Differential Expression Analysis

- Statistical model used to estimate fold changes, test for significance
  - Negative binomial model: edgeR, DESeq/DESeq2, cuffdiff2
  - Gaussian (normal) model: limma-voom
- Input is raw counts, which are then normalized prior to (or as part of) analysis
  - Typically not FPKM or RPKM

Normalization

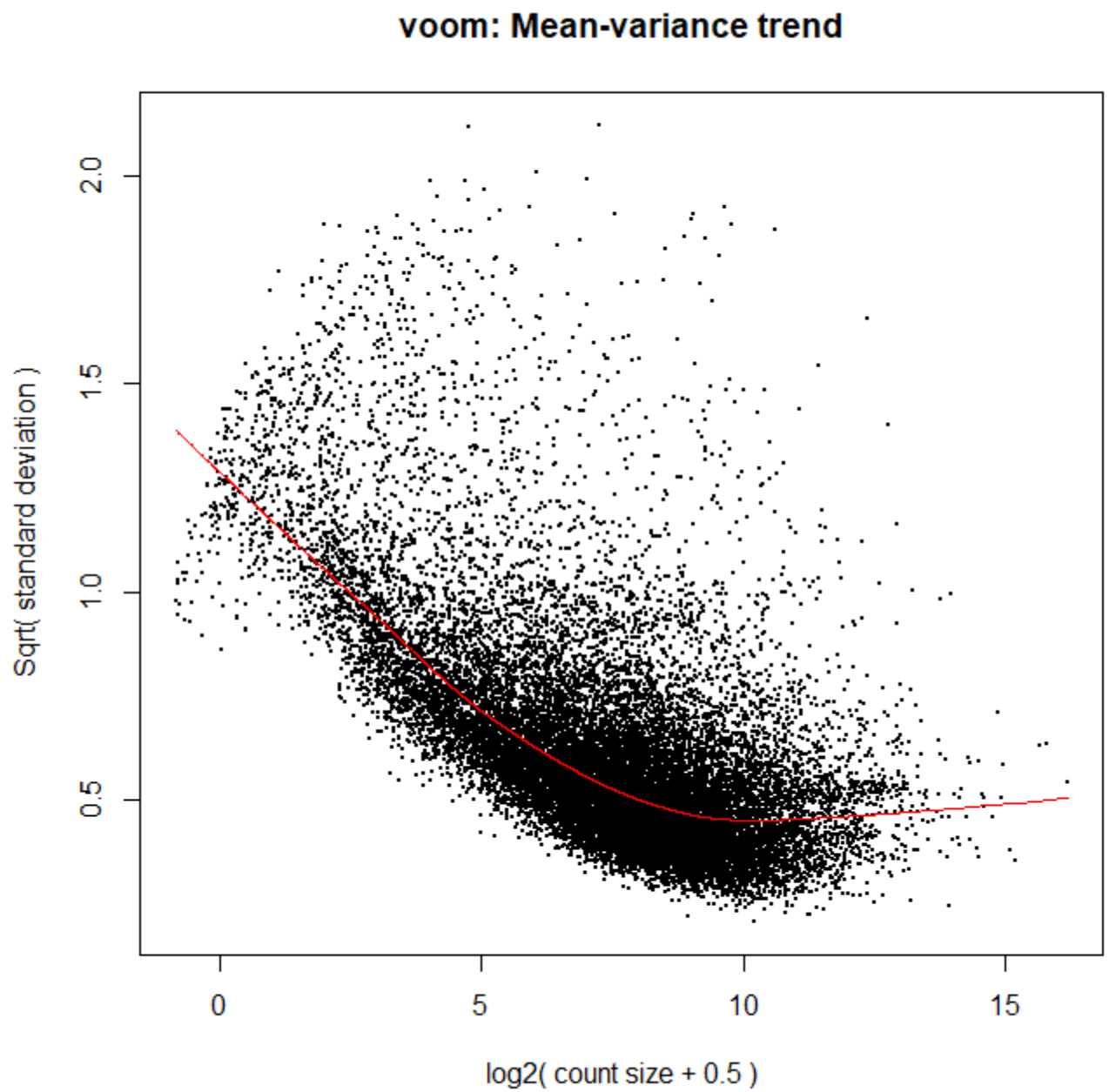
- For DE analysis, only sample-specific effects need to be normalized for
  - Sequencing depth is sample specific
  - GC content typically not sample specific
  - Gene length typically not sample specific
    - FPKM (Fragments per kb per million) and RPKM (Reads per kb per million) adjust for length
- Batch effects need to be accounted for, but often easiest to include batch in model

TMM Normalization

- edgeR uses TMM normalization by default
  - Scaling factors incorporated in statistical model
- TMM = trimmed mean of M values
- Based on log-fold changes between samples for medium-expression genes
- Many other procedures exist, TMM performs well by comparison
- Don’t use raw library size itself for normalization
  - Can be dominated by a few highly expressed genes

Voom Transformation and Weighting

- The voom function in edgeR modifies RNA-Seq data for use with the limma Bioconductor package
- Counts transformed to log2 CPM (counts per million reads)
  - “Per million” defined based on library sizes adjusted for normalization factors
- Linear models like those in limma usually assume constant variance
  - Log transformation fixes some of the mean-variance dependency
  - Variance weights from voom, which are then passed into limma, take care of the rest


voom: Mean-variance trend
Sqrt(standard deviation)
0.5
1.0
1.5
2.0
0
5
10
15
log2(count size + 0.5)

Why limma?

- Limma was developed for microarrays, based on Gaussian linear models
  - Why use such an old package?
- Negative binomial-based packages can suffer from an inflated FDR
- Limma-voom does not appear to have this problem
- Powerful, flexible, fast

General Steps in a Limma-Voom Analysis

- Calculate normalization factors
- Filter low expressed genes
- Define linear model
- Calculate voom weights
- Fit per-gene linear models
- Fit contrasts
- Use empirical Bayes smoothing to get better s.e. estimates for contrasts than the per-gene ones
- Adjust p-values for false discovery rate (Benjamini-Hochberg by default)
- Assess results (do they make sense?), enrichment analysis, etc.