github.com/rodriguezmDNA/bioinfo_primers/

Differentially Expressed

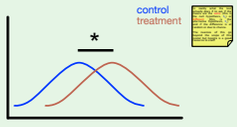
or, How I stopped worrying and learned DE.

The best way to describe the goal for this infographic is to state what this is not about: This is not about how to analyze expression data. Much less on how to process it or what goes behind these steps (think quality control, or trimming in the case of RNA-Seq), or normalization. This is not about copying and pasting code to get a result.

My goal is to provide (to the best of my understanding) a simplified way of thinking about what DE is and how these results came to be.

When I started dipping my feet in the waters of bioinformatics, one of my first tasks was to analyze expression data. I had just been learning how to use R and the swift change was kind of a splash of cold water. "*Replicates, samples, t-test, regression, normalization, differential expression analysis*"…before going deeper into the code, I spend a good amount of time trying to make sense of it all.

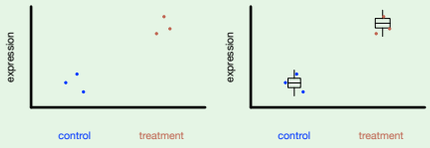
One of the common explanations I kept finding was along the lines of: "you test if a treatment has an effect on the expression of a gene and you do this by comparing the means" often accompanied by a figure like this:



The following is (in my opinion) a better way to visualize expression data for a **single gene** across all the samples.

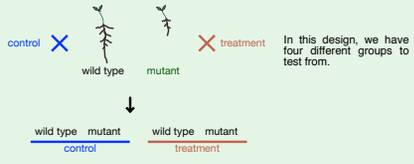
Each dot represents a sample and they're grouped by treatment.

A boxplot summarizes the data nicely.

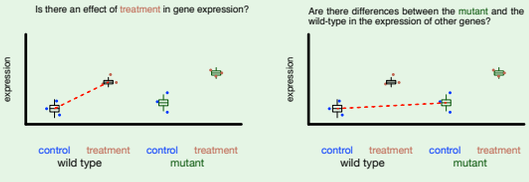


a motivating example

To root this example in real-life, we'll be looking at expression coming from a 2014 study of Arabidopsis root's response to nitrogen. This study compares the combined effects of genotype (wild type or a mutant) and nitrate addition (control or treatment) in gene expression.

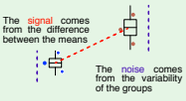


In this design, we have four different groups to test from.

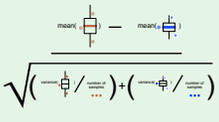


What are we testing for?

When we ask for "is there an effect of treatment or genotype in gene expression", a common approach is to use a t-test. This compares the signal against the noise in the data to assess if a difference exists.



This ratio results in a **t-value**. It measures the difference between two samples relative to the variation in the data.



In its simplest case, this is what differential expression tests for. The logFC is just the difference between the means.

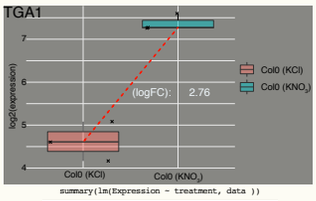
In real life

In the study, the authors looked at the effect of both a nitrate treatment and mutant genotype on gene expression. We can ask what is the effect of each separately or if the combination has an effect (an interaction).

First, let's look at how nitrate addition influences expression of a gene:

Expression	Genotype	Treatment
5.082	Col0	KCl
4.611	Col0	KCl
4.166	Col0	KCl
7.274	Col0	KNO_3
7.605	Col0	KNO_3
7.255	Col0	KNO_3

$$\frac{\text{signal}}{\text{variation}} = \frac{7.378 - 4.619}{\sqrt{(0.03878/3) + (0.2098/3)}} = 9.58 \quad \text{t-value}$$




```
summary(lm(Expression ~ treatment, data ))
```

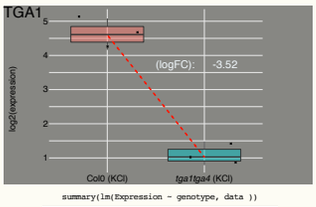
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)     4.619      0.204   22.70  2.2e-05 ***
treatmentKNO3   2.759      0.288    9.58  0.00066 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Following the t-test formula we get that the ratio between signal and noise is larger (indicative of the effect size of the treatment **alone**). The *t.test()* funciton in R provides more information about the logFC (estimate) and the p-value.

Next, we can see the effect of the genotype **only** (*tga1tga4*) on the expression of the same gene. Here we'd expect the expression to decrease relative to wild type in control conditions (without any nitrate treatment):

Expression	Genotype	Treatment
5.082	Col0	KCl
4.611	Col0	KCl
4.166	Col0	KCl
1.481	tga1tga4	KCl
1.027	tga1tga4	KCl
0.790	tga1tga4	KCl

$$\frac{\text{signal}}{\text{variation}} = \frac{1.099 - 4.619}{\sqrt{(0.1234/3) + (0.2098/3)}} = -10.56 \quad \text{t-value}$$

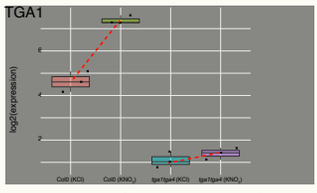



```
summary(lm(Expression ~ genotype, data ))
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)     4.619      0.236   19.6   4e-05 ***
genotypetga1tga4  -3.520      0.333  -10.6  0.00045 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

What is an interaction effect?

Is the effect of treatment **and** genotype (together) on gene expression different than the sum of either factors alone? When we talk about interactions we essentially ask if the difference between the effect of treatment and the effect of the genotype is not zero.




```
summary(aov(Expression ~ genotype * treatment, data ))
```

```
                    Df Sum Sq Mean Sq F value   Pr(>F)
genotype             1   67.7    67.7   614.1 7.5e-09 ***
treatment            1    7.0     7.0    63.7 4.4e-05 ***
genotype:treatment   1    4.5     4.5    41.1 0.00021 ***
Residuals            8    0.9     0.1
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Here I'm introducing a new way to look at these difference with the **an**alysis **o**f **va**riance (anova). With an anova , we look at the variance **between** groups (signal) and **within** groups (noise) - I'll likely update this poster to provide more insight into how it works.

These are a couple of ways to look at changes in expression at the genomic level. We can either test for single effects (ie, genotype or treatment) or for a combination of both (the interaction). When doing these tests on all of the genes in the genome it is important to consider the effects of repeated testing to avoid false positives. In the next version of this poster I will add a few ways to account for this using either Bonferroni correction or calculating the false discovery rate (FDR).

@rodriguezmDNA

an informational poster by Joel Rodriguez-Medina

ASPB/Conviron Scholar
Final independent project

v1.0 06.2019