- Testing many genes: multiple testing, filtering, variance shrinkage
- More complicated experimental designs: linear models
- Mean-variance relationship

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.

Filtering

Usually the first step is some kind of filtering; common goals are the removal of

- genes which are unexpressed
- genes which have low variance (ie. have constant expression across the samples)

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

(We will return to filtering.)