

Summer Institutes of Statistical Genetics, 2022

Module 6: GENE EXPRESSION PROFILING

Greg Gibson and Peng Qiu

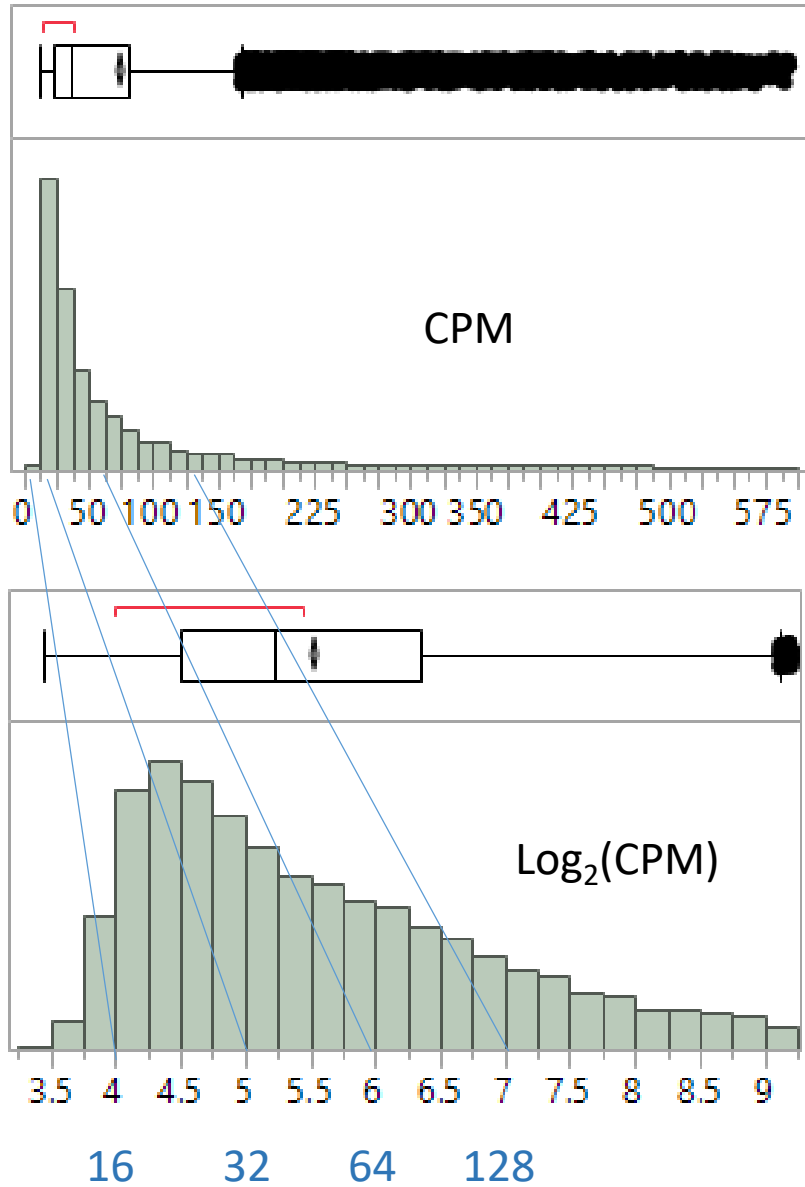
Georgia Institute of Technology

Lecture 4: NORMALIZATION

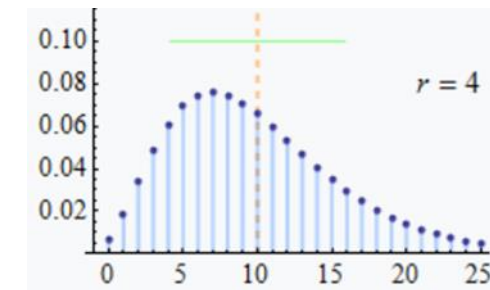
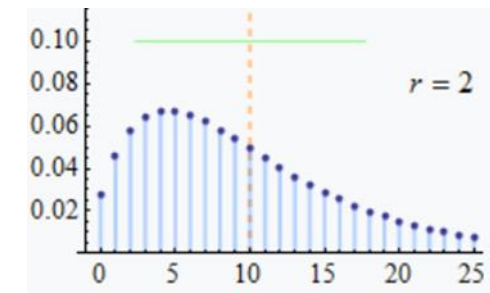
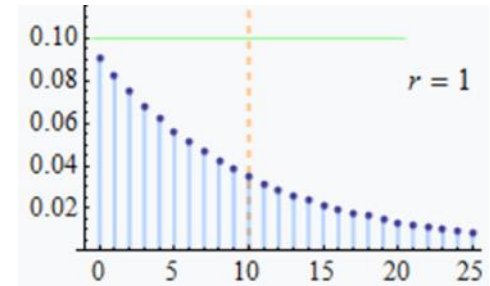
Why do we work on the \log_2 scale ?

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



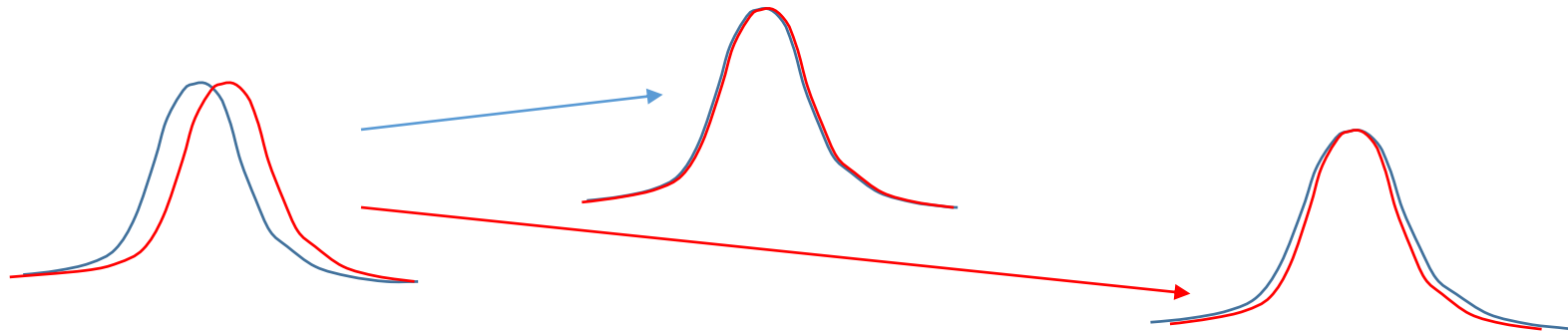
Negative Binomial Distribution
(one low abundance transcript)



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

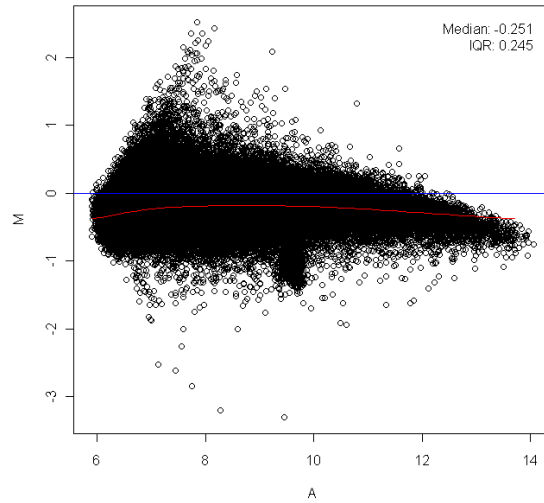


Additional adjustments like TMM account for biases due to variable abundance of a small number of highly expressed transcripts, like HBB or Ribosomal or Mitochondrial components. If they account for 50% of the transcripts in one sample but 30% in another, then the CPM will all be higher on the second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

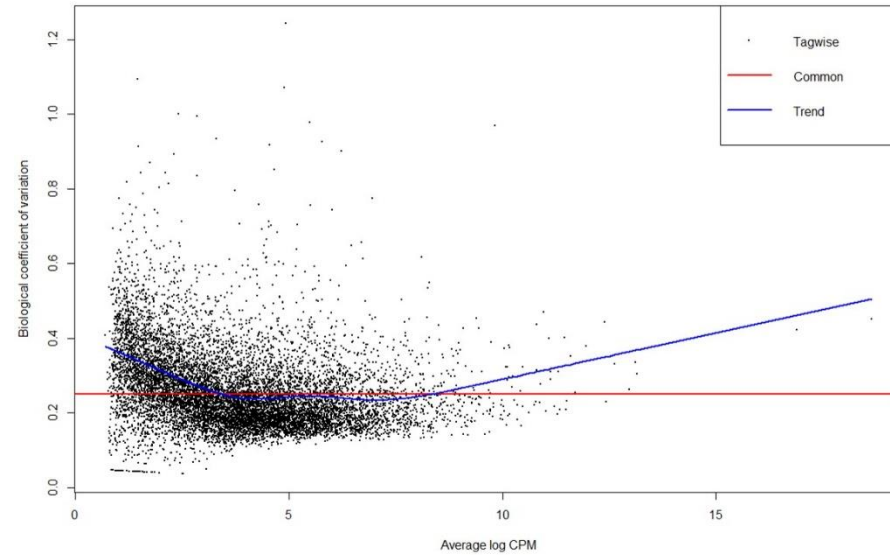
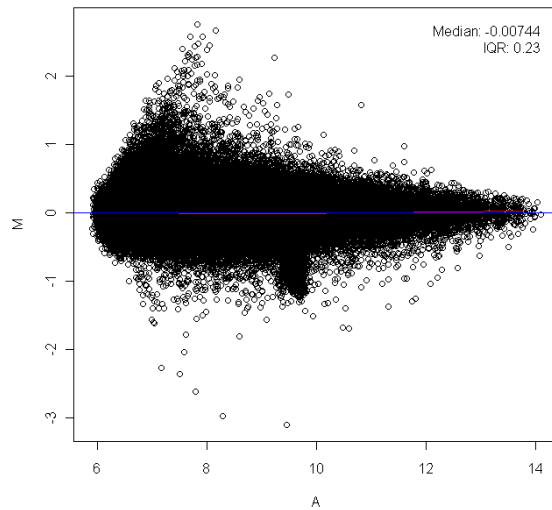
Pre-Norm Dilutions Dataset (array 20B v 10A)



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

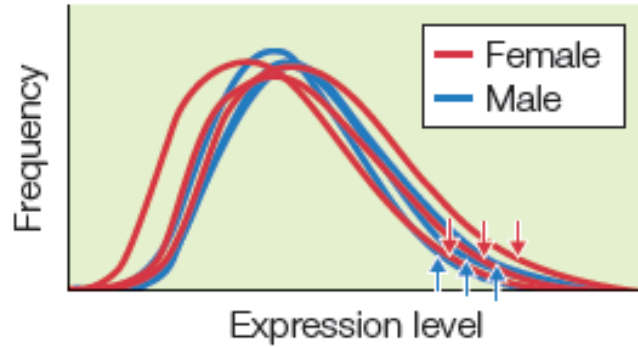
$$A = \frac{1}{2} \log_2(RG) = \frac{1}{2} (\log_2(R) + \log_2(G))$$

Post-Norm: Dilutions Dataset (array 20B v 10A)



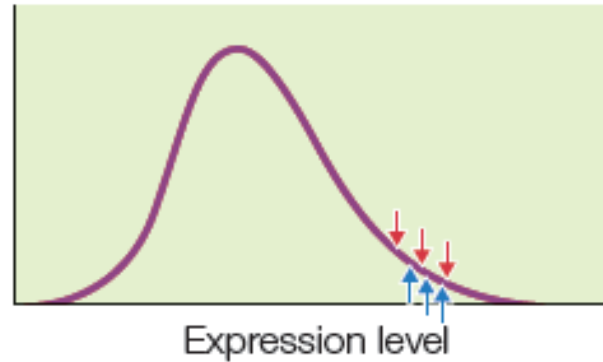
Relative and Absolute Normalization

(A) Raw distribution



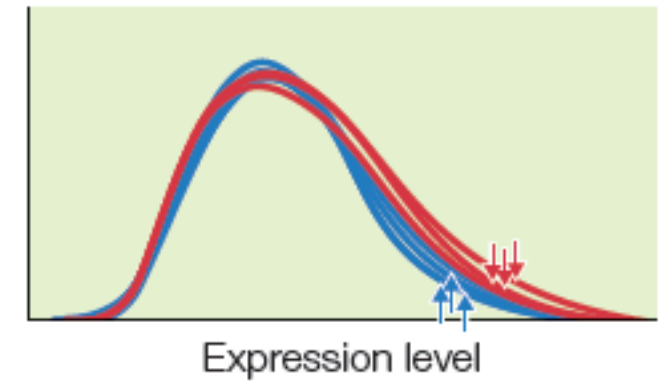
Raw data:
no effect

(B) Quantile normalization



Variance transformed:
no effect

(C) Supervised normalization



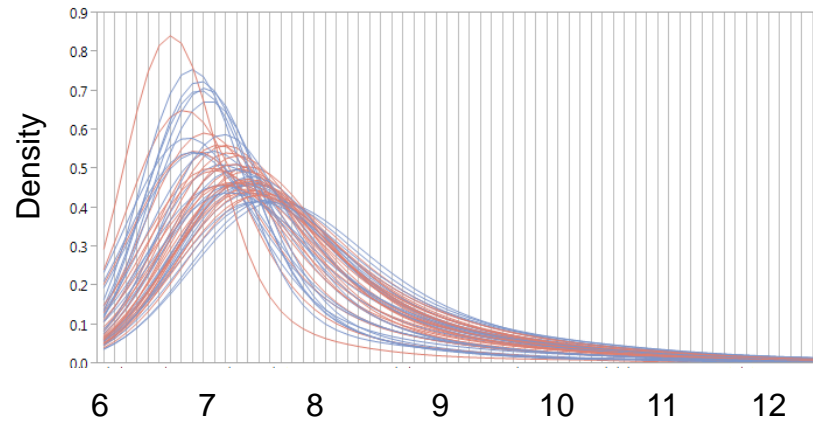
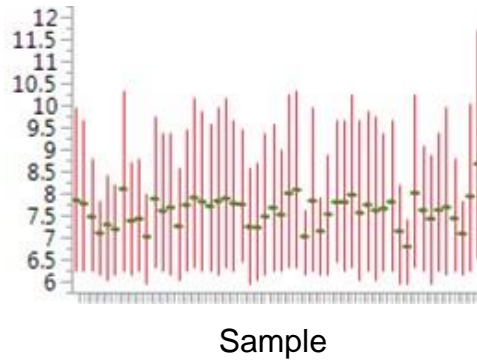
Mean centered:
significant effect

Approaches to Normalization

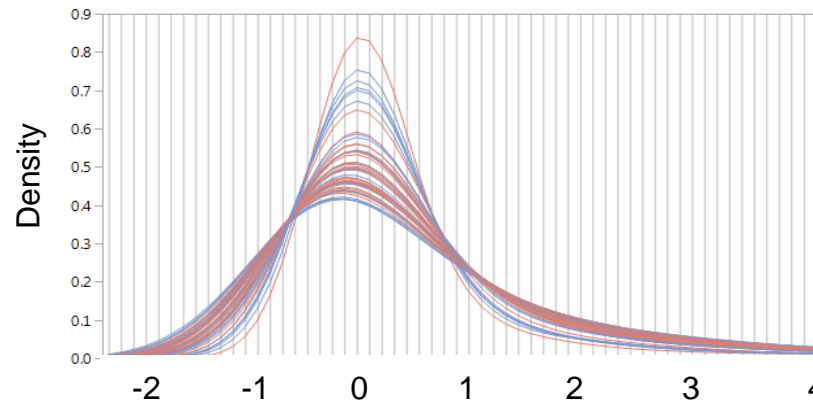
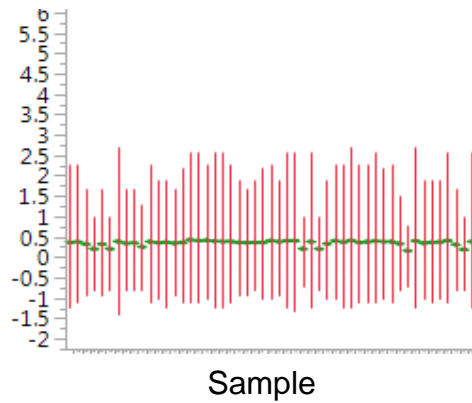
- Mean or Median transform, simply centers the distribution
 - Something like this is essential to control for overall distributional effects (eg RNA concentration)
- Variance transforms, such as standardization or inter-quartile range
 - Depends on whether you think the overall distributions should 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

Effect of Median Centering

Raw Profiles

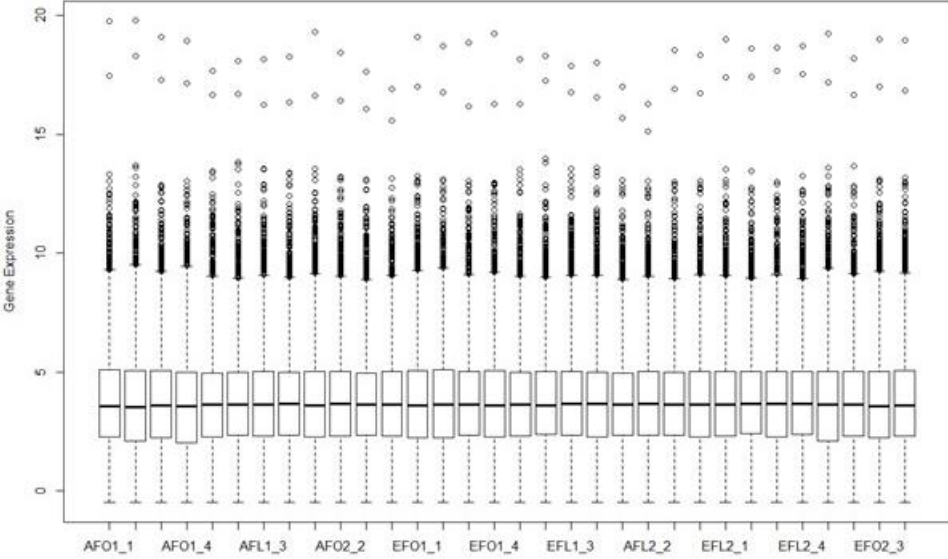
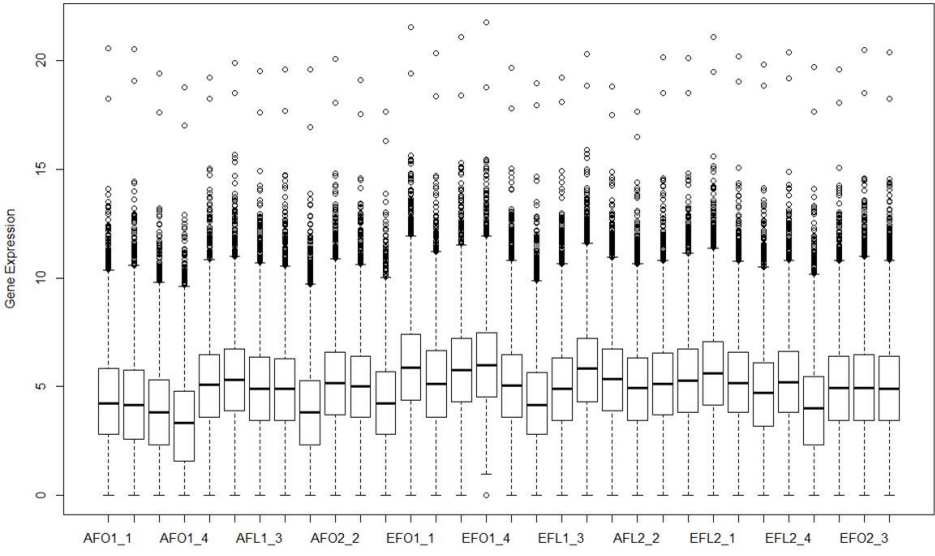


Median Transform

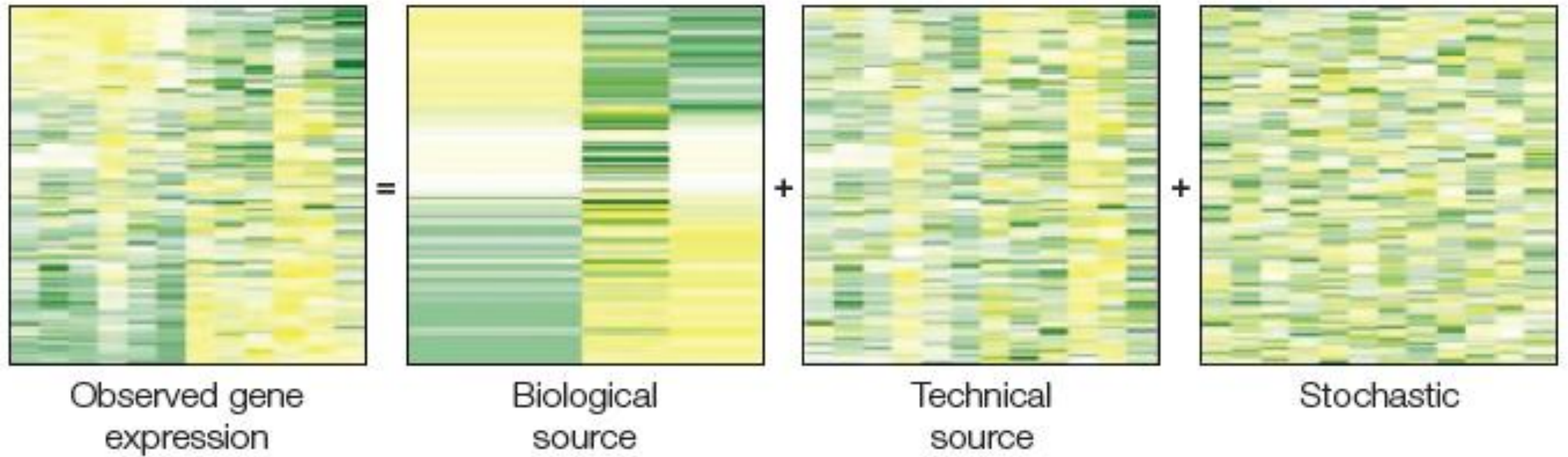


For RNASeq data, CPM essentially does this: $\text{cpm} = 1,000,000 \times \text{reads}/\text{total reads}$

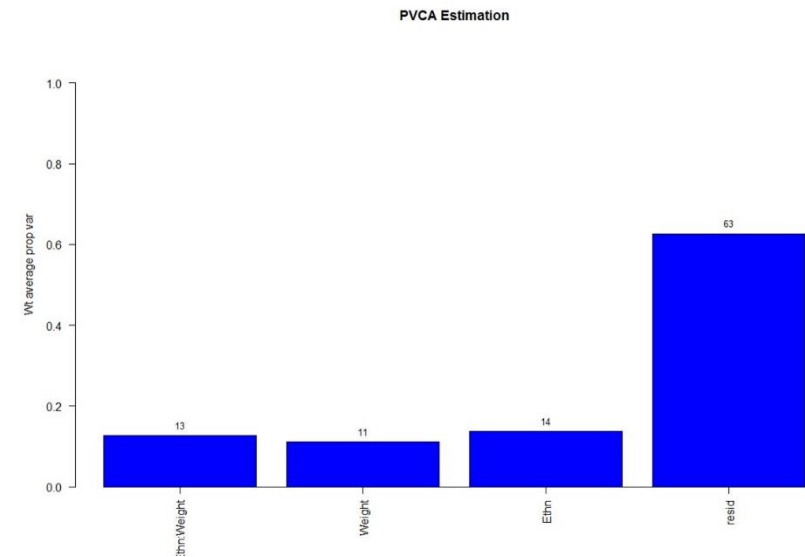
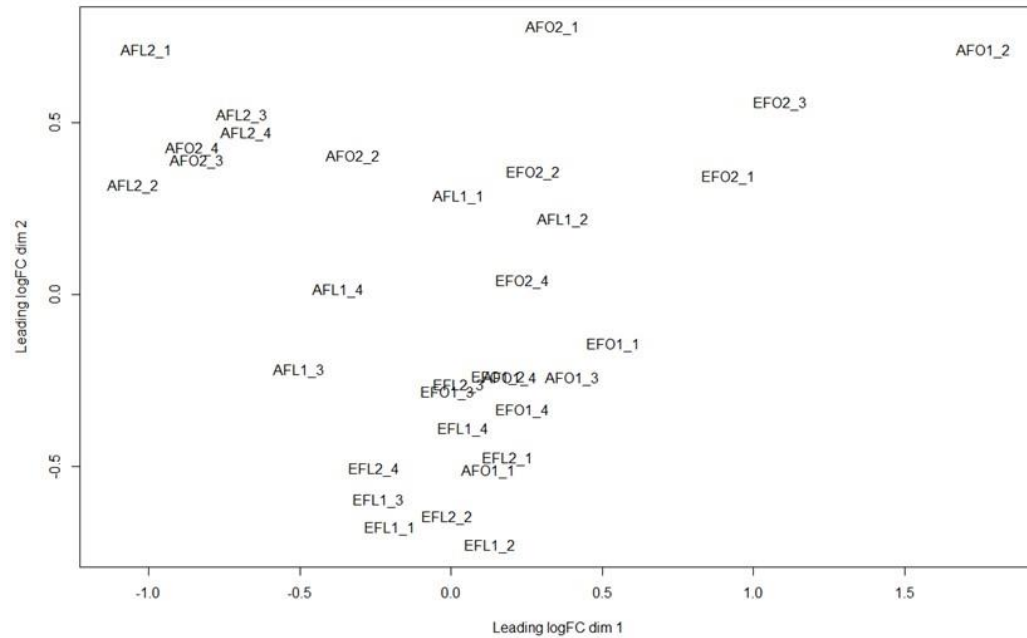
Effect of Variance Scaling



The Normalization Challenge



Principal Component Variance Analysis



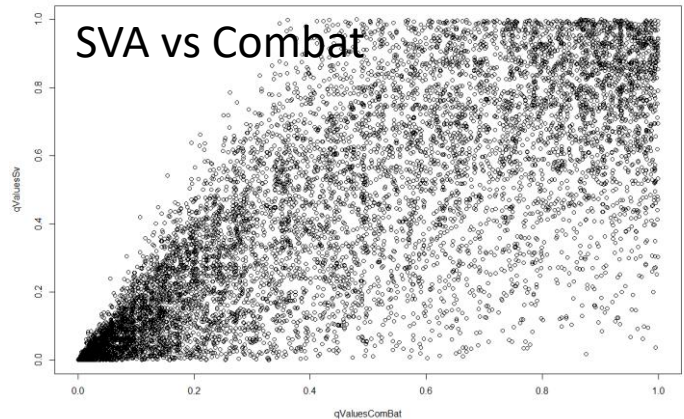
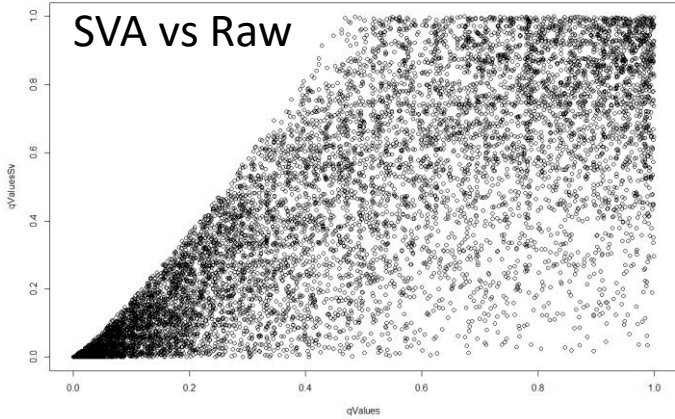
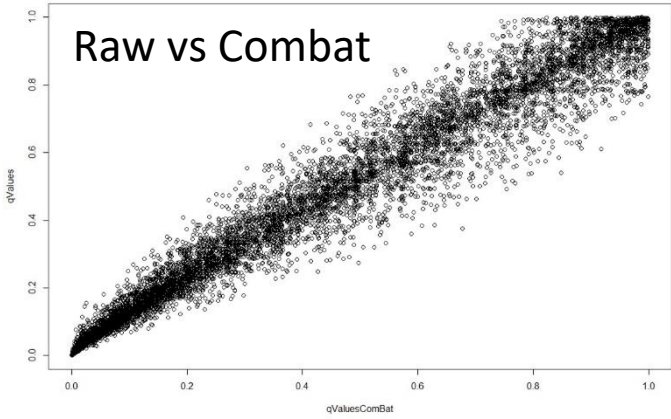
It is always a good idea to start by asking what biological and technical factors dominate the variation in your samples. Then you can choose which ones to adjust for in your modeling.

Surrogate Variable Analysis

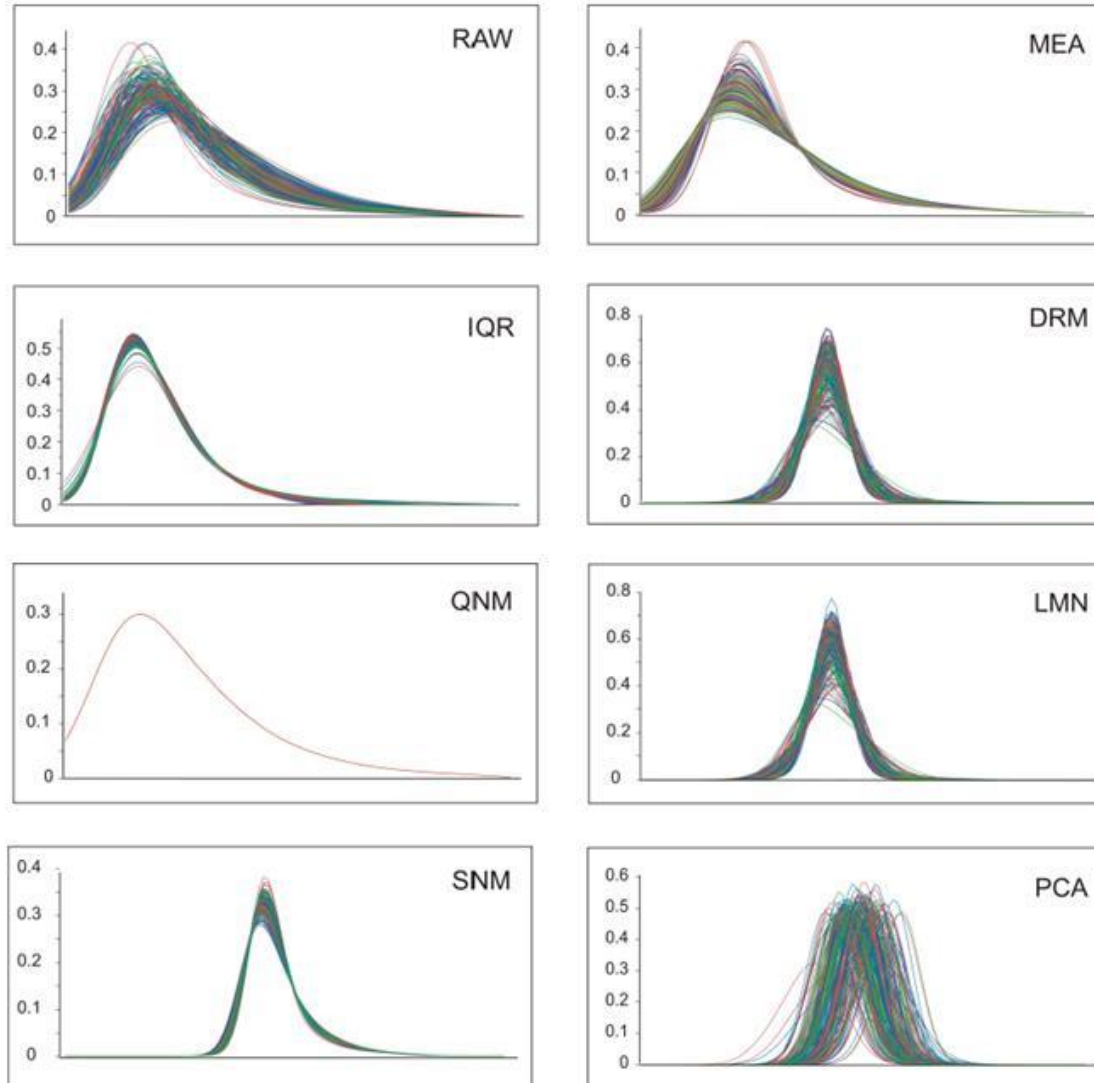
```
> summary(fit2)
      Df Sum Sq Mean Sq F value    Pr(>F)
ethn    1  0.0070  0.00699    0.708    0.409
weight  1  0.0015  0.00153    0.155    0.698
visit   3  0.0175  0.00585    0.593    0.626
person  5  0.7668  0.15335   15.545 1.92e-06 ***
Residuals 21  0.2072  0.00987
---
Signif. codes:  0 '***' 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

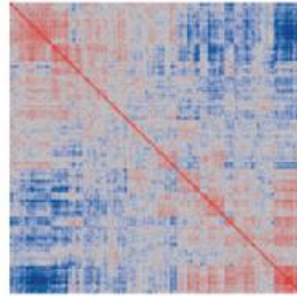


Effect of Normalization on Distributions

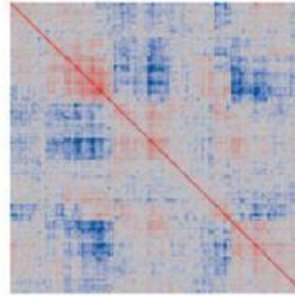


Effect of Normalization on Covariance

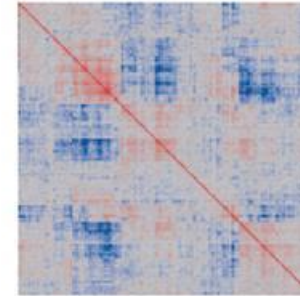
RAW



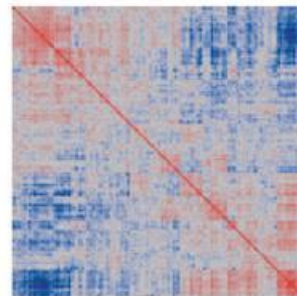
dr3



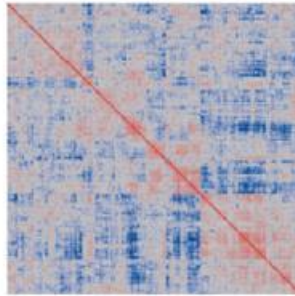
DRM



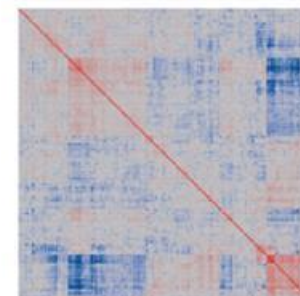
MEA



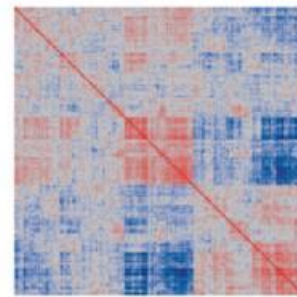
QNM



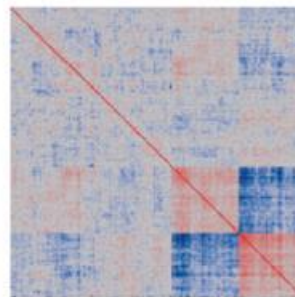
LMN



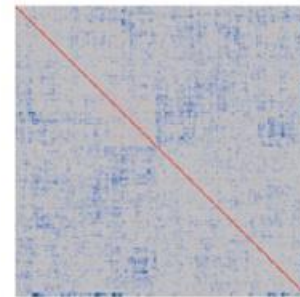
IQR



SNM



PCA



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”).