The Nature of RNA-Seq Data

Vince Buffalo
Bioinfomatics Core
UC Davis Genome Center

February 9, 2012

Outline

What I'll (quickly) discuss

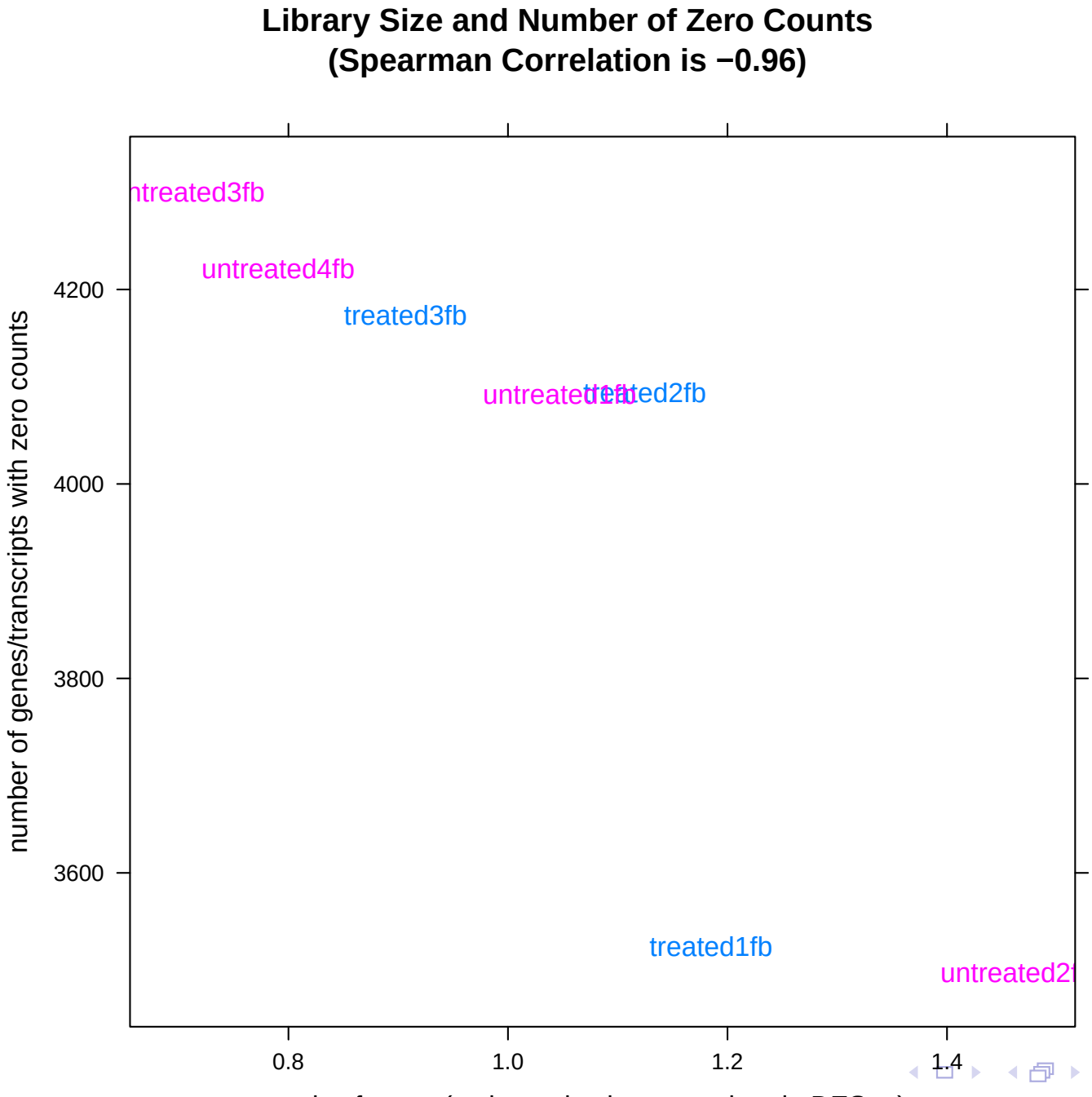
- What RNA-seq data looks like.
- Issues with RNA-seq data.
- Issues with normalizing RNA-seq data.

What factors contribute to better detection of expressed genes?

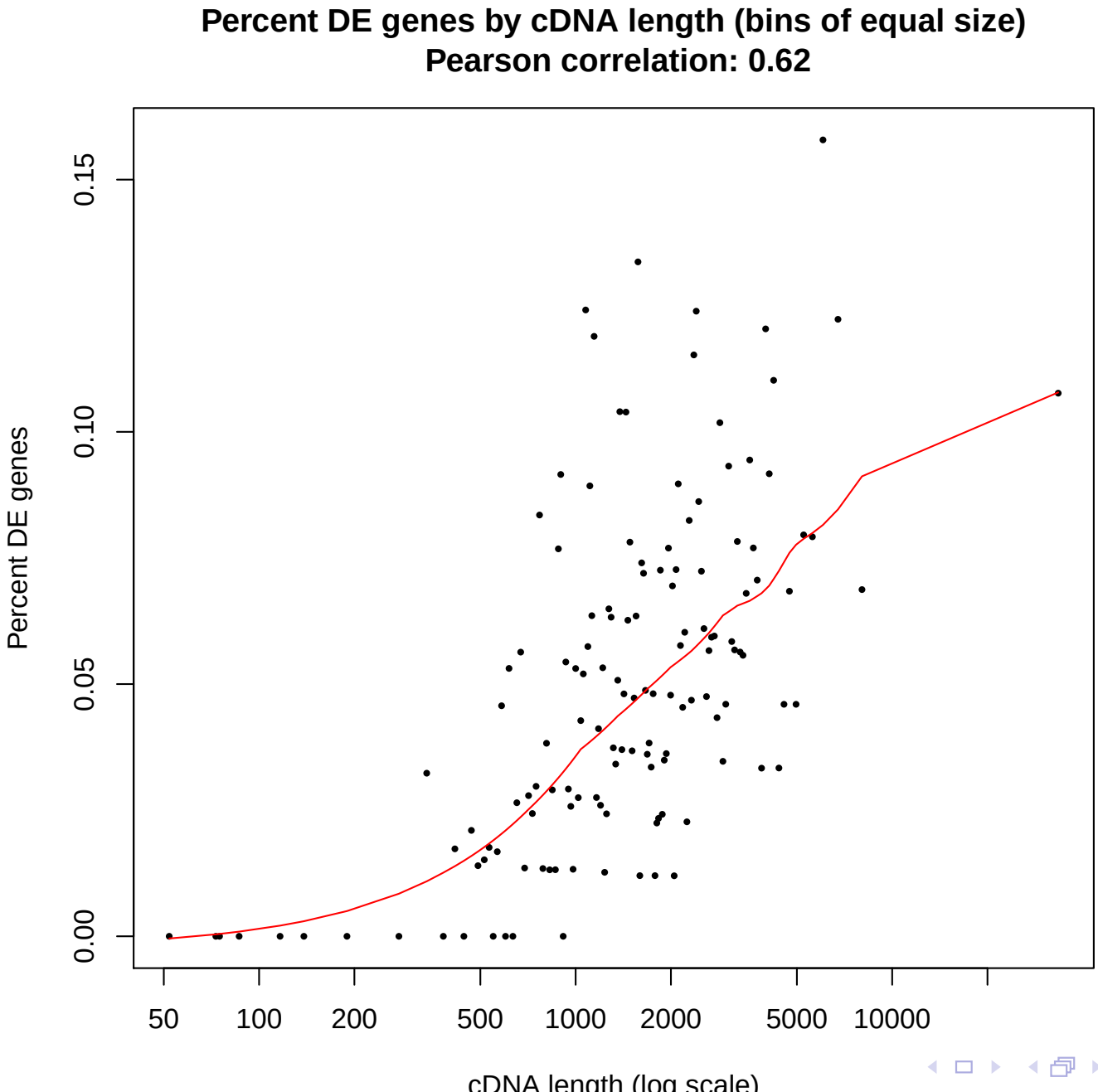
- Gene/transcript length: longer are more easily detected.
- Sample concentration: higher concentration leads to better detection of lowly-expressed genes, and better estimate of expression.

- The concentrations of cDNA samples vary. What impact does this have?
  - Increased coverage of all expressed genes (n counts to cn counts).
  - Increased detection of lowly-expressed genes (0 counts to n counts).

Library size and detection of low-expressed genes



Gene Length and Differential Expression

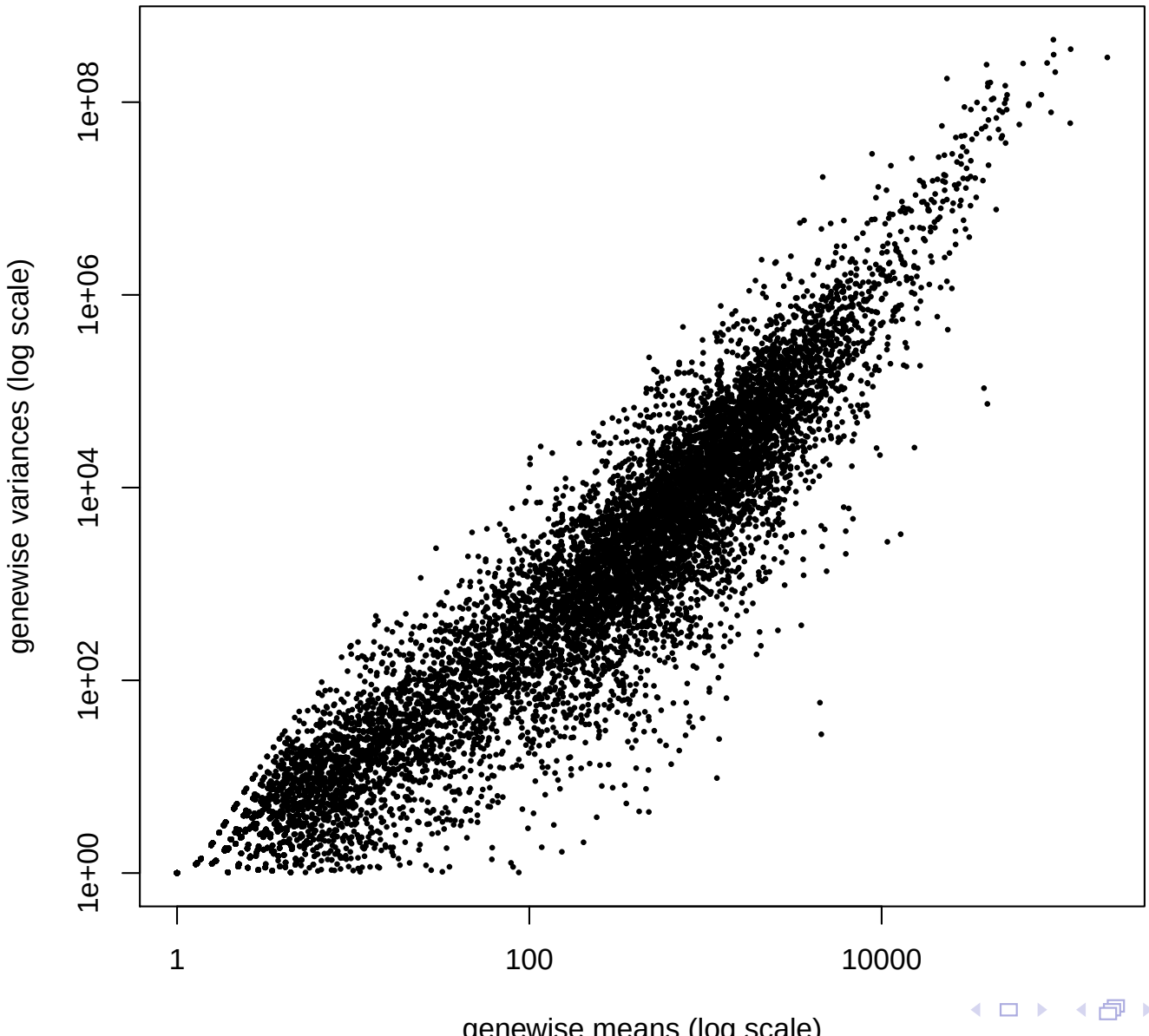


Gene Length Bias

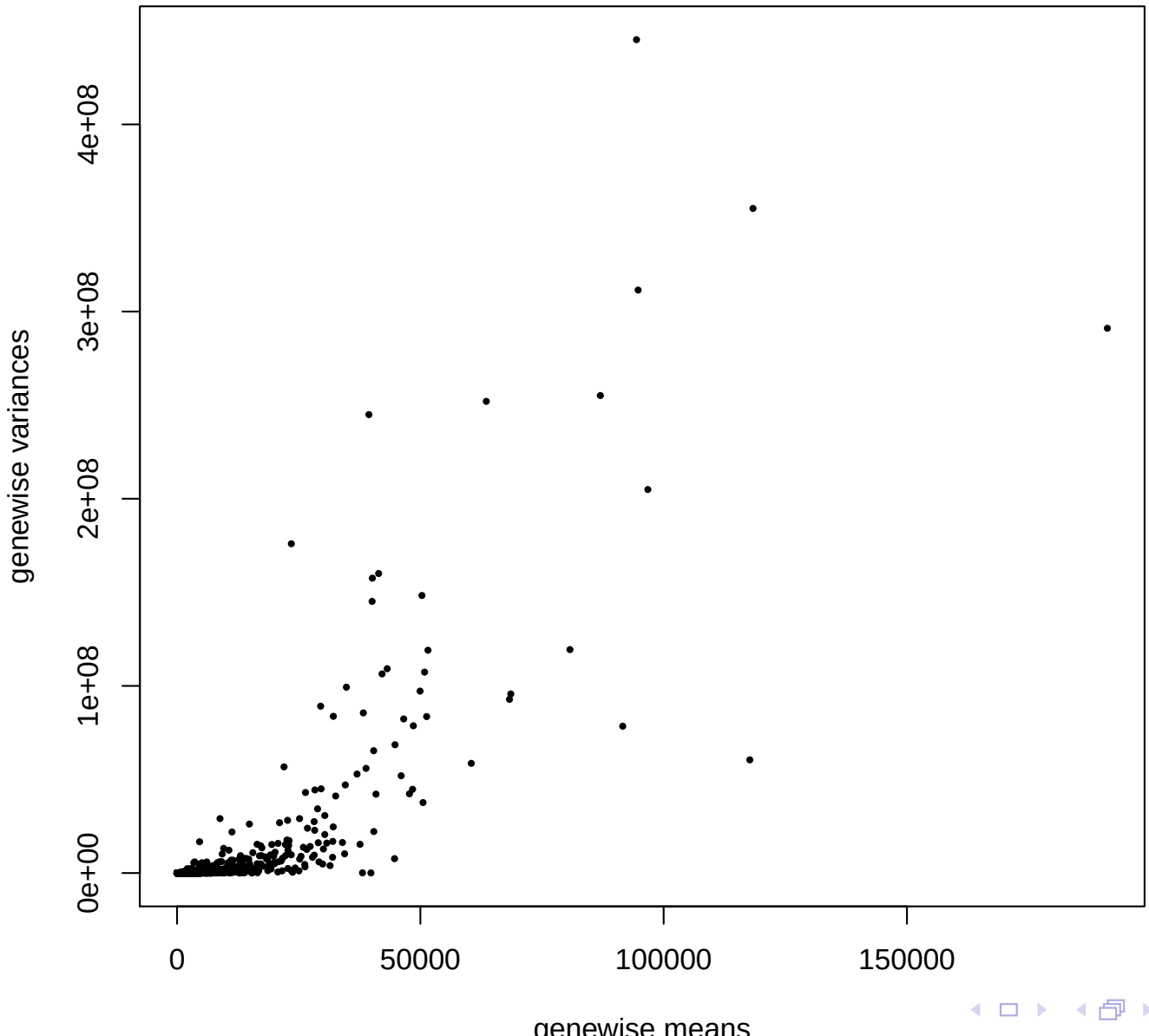
Oshlack, et al. 2009 talk about this bias extensively.

- How do we get around this?
- Will all findings be confounded by differing power due to gene length?
- How can we handle this effect in cross-gene, within-sample comparisons? Can we use RNA-seq at all?

The Variance-Mean Relationship



The Variance-Mean Relationship (not log scale)



Why does this matter?

- We need to model this explicitly (`DESeq`, `edgeR`, etc).
- This variance is *greater* than the mean (overdipsersion) in almost all cases in which there are biological replicates. This is due to biological heterogeneity in individuals.
- Highly expressed genes have high variance; lowly expressed genes have low variance. Any machine learning methods that use variance or distance (PCA, sparsePCA, the Lasso, distance-based clustering) will be negatively affected by this.

Distance-based methods

$d(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \ldots + (x_n - y_n)^2}$

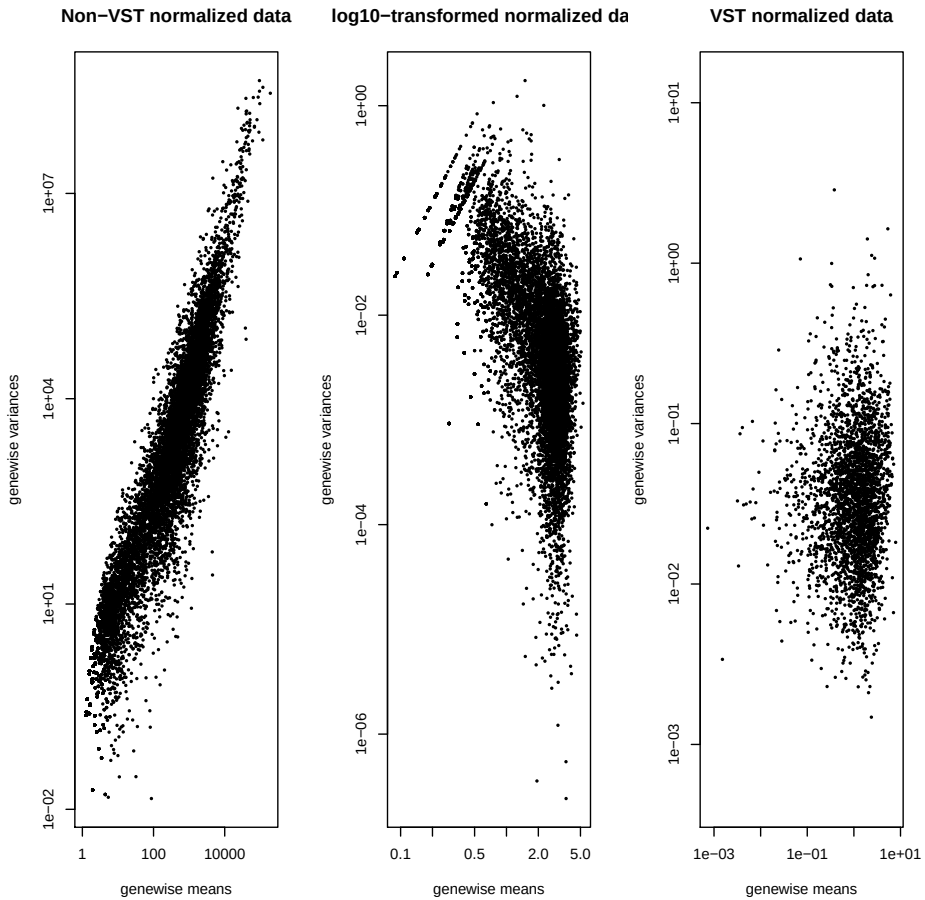
Suppose x and y are two *replicates*. If y_1 and x_1 are the expression values for a highly expressed gene, we know that it will have high variance, and these values will likely be very different.

If x_2 and y_2 are the expressed values for a lowly expressed gene, their difference will likely be less than that of y_1 and x_1.

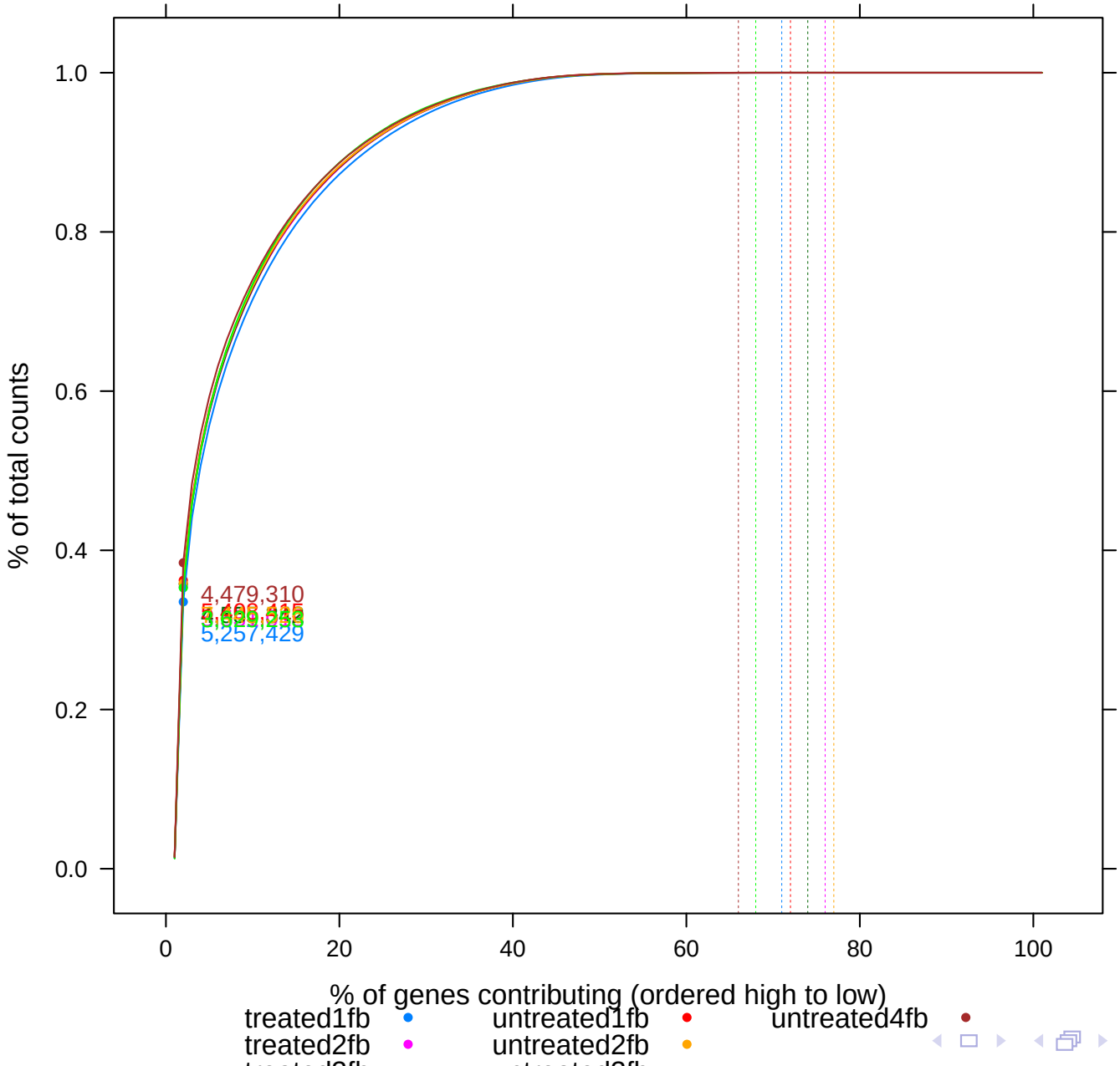
- What drives this distance calculation?
- Is this desirable?

Variance Stabilizing Transformations

Note: every scale is log-transformed for comparison to the original data!



Gene counts as a proportion of total lane counts



RPKM

$$\frac{\text{reads mapped}}{\text{mapped reads (in millions)} \cdot \text{gene length (in KB)}}$$

The RPKM motivation

Suppose we have one replicate with counts q_1, q_2, etc. Replicate two has counts p_1, p_2, etc; we know *a priori* that we put twice as much sample into replicate two as one. Thus, a global scaling factor approach works.
RPKM approach assumes that total lane counts accurately estimates sample concentration in all cases. Is this true?

Highly expressed genes

It's not. Some genes can dominate lane counts.
The top 1% of highly-expressed genes can make up a huge proportion of total lane counts. Scaling by total lane counts then can bias differential expression results.
Thought experiment: if 400 genes (of 30,000) made up 80% of lane counts, would you really want to scale the remaining 29,600 genes' counts by a value that's 80% composed of 1.3% of the genes' expression?

Better normalization techniques

- Quantile normalization (not a scaling factor technique).
- DESeq's method (use a more robust scaling factor):
  1. Take the geometric mean of all rows (across samples, per gene) to create a reference sample.
  2. Calculate the ratio of a sample's counts to the reference sample counts, for each gene.
  3. Find the median of all these genewise ratios to get the relative library depth.