Summer Institutes of Statistical Genetics, 2023

Module 6: GENE EXPRESSION PROFILING

Greg Gibson and Peng Qiu

Georgia Institute of Technology

Lecture 2: HYPOTHESIS TESTING and NORMALIZATION

greg.gibson@biology.gatech.edu

http://www.cig.gatech.edu

Why do we work on the $\log_2$ scale ?

1. Log transformation makes the data more normally distributed, minimizing biases due to the common feature that a small number of genes account for over half the transcripts

2. Log base 2 is convenient, because in practice most differential expression is in the range of 1.2x to 8x, depending on the contrast of interest and complexity of the sample.

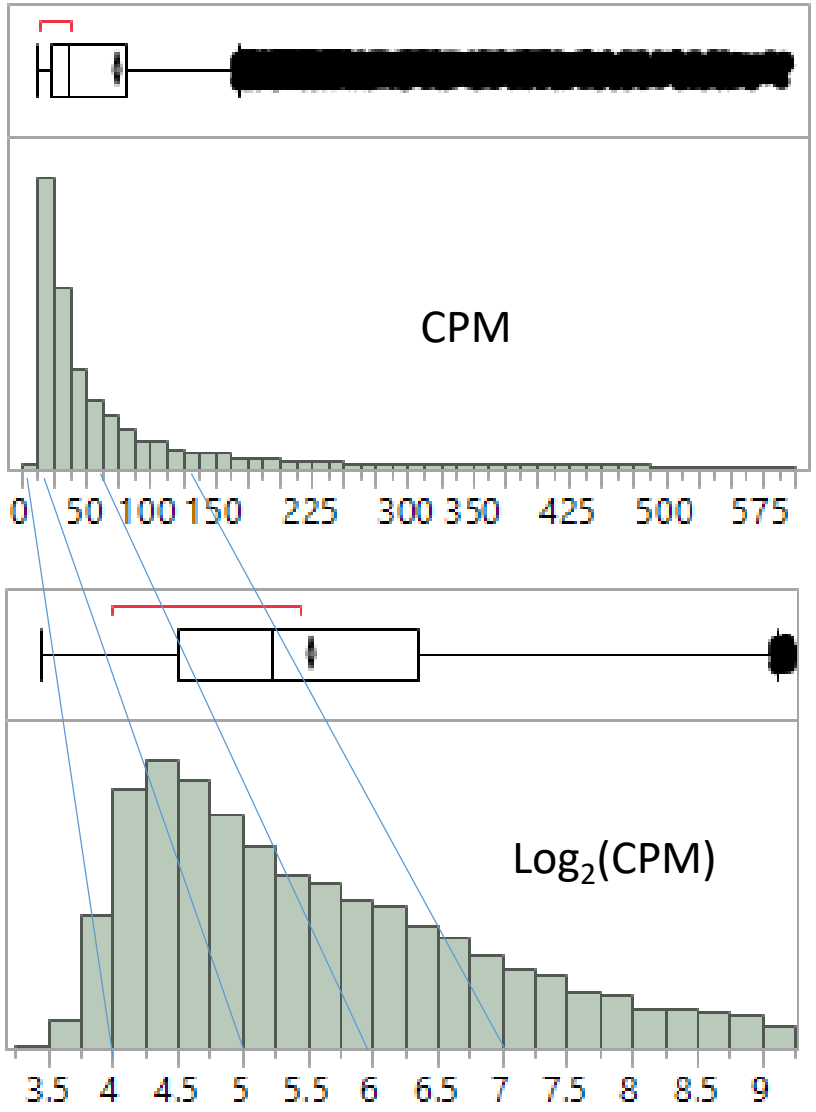
3. It is also intuitively simple to infer fold changes in a symmetrical manner:

   A difference of -1 unit corresponds to half the abundance, and +1 to twice the abundance
   A difference of -2 units corresponds to a quarter the abundance, and +3 to 8-times the abundance

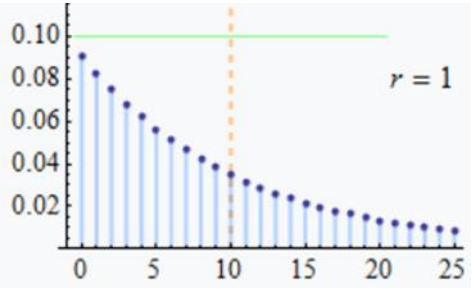
4. The log scale is insensitive to mean centering, so it is simple to just set the mean or median to 0, preserving the relative abundance above or below the sample average

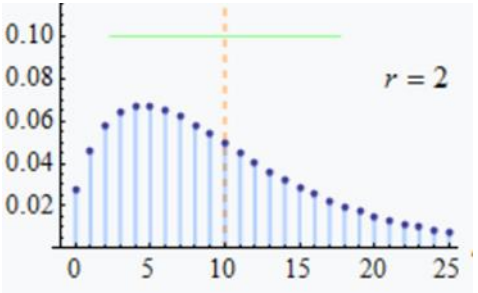
5. It is generally useful to add 1 to all values before taking the log, to avoid “0” returning #NUM! (but this step is built into most code)

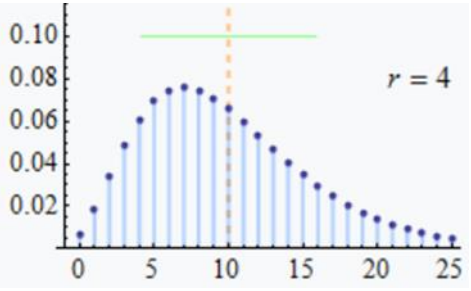
Data Distributions



Negative Binomial Distribution
(one low abundance transcript)



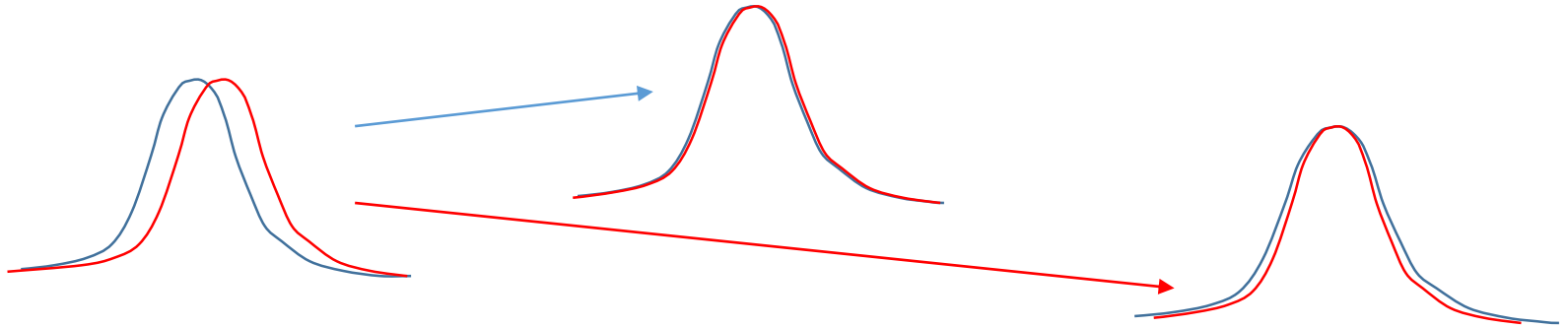




Sample-specific Normalization

In the Microarray days, we generally used additive adjustment to center the mean or median

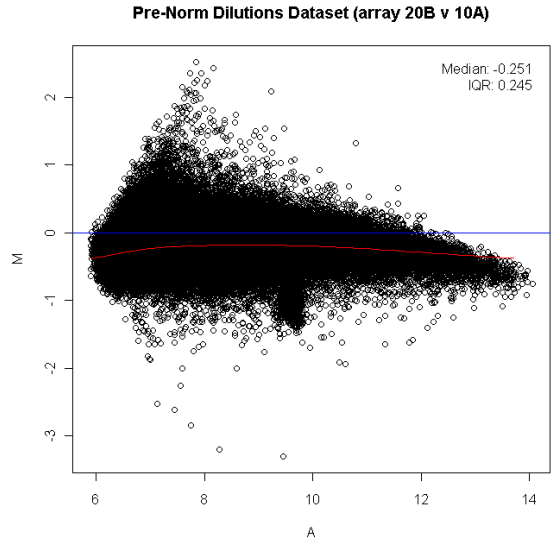
When RNAseq took over, the emphasis shifted to multiplicative scaling to total counts

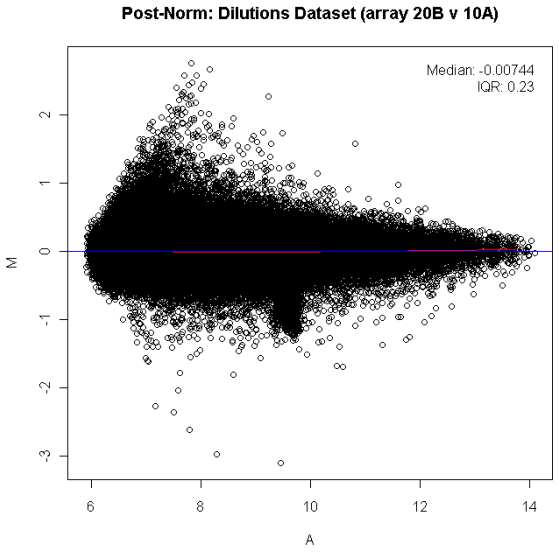


Additional adjustments like TMM account for biases due to variable abundance of a small number of highly expressed transcripts, like HBB or Ribosomal or Mitochondrial components. If they account for 50% of the transcripts in one sample but 30% in another, then the CPM will all be higher on the second sample.

Also for RNAseq data, adjustment is made for the high “zero-count” (drop-out) rate for low-abundance transcripts: the data is said to be negative binomially distributed.

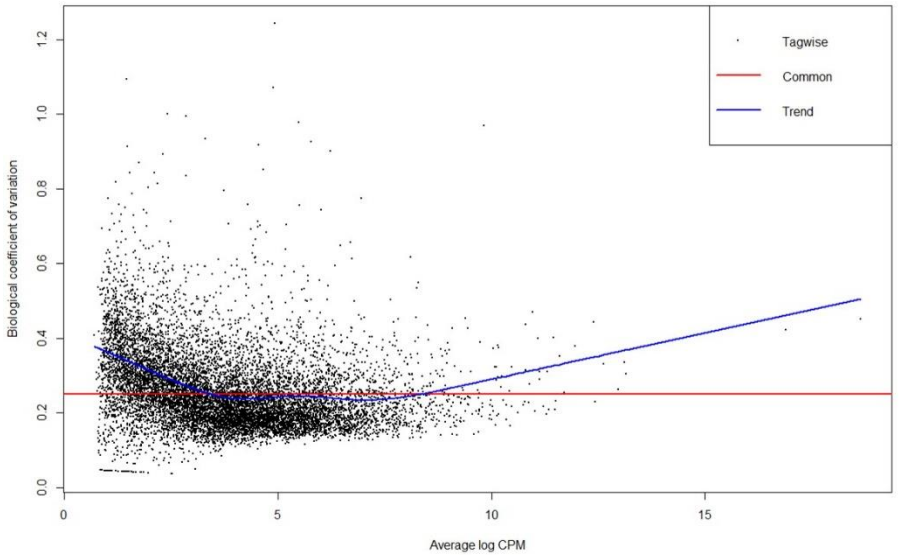
MA Plots: Magnitude vs Abundance; and Dispersion





$$M = \log_2(R/G) = \log_2(R) - \log_2(G)$$

$$A = \frac{1}{2}\log_2(RG) = \frac{1}{2}(\log_2(R) + \log_2(G))$$



Types of Hypothesis Test

T-test **Pairwise Single Contrasts**

Assumes the data is normally distributed
Evaluates whether the sample means differ
Paired sample options also available

$$t = \frac{Z}{s} = \frac{\bar{X} - \mu}{\widehat{\sigma}/\sqrt{n}}$$

ANOVA **Analysis of Variance**

Evaluates within and between sample variance with an F-test
Handles complex designs with 2 or more fixed effects
Interaction effects evaluate conditional significance
One-way and Nested designs are special cases
Contra Deseq/EdgeR no reference required
Mixed models include random effects

$$F = \frac{\text{variance between treatments}}{\text{variance within treatments}}$$

$$F = \frac{MS_{\text{Treatments}}}{MS_{\text{Error}}} = \frac{SS_{\text{Treatments}}/(I-1)}{SS_{\text{Error}}/(n_T - I)}$$

where MS is mean square, I = number of treatments and n_T = total number of cases

Types of Hypothesis Test

Likelihood Ratio-test **Model Comparisons**

Evaluates whether inclusion of a parameter explains the data better than not having it in the model
-2ln(ratio) converges on χ^2 distribution with large n
Gives you control over what contrasts to perform:
Simplest model is the null
Full models can include multiple parameters

$$\lambda_{LR} = -2 \ln \left\{ \frac{\text{(Likelihood of data with Parameter)}}{\text{(Likelihood of data with Simpler Model)}} \right\}$$

Rank-Sum tests **Non-parametric**

Does not assume anything about the distributions
Tests for symmetry of ranks around the mean, but without caring about the actual deviations
Wilcoxon Signed-Rank for paired samples
Mann-Whitney U = Wilcoxon Rank-Sum for independent samples

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.

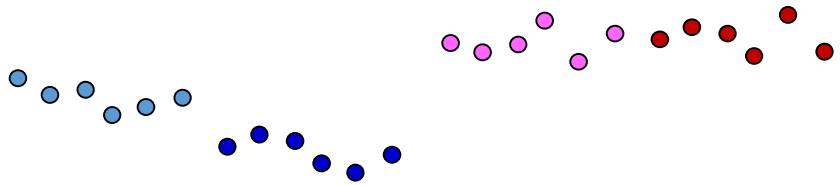
2. edgeR shrinks the variance of low abundance transcripts by fitting the distribution to the "negative binomial" expectation. Basically it adds values to account for sampling error at the low end so that comparing 0, 1 and 2 is more like comparing 10, 11 and 12.

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.

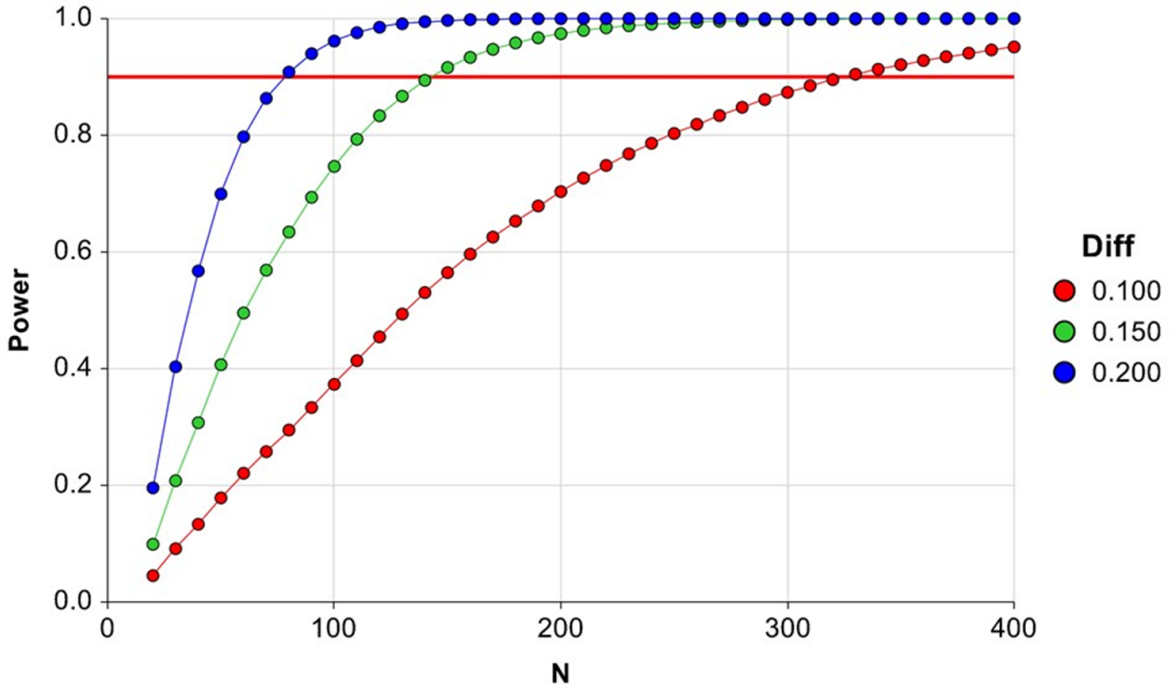
I can ask, are pink and red different from light and dark blue (is there a COVID effect)
Or, are dark blue and red different from light blue and pink (is there a B vs T cell effect)
Or, are dark red different from all others (are COVID B-cells affected), which gives a different answer than are red different 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

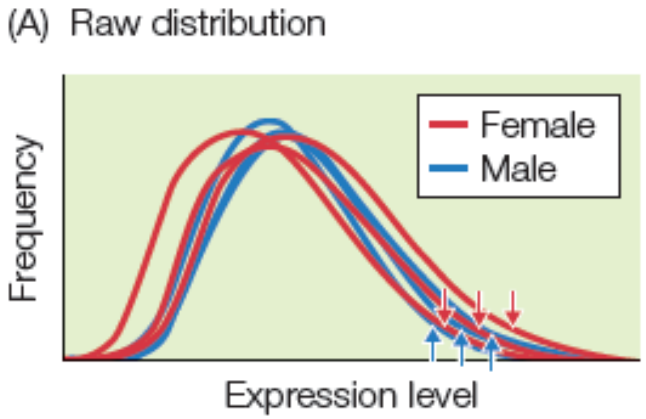
Yu L, Fernandez S, Brock G. (2017) *BMC Bioinformatics* **18**: 234
"Power analysis for RNAseq differential expression studies"

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

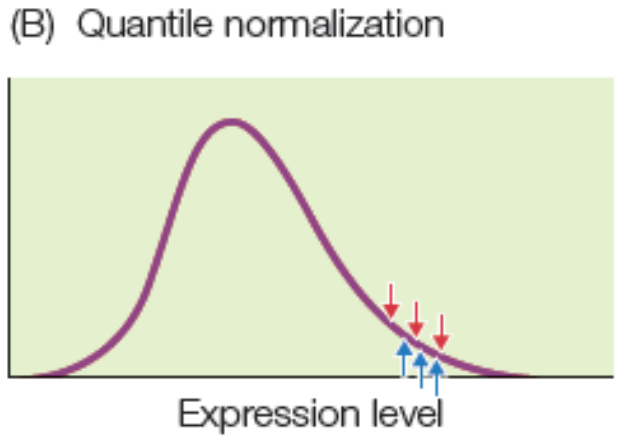
Approaches to Normalization

- Mean or Median transform, simply centers the distribution
  - Something like this is essential to control for overall distributional effects (eg RNA concentration)
- Variance transforms, such as standardization or inter-quartile range
  - Depends on whether you think the overall distributions should have similar variance
- Quantile normalization
  - Transforms the ranks to the average expression value for each rank
- Gene-level model fitting
  - Remove technical or biological effects before model fitting on the residuals
- Supervised normalization
  - Optimally estimate the biological effect while fitting technical factors across the entire experiment

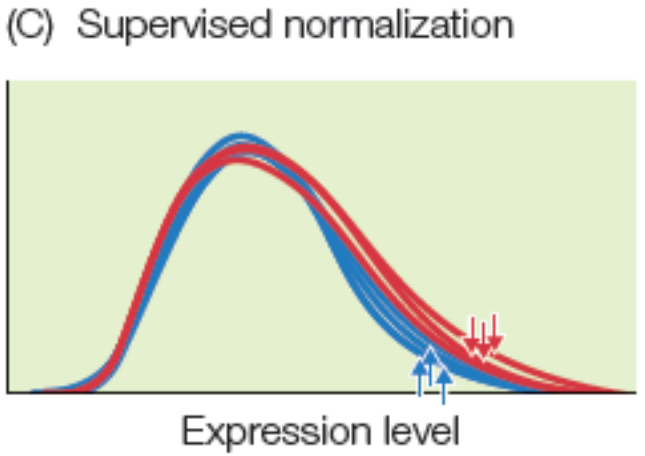
Relative and Absolute Normalization



Raw data:
no effect



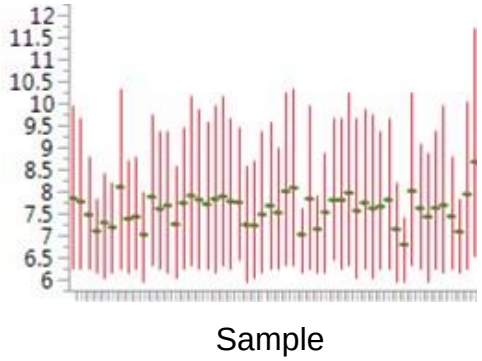
Variance transformed:
no effect

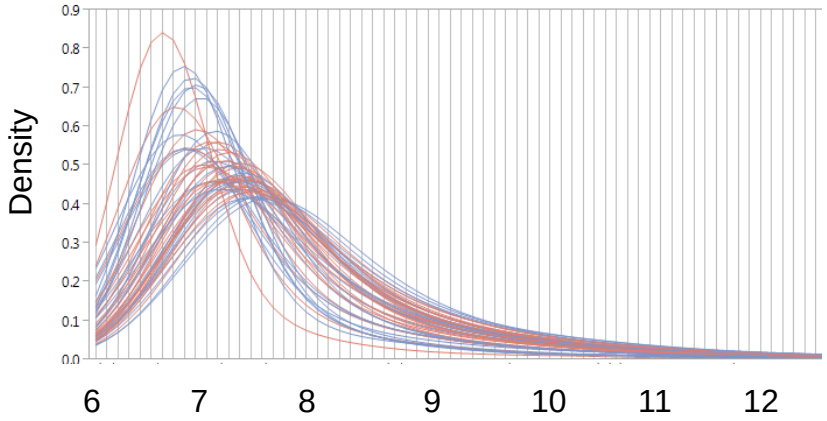


Mean centered:
significant effect

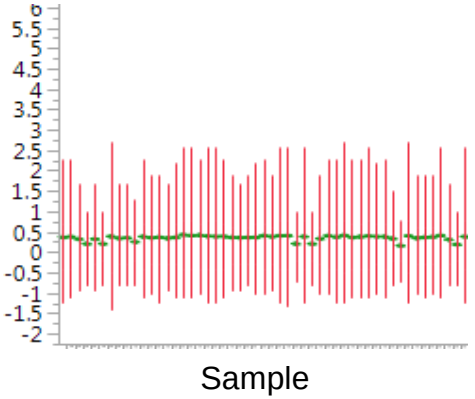
Effect of Median Centering

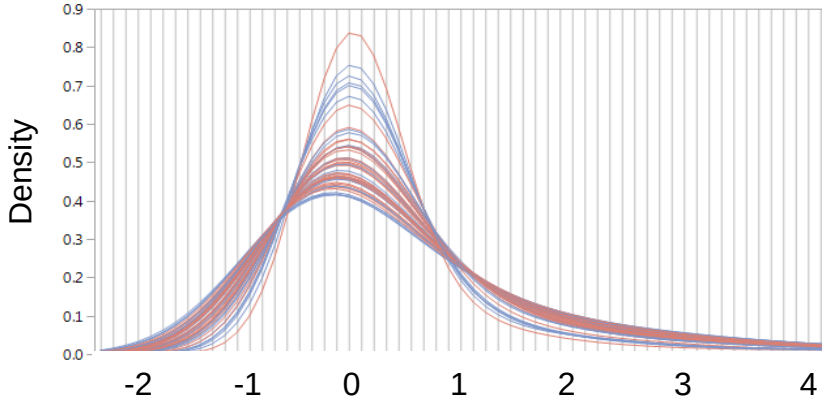
Raw Profiles





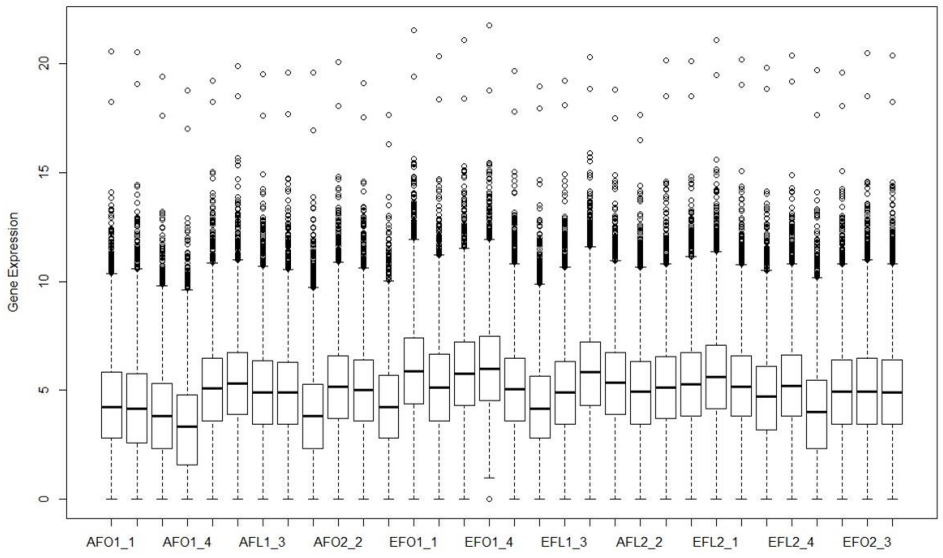
Median Transform

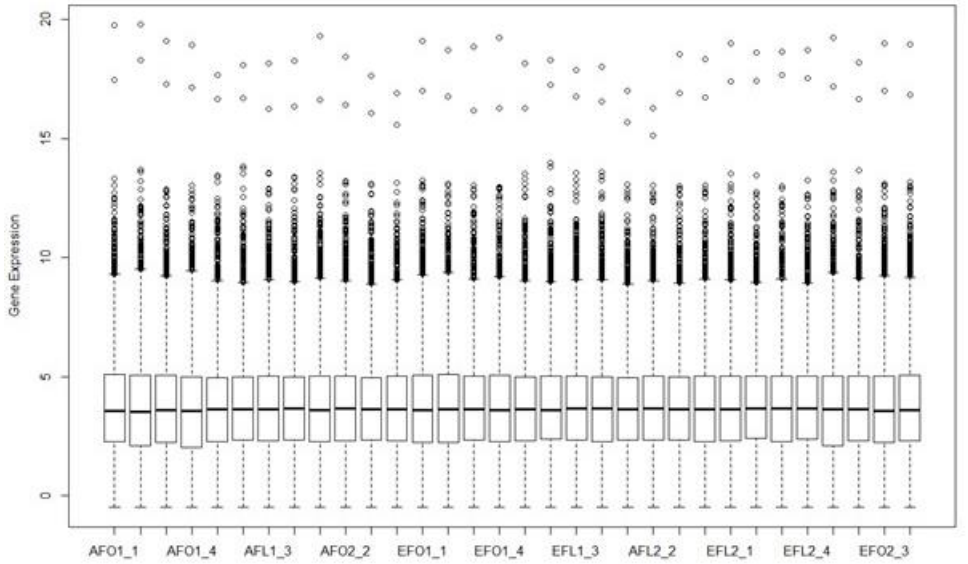




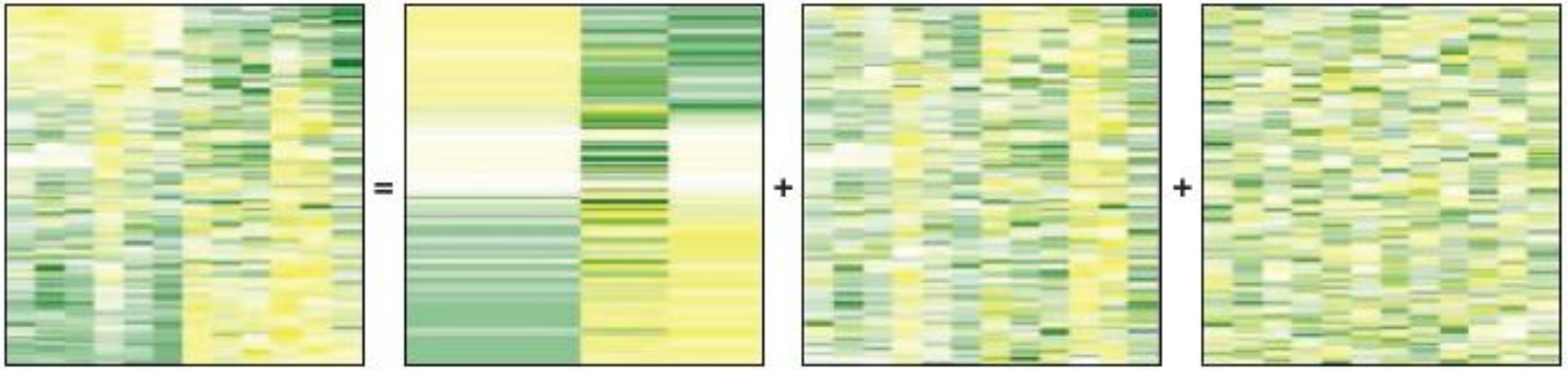
For RNASeq data, CPM essentially does this: cpm = 1,000,000 x reads/total reads

Effect of Variance Scaling

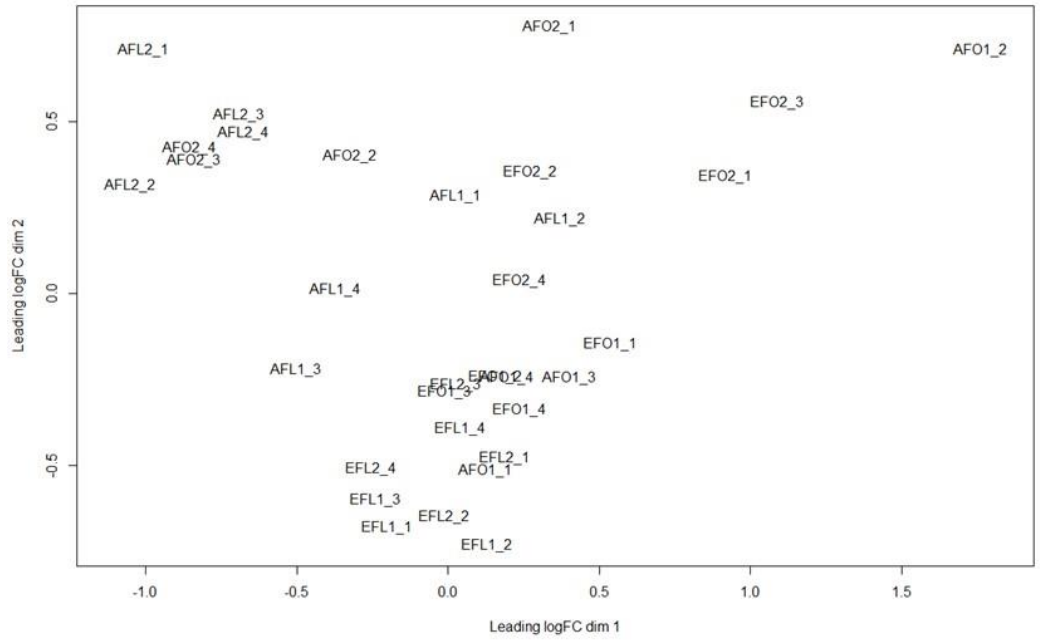


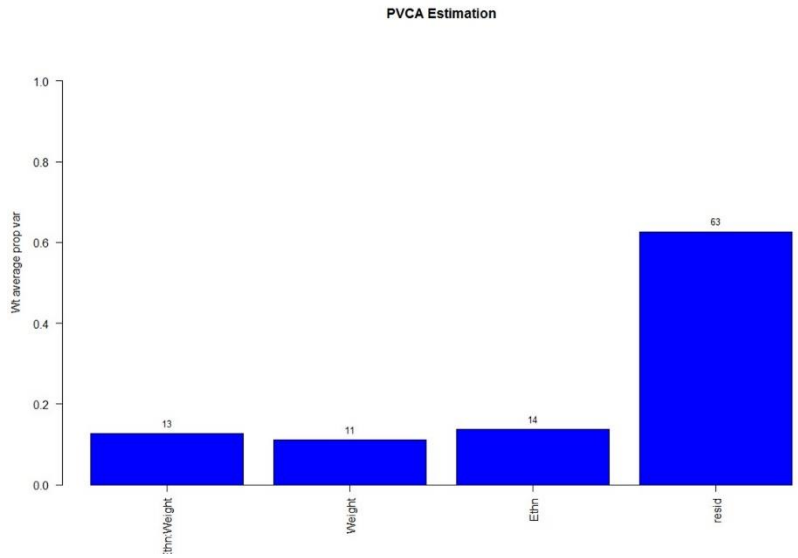


The Normalization Challenge



Principal Component Variance Analysis





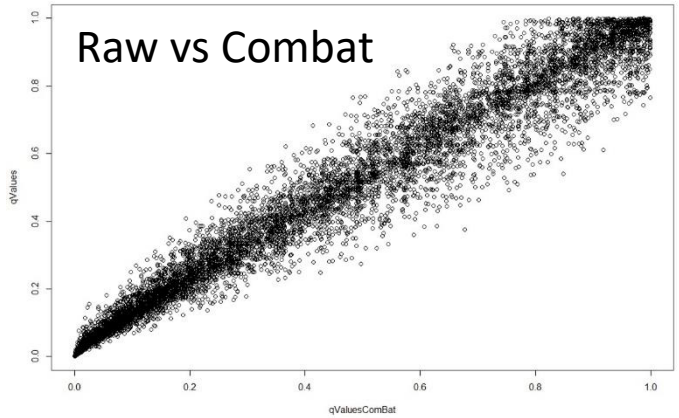
It is always a good idea to start by asking what biological and technical factors dominate the variation in your samples. Then you can choose which ones to adjust for in your modeling.

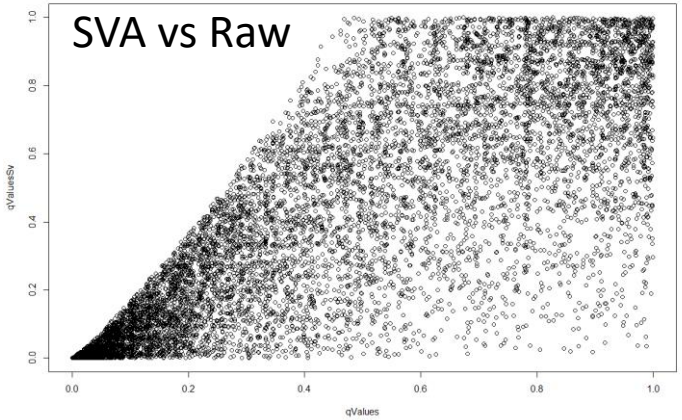
Surrogate Variable Analysis

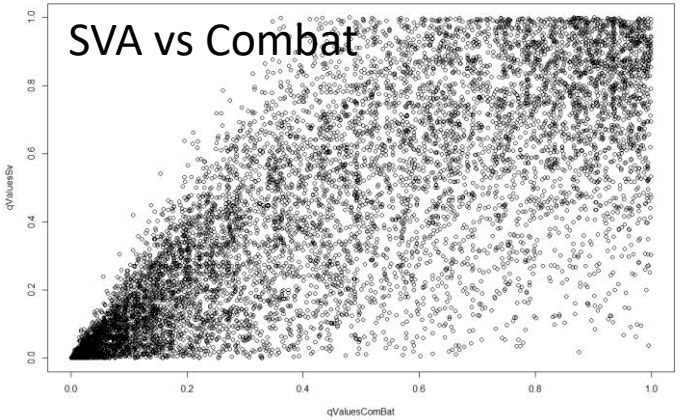
```
> summary(fit2)
            Df Sum Sq Mean Sq F value   Pr(>F)
ethn         1 0.0070 0.00699   0.708    0.409
weight       1 0.0015 0.00153   0.155    0.698
visit        3 0.0175 0.00585   0.593    0.626
person       5 0.7668 0.15335  15.545 1.92e-06 ***
Residuals   21 0.2072 0.00987
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> fit3 <- aov(SV3 ~ ethn + weight + visit + person, data=phEIGHT)
> summary(fit3)
            Df Sum Sq Mean Sq F value   Pr(>F)
ethn         1 0.5516  0.5516  74.451  2.4e-08 ***
weight       1 0.0005  0.0005   0.061 0.806947
visit        3 0.0188  0.0063   0.846 0.483938
person       5 0.2735  0.0547   7.383 0.000392 ***
Residuals   21 0.1556  0.0074
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

COMBAT is a batch correction method: you remove the effects of technical confounders
PEER factor analysis is a Bayesian approach that by default automatically adjusts for latent variables
SVA (Surrogate Variable Analysis) gives you control over which variables to adjust for
SNM (Supervised Normalization of Microarrays) iteratively adjusts for biological and technical factors

Normalization matters







Recommended Approach

1. Normalize the samples, paying attention to the distributions of overall profiles

2. Extract the Principal components of gene expression, and ask whether the major PC are correlated with technical covariates such as Batch or RNA quality; or with Biological variables of interest

3. If they are, renormalize to remove those effects

4. As much as possible, analyze the dataset in several different ways to

(i) confirm that the findings are not sensitive to your analytical choice, and

(ii) gain insight into what may cause differences, eg find confounding factors

5. Compare the final p-value distributions, and perform gene ontology analysis to evaluate which strategy is giving you biologically plausible insight.

Biological Interpretation

Variance transforms tend to focus attention on relative changes in abundance, but it is not clear biologically whether it matters to a cell whether it is the ratio of gene A to gene B, or absolute levels that matter. Enzymes and kinases and cytoskeleton and growth factors may differ.

There is no reason to assume that all samples have the same overall distribution of transcript abundance. All approaches accept that there is pervasive covariance. Supervised approaches assume that biologically similar samples are transcriptionally more similar. But this implies that different analysis pipelines will yield different results, which some analysts are uncomfortable with.

Different genes have different variance, related to technical, biological, and abundance effects. How we model these also affects the significance and hence DE (differential expression) estimates.

It is inevitable that choice of analytical strategy will affect the outcome. Survey of the ontology of the DE genes is one way to assess whether your strategy worked, but “it makes biological sense” could suffer from positivist bias (“I saw what I expected, so it is correct”).