Voom: precision weights unlock linear model analysis tools for RNA-Seq read counts

LAW CW, CHEN Y, SHI W, AND SMYTH GK
GENOME BIOLOGY 2014

PRESENTED BY LAURA SABA

COPIES OF SLIDES AVAILABLE AT HTTPS://GITHUB.COM/LAURASABA

LIMMA Package

- Original Paper - Smyth, G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology, 3, Article3.*
- R Package available thru BioConductor: *http://www.bioconductor.org/packages/release/bioc/html/limma.html*
- Averages over 10,000 downloads per month

LIMMA Package – General Idea

- Small sample sizes → bad variance estimates
- Empirical Bayes Method "borrows" information across genes to adjust variance estimates for a gene.
  - Moderated t-statistic or F-statistic replaces the estimated residual variance with a posterior estimate of the residual variance.
  - It "shrinks" the gene-specific original variance to the typical variance across genes, i.e., the prior for the estimate of the residual variance is derived from the data across genes (empirical part).

LIMMA – Example Code

```
###################################
####  Differential Expression  ####
###################################

library(limma)

# pre-Analysis filtering - eliminate control probes
filtered.data <- exprs(rma.data)[!(1:nrow(rma.data) %in% grep("AFFX",rownames(exprs(rma.data)))),]
filtered.calls <- exprs(calls)[!(1:nrow(rma.data) %in% grep("AFFX",rownames(exprs(rma.data)))),]

# pre-Analysis filtering - eliminate probes that are not present in any sample
filtered.data <- filtered.data[rowSums(filtered.calls == "P")>0,]
filtered.calls <- filtered.calls[rowSums(filtered.calls == "P")>0,]

# calculated differential expression between strains using empirical Bayes method
strain <- as.factor(pData(filenames)$strain)
design <- model.matrix(~ -1 + strain)                                    # fit means model
contrast.matrix <- makeContrasts(strainC57 - strainDBA,levels=design)    # identify comparison of interest
fit <- lmFit(filtered.data,design=design)
contrast.results <- contrasts.fit(fit,contrasts=contrast.matrix)
eBayes.results <- eBayes(contrast.results)

# apply FDR as a multiple testing correction
fdr.values <- p.adjust(eBayes.results$p.value,method="BH")

# create a data set with strain means, fold change, p-value, and FDR
results <- as.data.frame(cbind(fit$coefficients,contrast.results$coefficients,eBayes.results$p.value,fdr.values))
colnames(results) <- c("C57.Mean","DBA.Mean","Fold.Change","p.value","FDR")

# output all data to csv file
output <- cbind(rownames(results),results)
colnames(output)[1] <- "ProbeID"
write.table(output,file="All.Output.csv",sep=",",quote=FALSE,row.names=FALSE)

# output significant probe sets to txt file (this file can be directly uploaded to PhenoGen
output2 <- rownames(results)[results$FDR < 0.05]
write.table(output2,file="SigProbeSets.txt",sep="\t",quote=FALSE,row.names=FALSE,col.names=FALSE)
```

LIMMA with RNA-Seq

- LIMMA is based on the normal distribution of residuals → even logged values of counts are still not normal and there is a dependence between read count and variance
- Most current methods for RNA-Seq differential expression use the Negative Binomial distribution which is better suited for count, but you lose some of your flexibility for types of comparisons that can be conducted.
  - The 'dispersion' factor is what causes the problems.
  - Estimating this factor is tricky and assuming it is a fixed value is problematic.

Comparison of Variance vs. Read Count

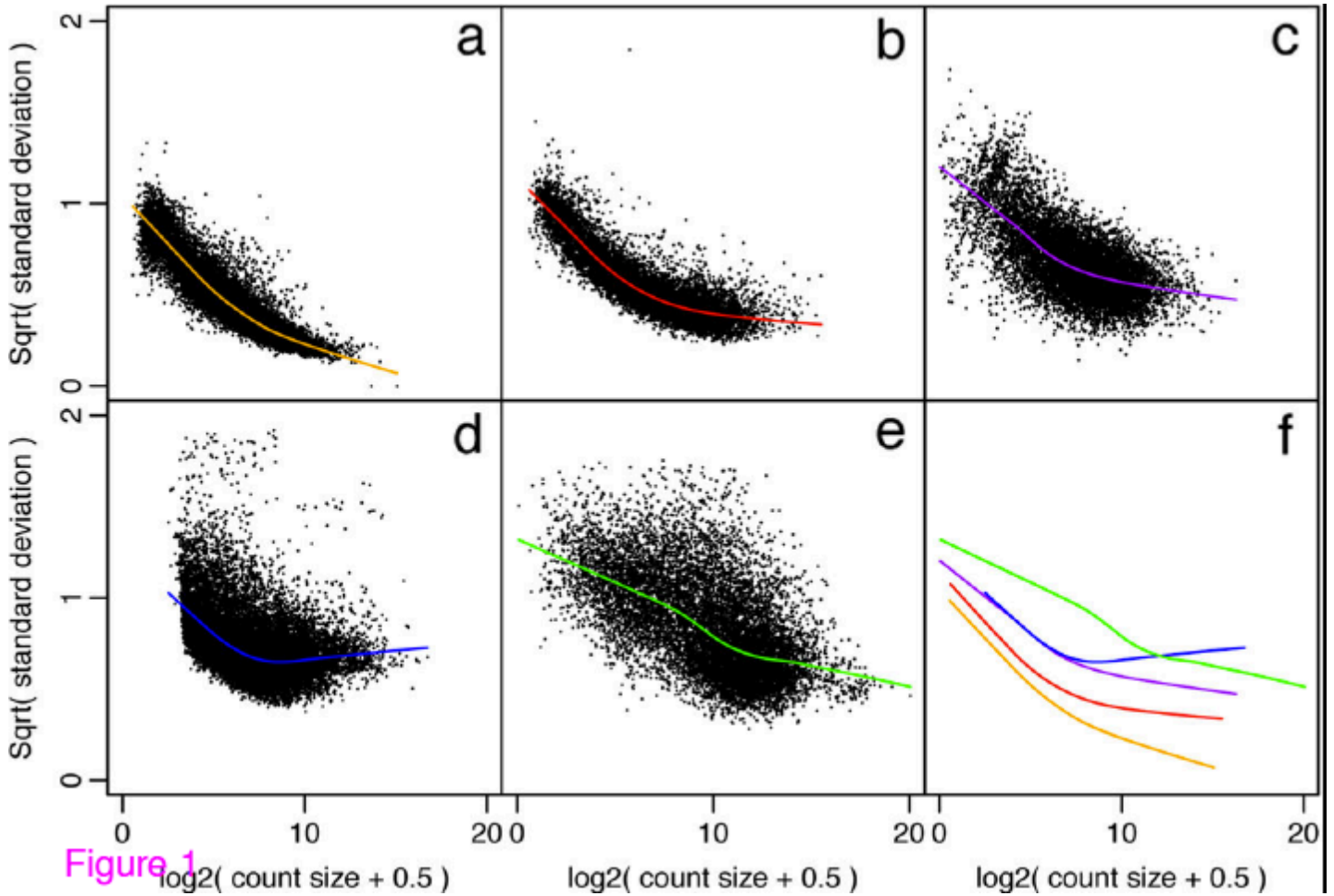


Figure 1

Figure 1 mean-variance relationships. Gene-wise means and variances of RNA-Seq data represented by black points with a lowess trend. Plots are ordered by increasing levels of biological variation in datasets. Panel (a), voom trend in HBRR and UHRR genes in Sample A, B, C and D of SEQC project; technical variation only. Panel (b), C57BL/6J and DBA mouse experiment; low-level biological variation. Panel (c), simulation study in the presence of 100 up-regulating genes and 100 downregulating genes; moderate-level biological variation. Panel (d), Nigerian lymphoblastoid cell lines; high-level biological variation. Panel (e), Drosophila melanogaster embryonic developmental stages;
very high biological variation due to systematic differences between samples. Panel (f), lowess voom trends for datasets a–e.

Proposed Solution

- Estimate the mean-variance relationship in the logged count data from the data across all genes and develop precision weights for individual sample/ gene observations
  - “correct modeling of the mean-variance relationship inherent in a data generating process is the key to designing statistically powerful methods of analysis”

ALGORITHM

2 methods explored

- *limma-trend* applies the mean-variance relationship at the gene level
- *voom (variance modeling at the observational level)* applies the weights at the individual level

Steps to VOOM Analysis

1. Calculate log-cpm values
2. Fit gene-wise linear models and calculate a residual standard deviation for each gene
3. Fit a robust trend to residual standard deviations (dependent variable) based on average log count (independent variable)
4. The gene-wise linear model in Step 2 is used to predict the log-cpm value for each sample.
5. The predicted log-cpm value is converted to a predicted log count for each sample based on library size.
6. The robust trend in Step 3 is used to predict the standard deviation of individual values based on their predicted count
7. The inverse squared predicted SD of each observation becomes its weight.

Step 1 – Calculate log-cpm values

- cpm = counts per million mapped reads
- log-cpm = log base 2 of cpm

$$\log - \mathrm{cpm}_{\mathrm{gi}} = \log 2\left(\frac{r_{gi} + 0.5}{R_i + 1.0} \times 10^6\right)$$

- don't need to adjust for transcript length because method is only intended for comparisons of the same gene across samples
- smoothly decreasing mean-variance trend with count size
- variance and mean are independent at **high** read counts

Step 2 – gene-wise linear models

The benefit of moving away from a NB model to a linear model is that this model have a variety of structures, e.g., 2-way ANOVA, repeated measures, etc.

1. Fit gene-wise linear models based on study design
2. Extract residual standard deviation from this model for each gene

Step 2 – gene-wise linear models

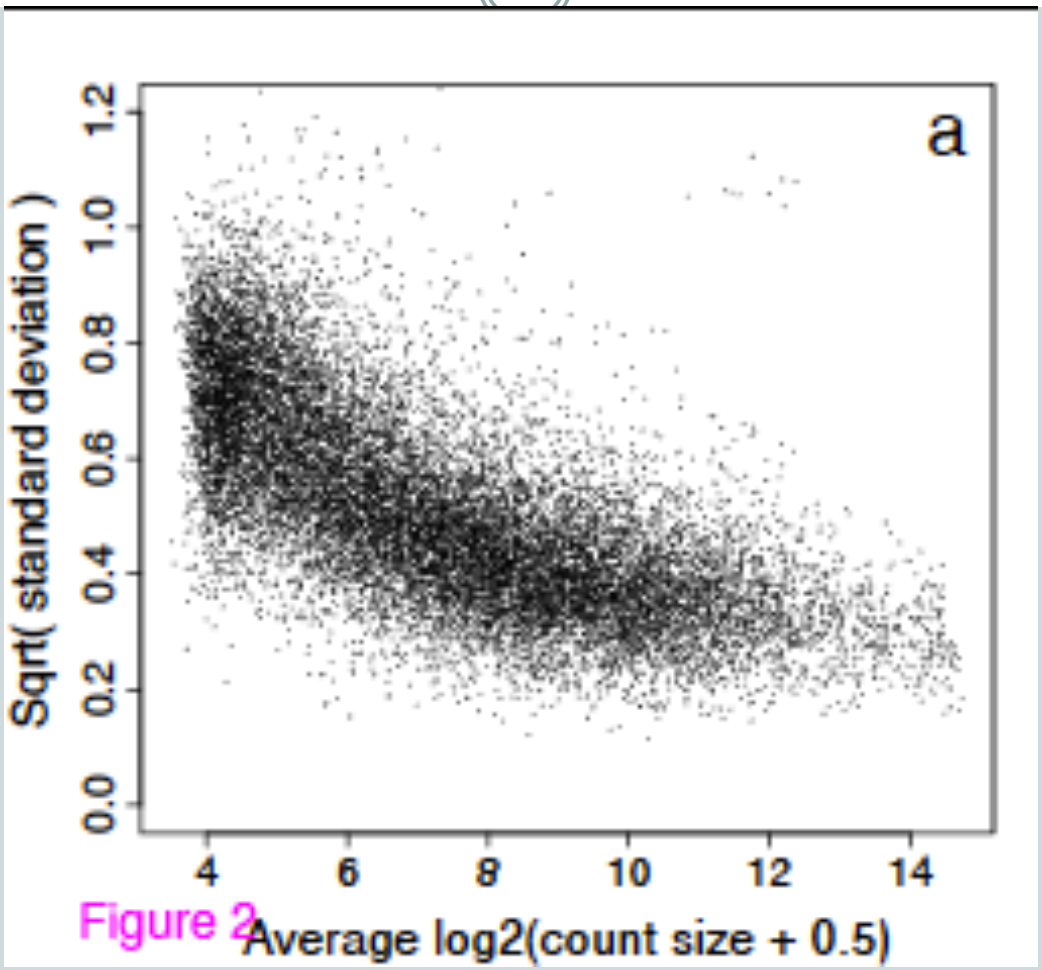


Figure 2

Step 3 – Robust Trend between SDs and Average Log Count

1. Convert Average log-cpm to average log count

$$\tilde{r} = \bar{y}_g + \log_2(\tilde{R}) - \log_2(10^6)$$

where $\tilde{R}$ = geometric mean of the library sizes plus 1

2. Fit LOESS curve to square root of the standard deviations as a function of the average log count.

$$s_g^{1/2} = lo(\tilde{r})$$

Step 4 – Estimate Fitted log-cpm Values

For each sample i and gene g, calculate $\hat{\mu}_{gi}$ from the gene-wise linear model estimated previously.

The exact formula for $\hat{\mu}_{gi}$ will depend on your study design

Step 5 – Convert fitted log-cpm value to fitted log count value

$$\hat{\lambda}_{gi} = \hat{\mu}_{gi} + \log_2\left(R_i + 1\right) - \log_2\left(10^6\right)$$

$\hat{\lambda}_{gi}$ is the fitted log count value for subject i, gene g

$\hat{\mu}_{gi}$ is the fitted log - cpm value for subject i, gene g

R_i is the library size for subject i

Step 6 – Predict Standard Deviation of Individual Values

Using the loess regression, lo(), estimated in Step 3, predict standard deviation of individual values.

$$\text{predicted standard deviation of } y_{gi} = \left[lo\left(\hat{\lambda}_{gi} \right) \right]^2$$

Step 7 – Calculate Weight for Each Observation

$$w_{gi} = lo\left(\hat{\lambda}_{gi}\right)^{-4}$$

The higher the predicted variance, the lower the weight

RESULTS

Control of Type 1 Error Rate

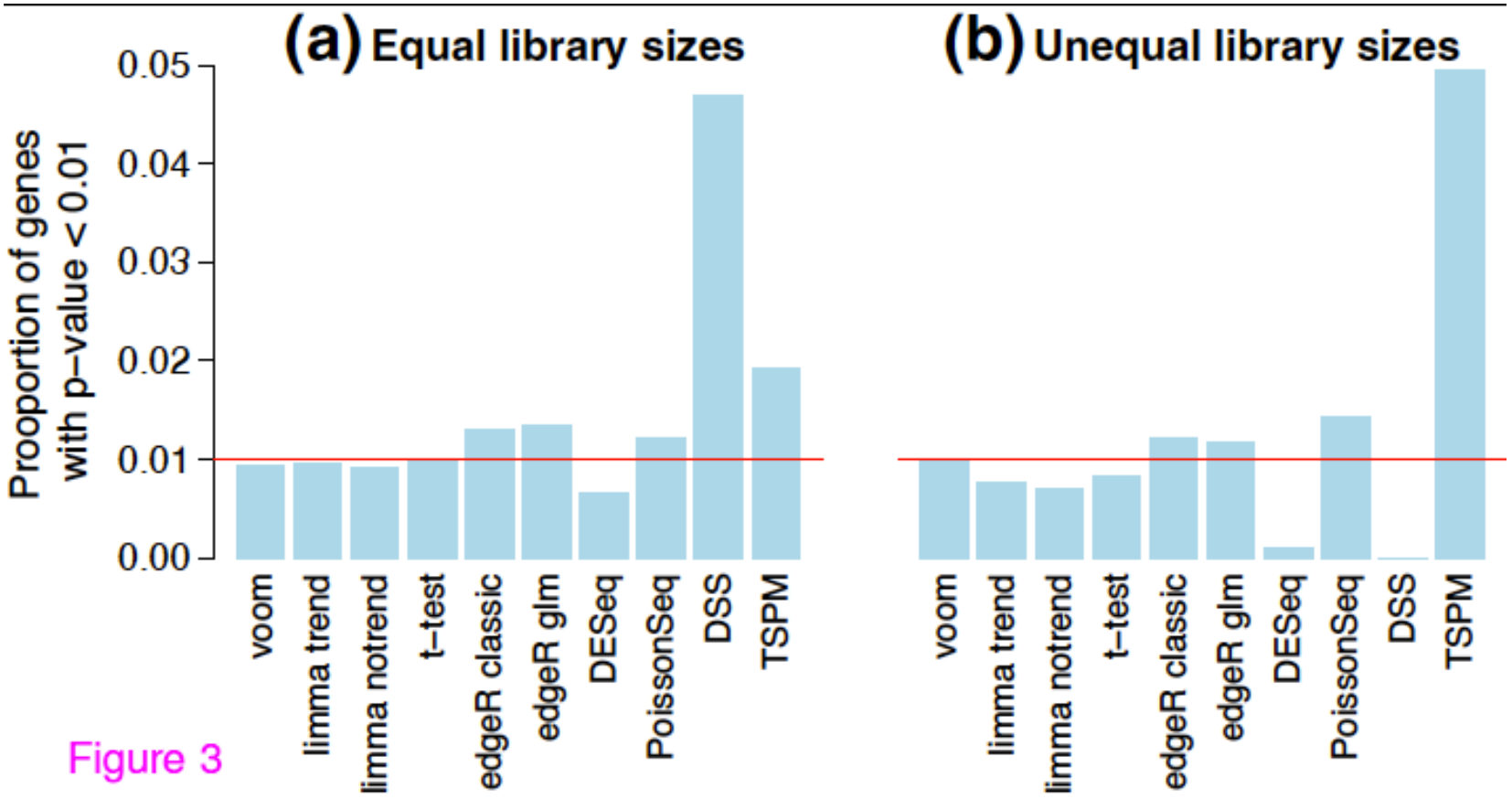


Figure 3 Type I error rates in the absence of true differential expression. The barplots show the proportion of genes with p-value < 0.01 for each method (a) when the library sizes are equal and (b) when the library sizes are unequal. The red line shows the nominal type I error rate of 0.01. Results are averaged over 100 simulations. Methods that control the type I error at or below the nominal level should lie below the red line.

Comparison of Power To Detect Differences

- 2-Group comparison
  - 100 genes 2-fold up-regulated in one group; 100 genes 2-fold down-regulated in the other group

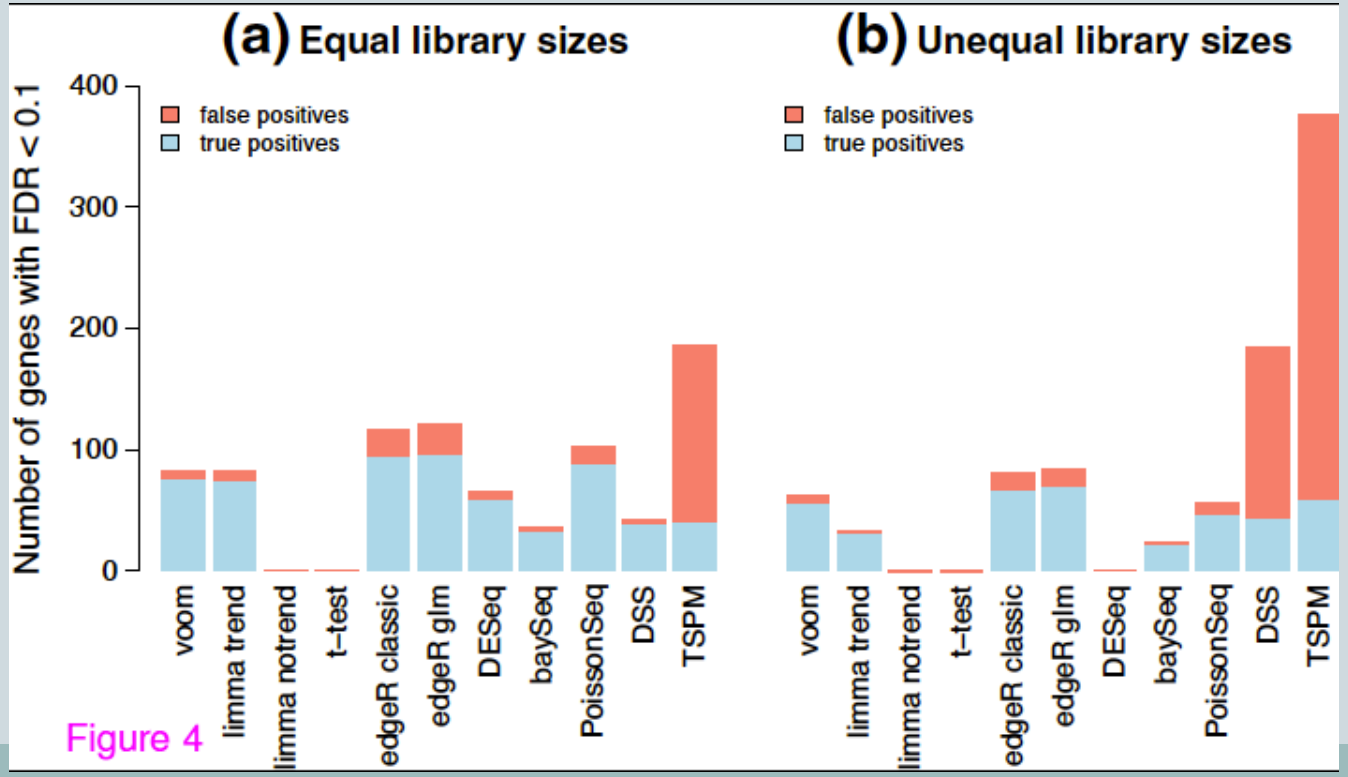


Figure 4

Comparison of FDR – Simulated Data

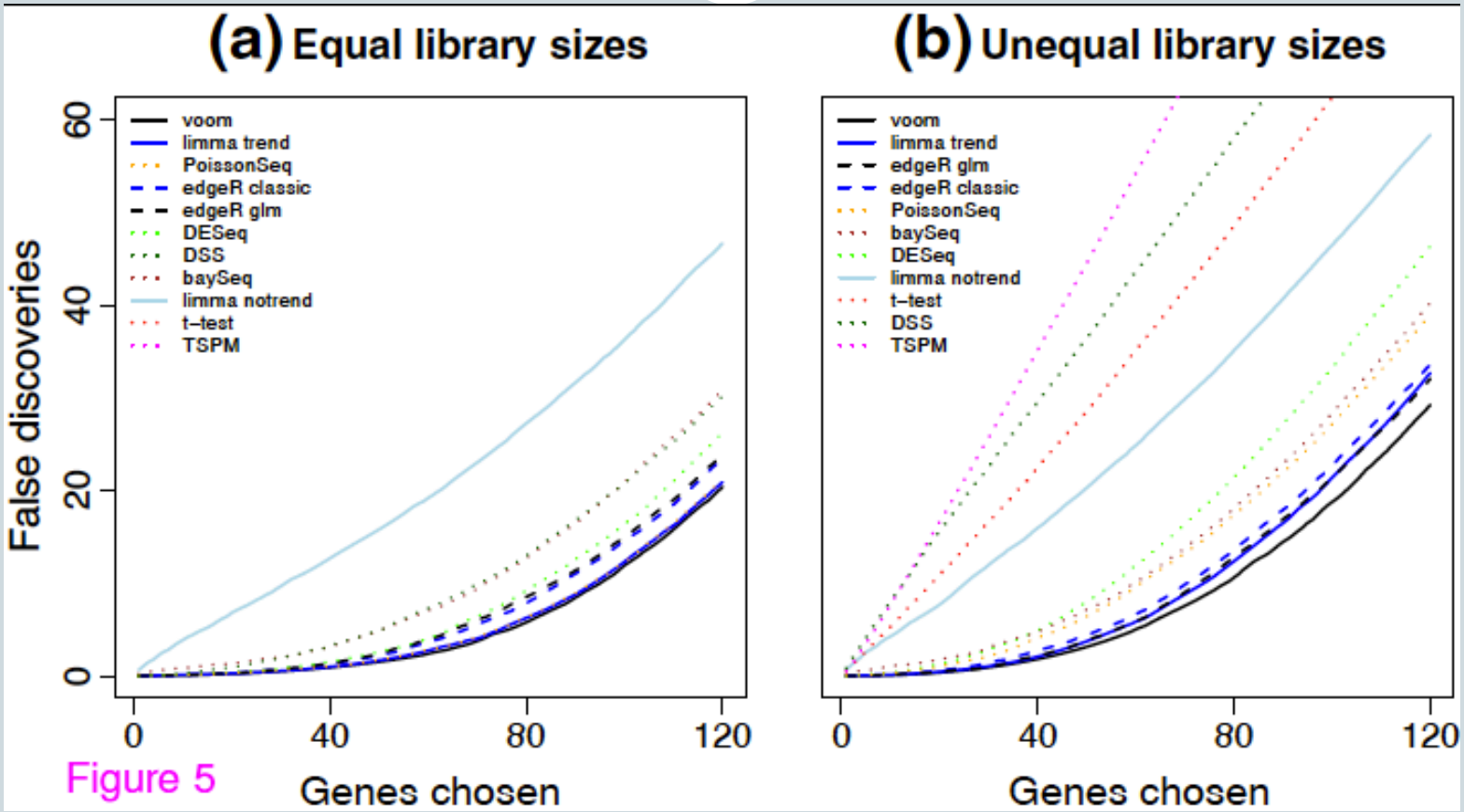


Figure 5 False discovery rates. The number of false discoveries is plotted for each method versus the number of genes selected as DE. Results are averaged over 100 simulations (a) with equal library sizes and (b) with unequal library sizes. Voom has the lowest FDR at any cutoff in either scenario.

Comparison of FDR – Spike Ins

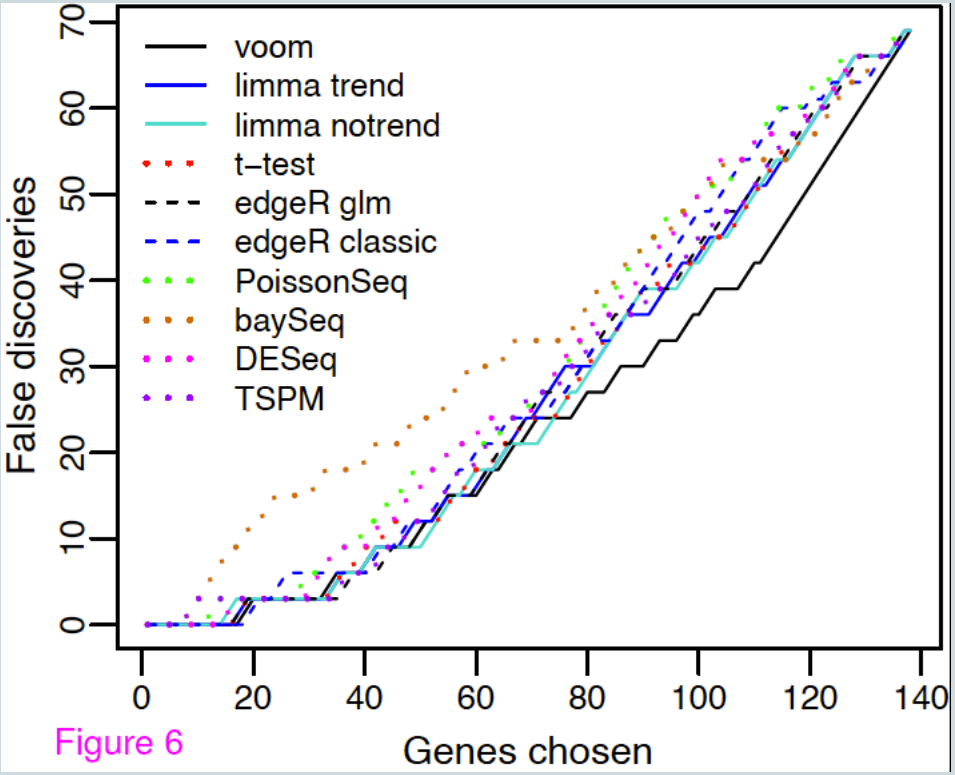


Figure 6 False discovery rates evaluated from SEQC spike-in data. The number of false discoveries is plotted for each method versus the number of genes selected as DE. voom has the lowest FDR overall.

Comparison of Computing Time

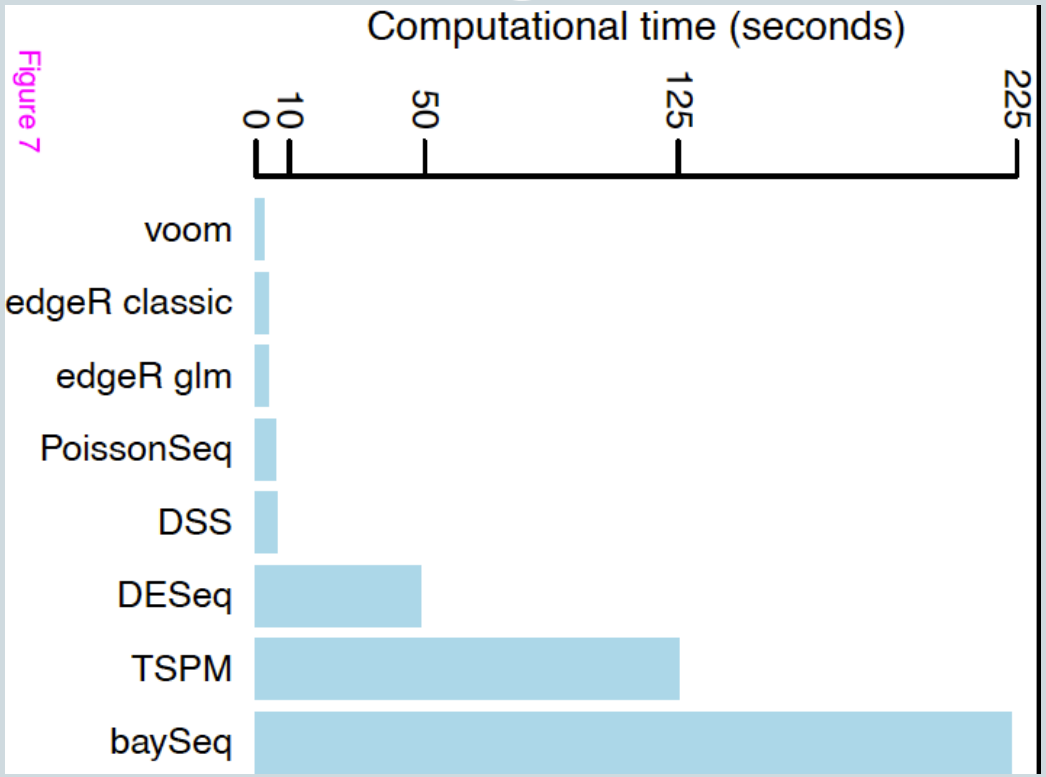


Figure 7 Computing times of RNA-seq methods. Bars show time in seconds required for the analysis of one simulated dataset on a MacBook laptop. Methods are ordered from quickest to most expensive.

Males vs. Females – Nigerian Individuals

- 16 genes up regulated in males / 43 genes up regulated in females
- 15 out of 16 top genes are from either the X or Y chromosomes
- Used ROAST and CAMERA for gene set enrichment
  - they account for inter gene correlations, so they cannot be used with NB models

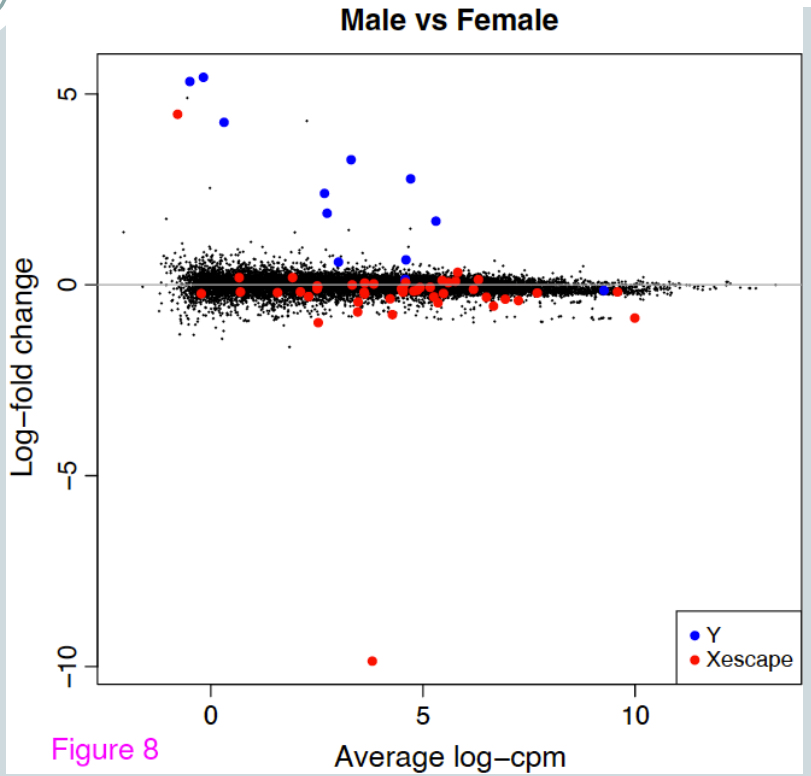


Figure 8

Figure 8 MA-plot with male and female specific genes highlighted. The log-fold change of each gene is plotted against its average log-cpm for the comparison between males and females. Genes on the male-specific region of the Y chromosome genes are highlighted blue and are consistently up-regulated in males, while genes on the X chromosome reported to escape X-inactivation are highlighted red and are generally down in males.

Time Trend Analysis – Fly Development

- 12 time points, but no replicates at any time point
- fit a quadratic trend with time
- won't go over results, but yay! they can fit a quadratic

Discussion

- Gene-Level Analysis
  - doesn't incorporate any methods for differentiating between isoforms
  - multiply aligned reads are not addressed (i.e., confidence in read estimates)
- log-cpm and log-rpkm will give the same results
- VOOM just needs to be 'as good' as the count-based methods, but it actually showed some improvements
- VOOM outperforms limma-trend when library sizes are uneven
- Ordinary t-statistics performed poorly proving that alternative models are needed.

Why is VOOM better?

1. The theoretical mean-variance relationships in either the Poisson distribution or the Negative Binomial distribution fit the observed data precisely.
2. VOOM models the mean-variance relationship to individual observations rather than at the gene-level making it more precise when library sizes are uneven.
3. Use of normal models allows us to borrow information across genes using the tractable empirical Bayes distribution theory.
4. The combination of normal distribution approximations and variance modeling is partially supported by generalized linear model theory.
5. More robust against outliers.
6. Can access many different types of models to fit study designs that are not simple two group comparisons

My Thoughts

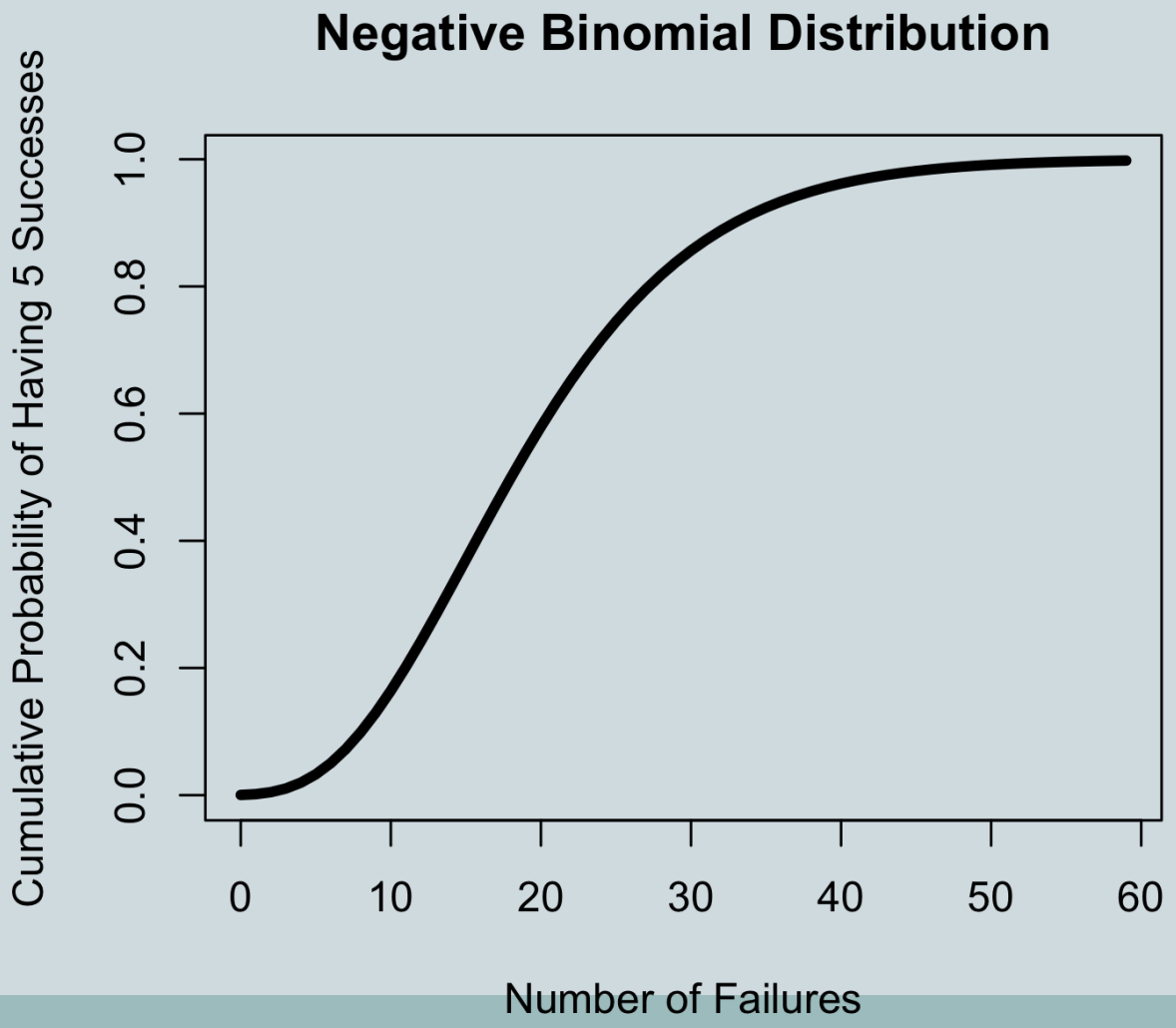
- I like LIMMA for microarrays, so I am a little bias towards this method.
- I also see lots of study designs that do not lend themselves well to anything besides simple two group comparisons.
- I am not real excited about the use of counts per million mapped reads, but I think this method allows for “tweeking” on how we want to calculate a library size, e.g., it might be easy to incorporate spike-ins.
- I LOVE that it is incorporated into an R package.
- Didn’t try it, so not endorsing it!!

Negative Binomial Distribution

- Goal – Find 5 people that have seen the movie Office Space
  - Probability of having seen Office Space is 20%

How many people will I have to ask before I find 5 people?

Negative Binomial Distribution



Methods For Differential Expression

- Can we estimate the "proportion of people who have seen Office Space" from how many failures before we have 5 success?
- Goal in RNA-Seq – identify genes expressed in different "proportions" in to populations

CuffDiff – uses a negative binomial distribution for error estimates and accounts for uncertainty in reads counts

Many methods out there with no clear WINNER!

Problem – Defining a "Failure"