



Summer Institute
in Statistical Genetics 2018

Integrative Genomics

2a. Dataset normalization

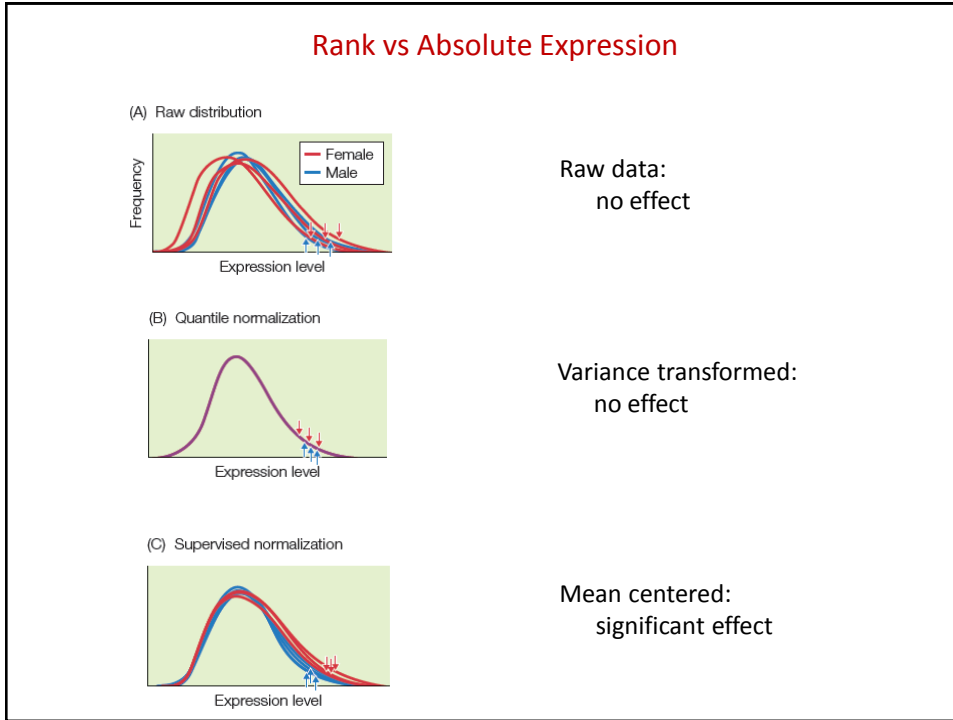


ggibson.gt@gmail.com

<http://www.cig.gatech.edu>

Gene Expression Data is analyzed on the log base 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 sometimes useful to add 1 to all values before taking the log, to avoid "0" returning #NUM! (but not advised with edgeR)



Quantile normalization

	S1	S2	S3	R1	R2	R3	Q1	Q2	Q3
Gene A	10.2	11.2	10.9	3	1	4	10.6	11.9	10.3
Gene B	9.6	10.7	8.9	5	2	6	9.6	11.2	9.2
Gene C	12.7	7.8	11.7	1	9	1	11.9	7.6	11.9
Gene D	9.5	10.0	9.9	6	4	5	9.2	10.3	9.6
Gene E	11.3	9.2	11.1	2	7	3	11.2	8.8	10.6
Gene F	7.7	7.7	6.5	9	10	10	7.6	6.7	6.7
Gene G	5.9	9.3	7.8	10	6	8	6.7	9.2	8.1
Gene H	8.8	8.2	8.4	7	8	7	8.8	8.1	8.8
Gene I	10.1	9.4	11.6	4	5	2	10.3	9.6	11.2
Gene J	8.2	10.6	7.2	8	3	9	8.1	10.6	7.6

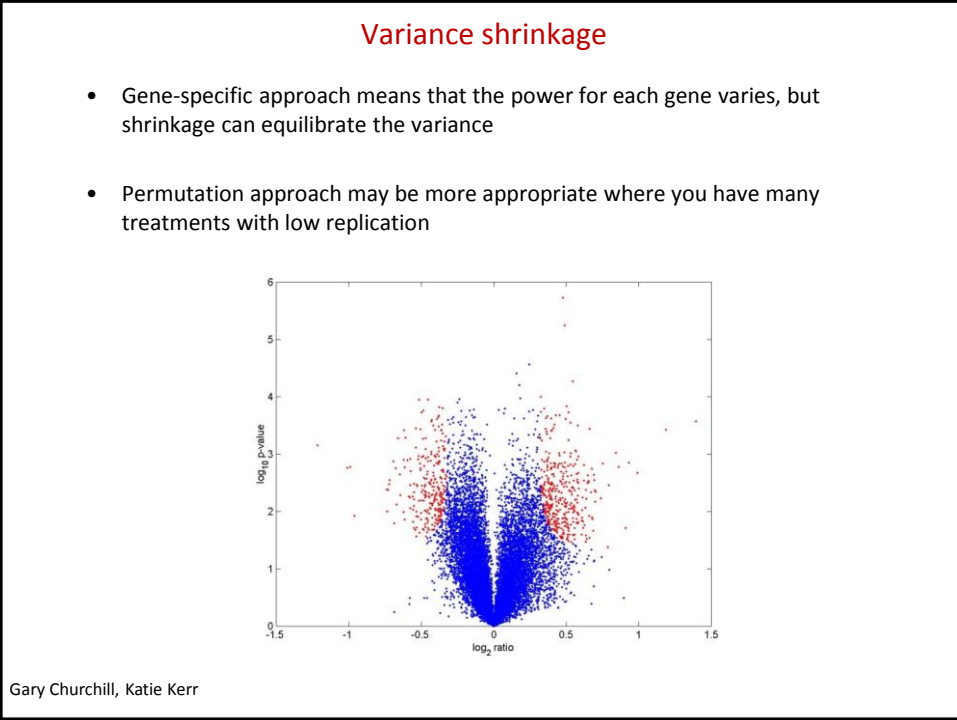
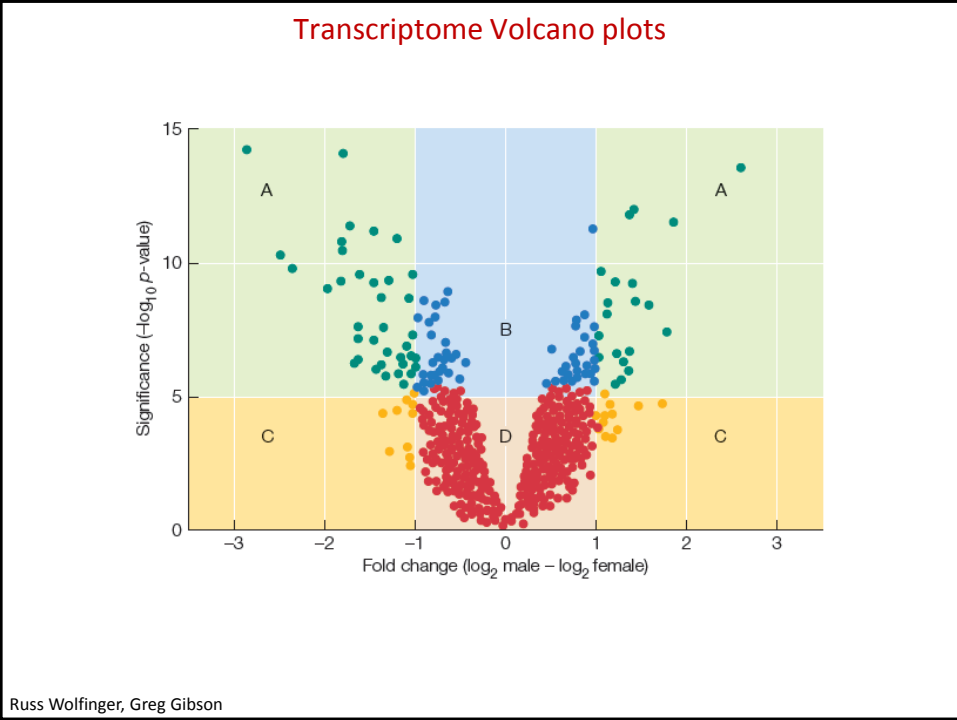
	S1	S2	S3	Avg
Rank 1	12.7	11.2	11.7	11.9
Rank 2	11.3	10.7	11.6	11.2
Rank 3	10.2	10.6	11.1	10.6
Rank 4	10.1	10	10.9	10.3
Rank 5	9.6	9.4	9.9	9.6
Rank 6	9.5	9.3	8.9	9.2
Rank 7	8.8	9.2	8.4	8.8
Rank 8	8.2	8.2	7.8	8.1
Rank 9	7.7	7.8	7.2	7.6
Rank 10	5.9	7.7	6.5	6.7

Step 1: Mean or Median Center Samples

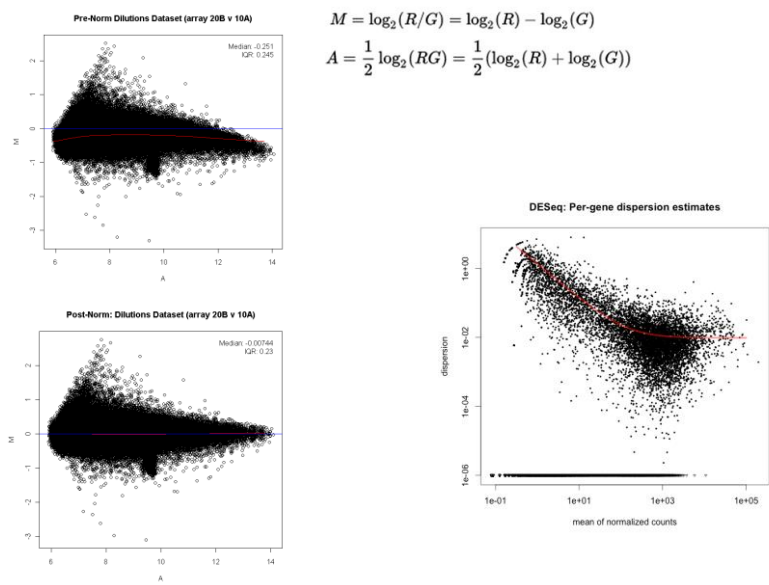
Step 2: Rank each Gene within each Sample

Step 3: Compute average of each Rank

Step 4: Reassign Average Rank to each Gene

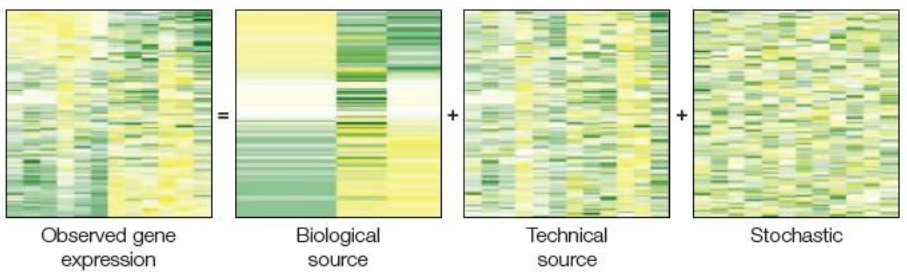


Deviation, Abundance, and Dispersion



Anders et al (2010) *Nature Protocols* 8: 1765-1786

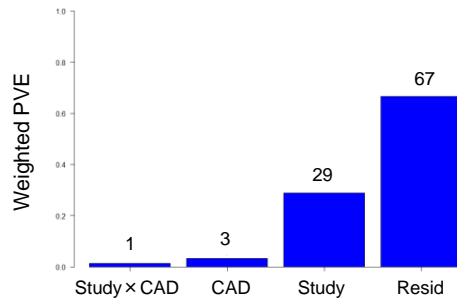
The normalization challenge



John Storey

Principal Variance Component Analysis (PVCA)

- Step 1: Compute the Principal Components (PC) of the dataset
- Step 2: Regress Technical and Biological Parameters on each PC to estimate the proportion of variance explained (PVE)
- Step 3: Sum the PVE, weighted by the strength of the PC
- This tells you how much of the overall variance is due to each factor.



Surrogate Variable Analysis (SVA)

The idea is to identify components of variation that are due to technical factors (batch effects, RNA quality, other unknown influences) and remove them so that the biological factors of interest are enhanced.

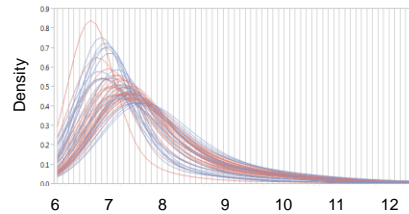
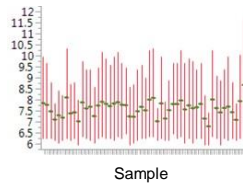
- Step 1: Compute the Surrogate Variables
 Fit the biological factors of interest to generate residuals
 Identify the PC of the residuals
 SV are weights applied to those genes that correlate with the PC
- Step 2: Ask which SV are correlated with Technical factors
- Step 3: Decide which ones to remove or at least adjust for

SNM (Supervised Normalization of Microarrays) optimizes this process in an iterative fashion, but has not been revised for RNASeq.

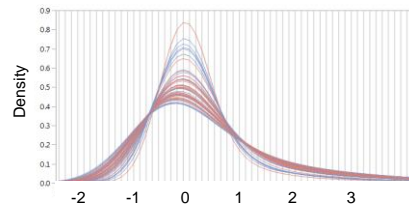
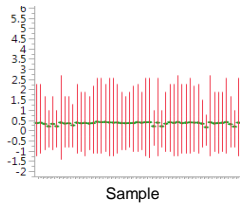
PEER factor normalization is a Bayesian version which automatically fits all of the surrogate variables, and is commonly used for cis-eQTL analysis.

Simple Centering

Raw Profiles



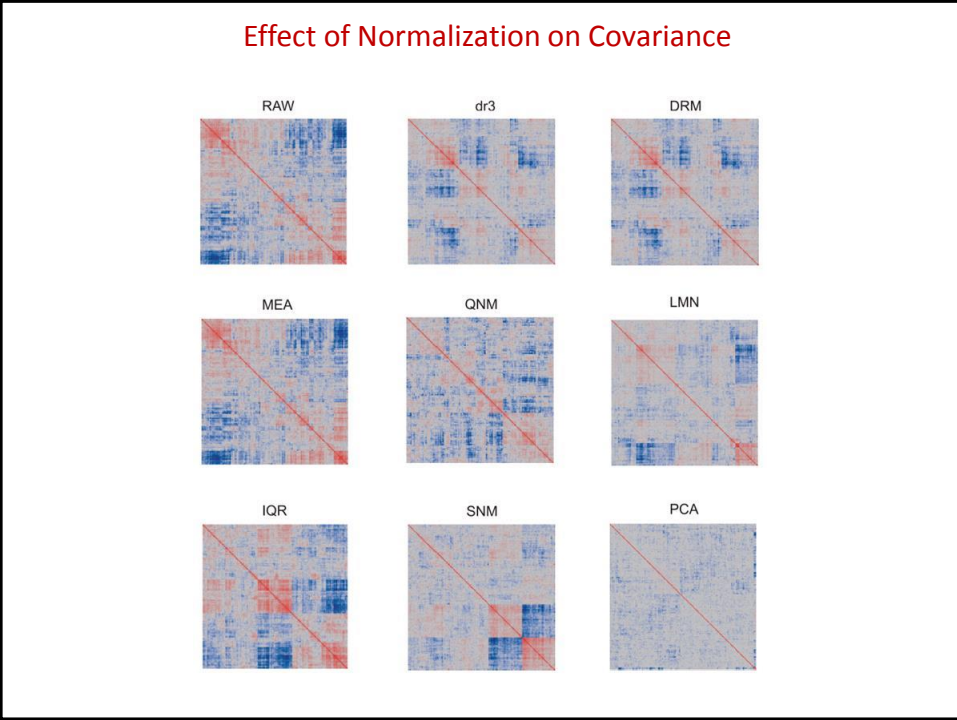
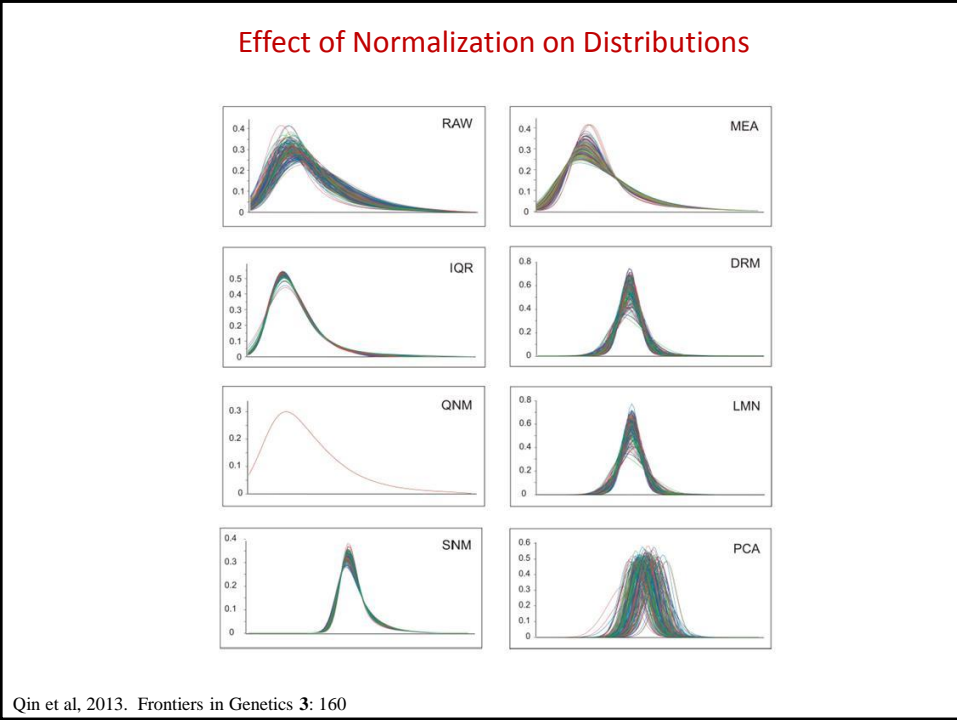
Median Transform

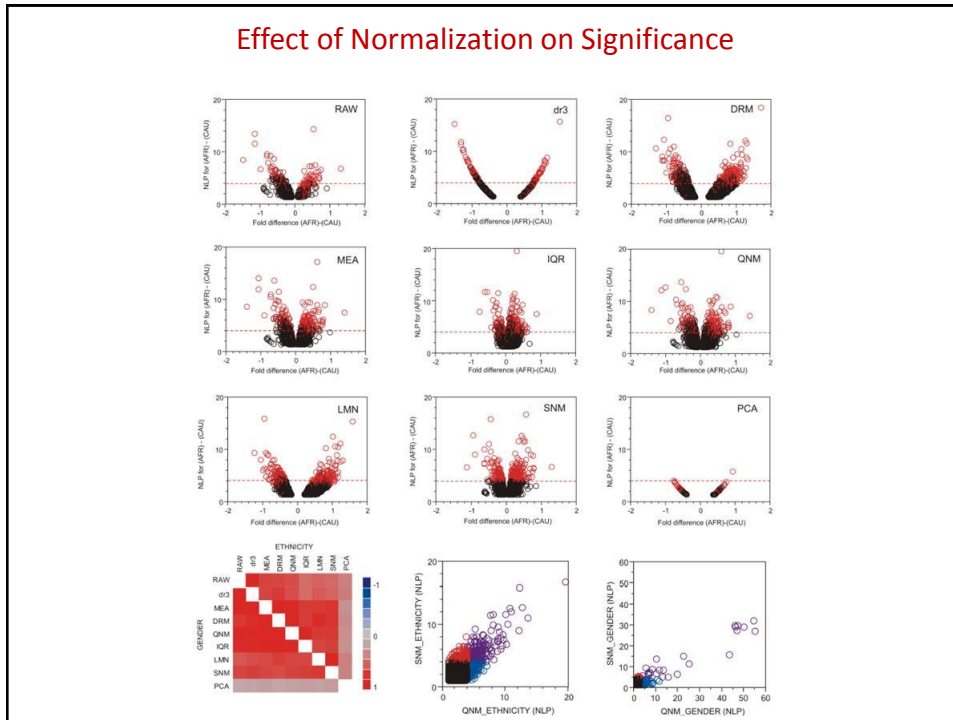


For RNASeq data, CPM essentially does this: $\text{cpm} = 1,000,000 \times \text{reads}/\text{total reads}$

Types of 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





Recommended analytical strategy

1. Normalize the samples
2. Extract the Principal components of gene expression
3. Ask whether the major PC are correlated with technical covariates such as Batch or RNA quality; or with Biological variables of interest
4. If they are, renormalize to remove those effects
 - (PEER factor normalization is a Bayesian approach to fitting Surrogate Variables;
 - SVA is a linear modeling approach often performed with COMBAT;
 - SNM is a supervised approach that allows you to retain Biological factors while fitting or removing technical ones)
5. 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