RNA-seq II: Differential expression

Paul Pavlidis

STAT/BIOF/GSAT 540 2016

Outline

- Look more closely at real data
- Motivation for new differential expression methods
- Weighted regression approach ('limma-voom')
- Methods specific for count data (EdgeR and DESeq)

Properties of data sets

Study	PMID	Species	Samples	UniqueAlignedReads	ReadsPerSample	Notes
bodymap	22496456	human	19	2,197,622,796	115,664,358	Illumina Human BodyMap 2.0 -- tissue comparison
modencodefly	21179090	fly	30	2,278,788,557	75,959,619	developmental time course
modencodeworm	19181841	worm	46	1,451,119,823	31,546,083	developmental time course
yang	20363980	mouse	1	27,883,862	27,883,862	hybrid cell line, X always inactive
trapnell	20436464	mouse	4	111,376,152	27,844,038	time course
mortazavi	18516045	mouse	3	61,732,881	20,577,627	tissue comparison
cheung	20856902	human	41	834,584,950	20,355,730	HapMap - CEU
hammer	20452967	rat	8	158,178,477	19,772,310	experimental vs. control at 2 time points
bottomly	21455293	mouse	21	343,445,340	16,354,540	2 inbred mouse strains
montgomery+pickrell	20220756	human	60	886,468,054	14,774,468	HapMap - CEU+YRI
wang	18978772	human	22	223,929,919	10,178,633	tissue comparison
gilad	20009012	human	6	41,356,738	6,892,790	liver; males and femlaes
core	19056941	human	2	8,670,342	4,335,171	lung fibroblasts
katz.mouse	21057496	mouse	4	14,368,471	3,592,118	control vs. CUG-BP1 knockdown myoblasts
nagalakshmi	18451266	yeast	4	7,688,602	1,922,151	priming technique comparison
sultan	18599741	human	4	6,573,643	1,643,411	cell type comparison

Modified from http://bowtie-bio.sourceforge.net/recount/; some additions since I made this table

Case study: The gilad data set

Letter

Sex-specific and lineage-specific alternative splicing in primates

Ran Blekhman,[1,4,5] John C. Marioni,[1,4,5] Paul Zumbo,[2] Matthew Stephens,[1,3,5] and Yoav Gilad[1,5]

[1] Department of Human Genetics, University of Chicago, Chicago, Illinois 60637, USA; [2] Keck Biotechnology Laboratory, New Haven, Connecticut 06511, USA; [3] Department of Statistics, University of Chicago, Chicago, Illinois 60637, USA

Genome Res. 2010 Feb;20(2):180-9

- Six human liver samples (3M 3F)
  - Also chimp and macaque, but will not discuss here
- Illumina GAII, two lanes per sample. 35bp
- 13,000 genes detected according to authors
- 627 genes reported as “sexually dimorphic” commonly in all three species.

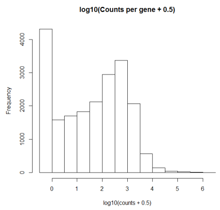
What I got via their supplement table 1: 20689 x 6 matrix with Ensembl gene IDs. (different version available through bowtie web site)

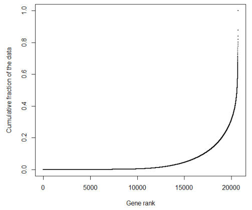
Gilad data set, cont'd

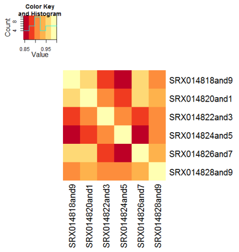
- Total read count: 20,679,864 (2.8 – 4.3 million per sample, mean=3.4 million)
- 4314 genes have 0 counts (total in 6 samples)
- 7599 have less than 10 counts total
- 196 genes have over 10000

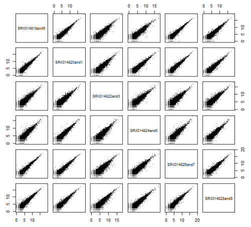
  → 11,527,345 counts for those genes (56%)

  Albumin (12%); complement, Jun, fibrinogen, serpins, APOs
- After some filtering, 10,720 genes.



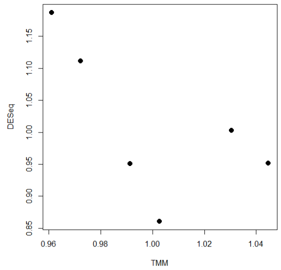






Scale factors for the Gilad data set

lib.size	TMM	DESeq	TMM, unfilt.	DESeq, unfilt.
2096011	1.031	1.002	0.99	1.00
2072827	0.991	0.951	0.92	0.95
1968729	1.045	0.951	1.15	0.963
1862868	1.003	0.866	1.02	0.858
2673491	0.961	1.18	0.92	1.185
2476156	0.972	1.11	0.99	1.115



Differential expression: Why we might need new methods

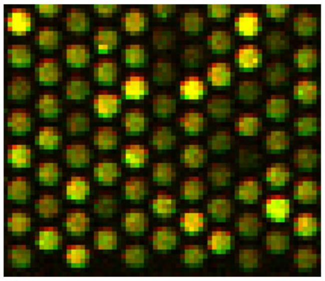
- **Goal**: accurate p-values for our hypothesis tests
- Properties relied upon for inference from *t* statistics shouldn't hold for count data.
- Perhaps most important: Heteroscedasticity
  - Strong mean-variance relationship *expected* with count data.

One challenge: Most evaluations are based on simulations. There are no widely used/accepted gold standards.

Properties of expression data: counts

Microarray

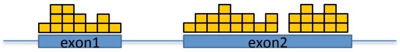
- Signal *is* fundamentally counts (deep down)
- But values are averaged across pixels and counts are high.
- Never really have zero: background ensures that values are not too small and thus "continuous"



http://www.genomics.agilent.com

Sequencing

- Unit of measurement is the read; no such thing as 0.1 read.
- Counts of reads start at 0
- As counts get high, the distinction should diminish



NOTE: We are focused on the distribution of expression values for a gene **across technical or biological replicates**

For this discussion we care less about comparing two genes **within a sample**.

Statistics of counts

- Say RNA for gene g is present “in the cell” at 1 out of 1,000,000 molecules.
  - Abundance $a = 1/1{,}000{,}000$ (1e-6)
- If we randomly pick $R_{lib} = 1{,}000{,}000$ molecules (“reads”), how many gene g RNAs will we see? (R_g)

E(R_g|R_{lib}) = _?_. But could get 0 or 5 “by chance”.

→ $R_g \sim Binomial(R_{lib}, a)$
Approximately: $R_g \sim Poisson\ (R_{lib} * a)$
As $R_{lib} * a$ gets large, approx: R_g~Normal($R_{lib} * a$,$R_{lib} * a$)

In all cases, variance is an increasing function of the mean

Options for doing differential expression on counts

Summary of the problem: Count data is expected to violate both normality and equal variance assumptions.

Possibilities for coping:

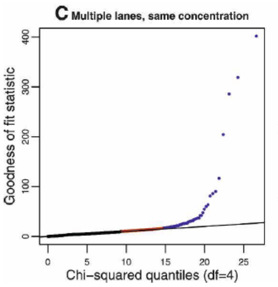
- Use a non-parametric test (e.g. SAM-seq – based on Wilcoxon; larger sample sizes needed, will not discuss further)
- Make adjustments and use standard methodology
- Use a model specific for count data

Some material from Mark Robinson (http://www.fgcz.ch/education/StatMethodsExpression/03_Count_data_analysis.pdf)

Poisson is appropriate for tech rep

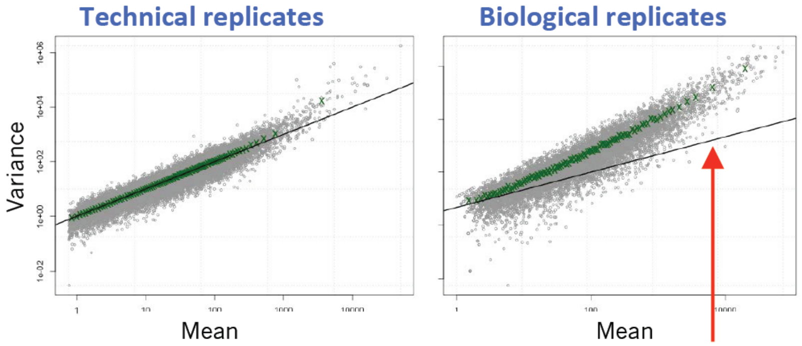
(Marioni et al.)

- Looked for “systematic differences between results for the same sample, sequenced at the same concentration in different lanes, over and above those expected from sampling error”
- Differences reasonably well explained by Poisson statistics, but does not account for biological variation (back to this later)



http://genome.cshlp.org/content/18/9/1509.long

Poisson does not capture biological variability



mean=variance
(Poisson assumption)

Data from Marioni et al. Genome Research 2008

Data from Parikh et al. *Genome Biology* 2010

http://www.fgcz.ch/education/StatMethodsExpression/03_Count_data_analysis.pdf

Impact of heteroscedasticity

- OLS: assume all errors have same variance
- If not true: higher variance regions get more weight in minimization of error than they should (since they are less precise)

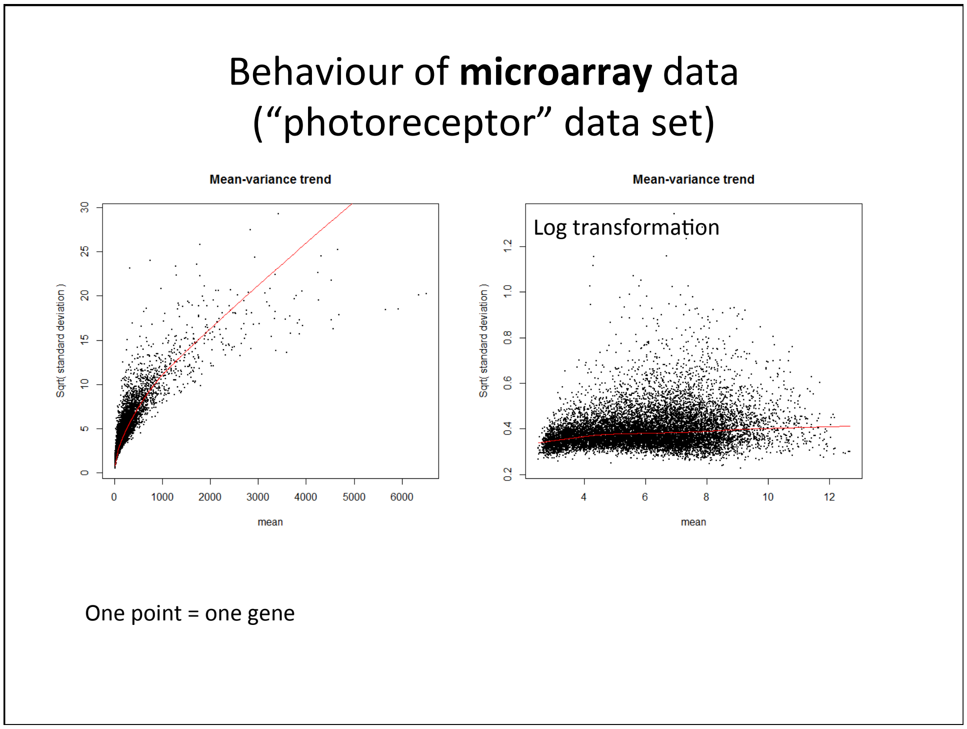
Standard errors of betas will be poor estimates

Recall: $t = \hat{\beta}/\hat{\sigma}$

… So p-values will also be wrong; In case of positive relationship, too small.

Transformation can help

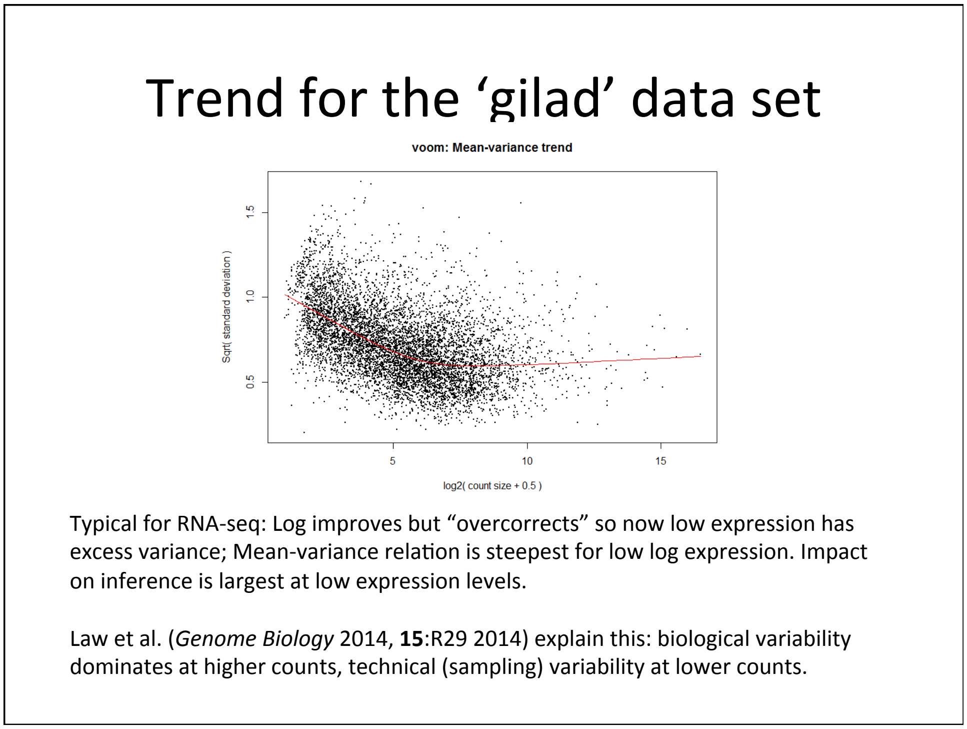
- log, square root, …
- For microarray data, taking logs is often deemed sufficient (but see “VSN” and other methods)
- None of these seem to adequately remove the trends in RNA-seq data



Behaviour of **microarray** data (“photoreceptor” data set)

Log transformation

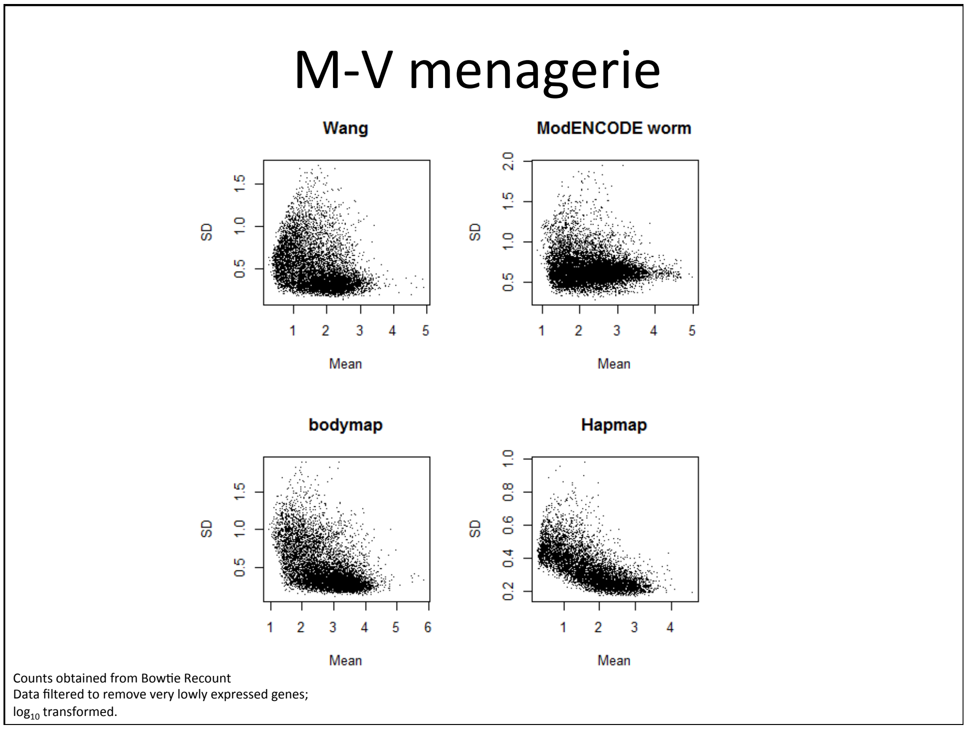
One point = one gene



Trend for the ‘gilad’ data set

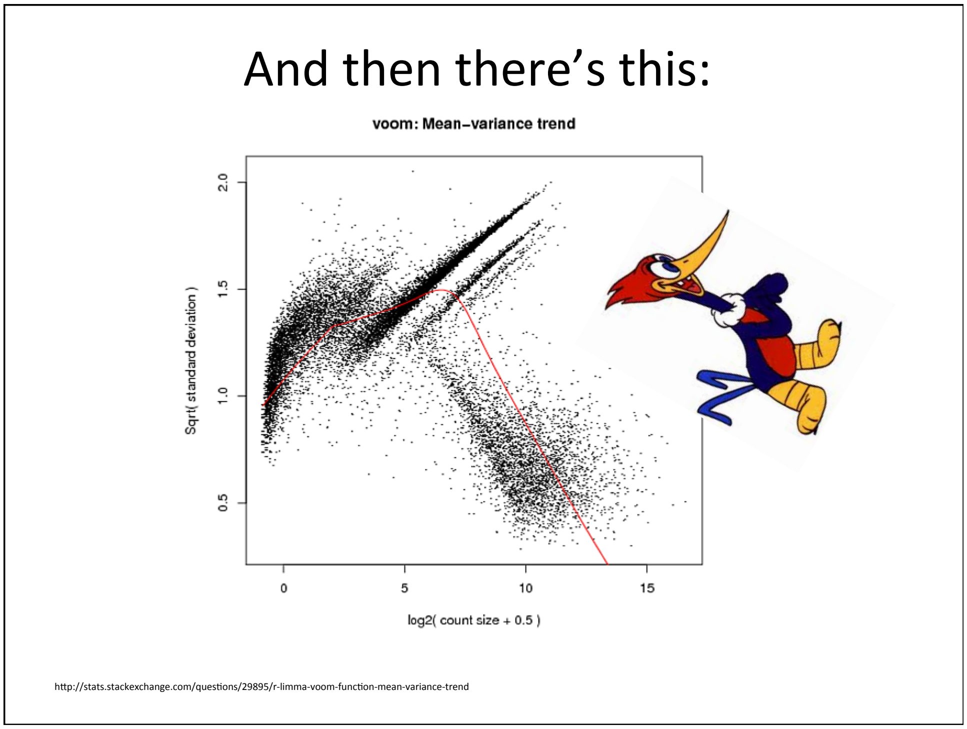
Typical for RNA-seq: Log improves but “overcorrects” so now low expression has excess variance; Mean-variance relation is steepest for low log expression. Impact on inference is largest at low expression levels.

Law et al. (*Genome Biology* 2014, **15**:R29 2014) explain this: biological variability dominates at higher counts, technical (sampling) variability at lower counts.



M-V menagerie

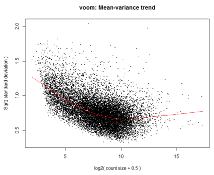
Counts obtained from Bowtie Recount
Data filtered to remove very lowly expressed genes;
$\log_{10}$ transformed.



And then there's this:

http://stats.stackexchange.com/questions/29895/r-limma-voom-function-mean-variance-trend

Voom



Transformation approach to allow use of limma.

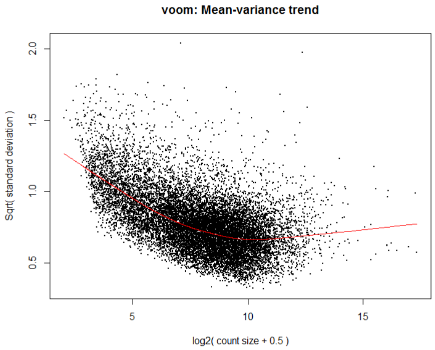
Key idea: Modeling the mean-variance relation is more important than getting the probability distribution exactly right.

Work with log2 counts per million (log-cpm)

Genome Biology 2014, **15**:R29

Rationales

- Why log transform: improves the mean-variance relationship but tends to "over-correct" so now low values are more variable than high values.
- Why quarter-root variance? Makes distribution more symmetric



Voom

"Voom is an acronym for 'mean-variance modelling at the observational level'"

1. Fit your linear model to the data ($\log_2$-transformed cpm)
2. Take the residuals. Their sqrt-stdev (quarter-root variance) per gene usually has a reasonable relationship with the mean; That is, consider

$$\hat{\mu} \sim \sqrt{\widehat{sd}(\varepsilon)}$$

3. Fit a lowess smoother to this relationship (red line in plots)
4. Use the lowess to estimate the variance for each (fitted): get weights

$$w_i = 1/\text{lowessfit}(\hat{c}_i)^4$$

where $\widehat{c}_i$ is the $\log_2$-transformed fitted cpm and lowessfit() provides the predicted sqrt-stdev.

Intuition: points where we are less sure of the actual value (higher variance) get lower weight in the analysis.

Why regress out the model first: Think of it as an iterative process. The first estimate of residuals will be "improved" by the weights computed. Those weights would be very poor estimates if the differential expression is large.

Getting observation-level estimates of variance

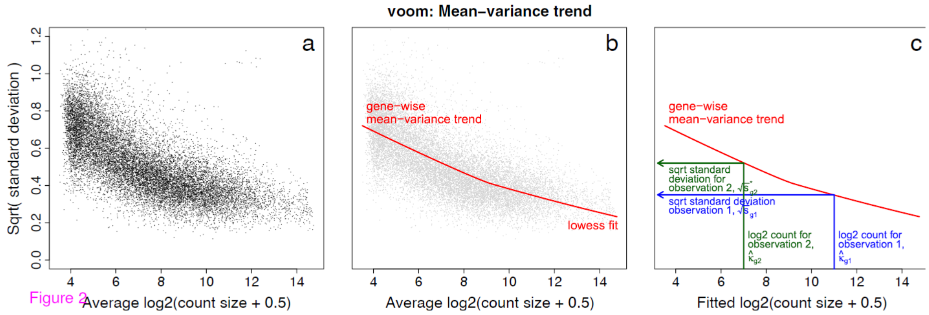


Figure 2 Voom mean-variance modeling. Panel **(a)**, gene-wise square-root residual standard deviations are plotted against average log-count. Panel **(b)**, a functional relationship between gene-wise means and variances is given by a robust lowess fit to the points. Panel **(c)**, the mean-variance trend enables each observation to map to a square-root standard deviation value using its fitted value for log-count.

Genome Biology 2014, **15**:R29

Weighted regression

R & Limma already supported weighted regression, so what it is?

Usual normal equations are

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

Modified to use weights:

$$\hat{\beta} = (X^T W X)^{-1} X^T W y$$

Where W is a diagonal matrix

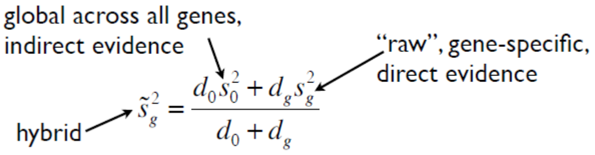
Intuition: In minimizing the residual, we want to “care less” about data points which are less precise.

$$argmin(\hat{\beta}) \sum_{i}^{n} w_i (X_i^T \hat{\beta} - y_i)^2$$

Thus the weights are expressed in terms of 1/variance.
Hard part is estimating the variance (we end up treating it as “known”)
But if values are right, assumptions of linear least squares are restored.

More about voom approach

- It does not modify the data. It only modifies the results of the lmFit call: the β values
- Residual standard error estimates are now (hopefully) better
- limma will further squeeze those:



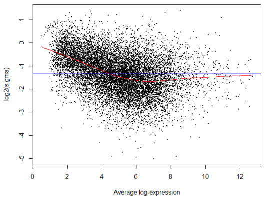
$$\tilde{s}_g^2 = \frac{d_0 s_0^2 + d_g s_g^2}{d_0 + d_g}$$

global across all genes, indirect evidence
“raw”, gene-specific, direct evidence
hybrid

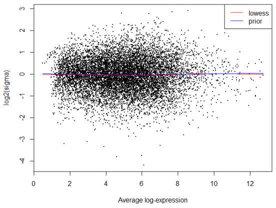
M-V plots from limma (gilad data set)

`plotSA()`

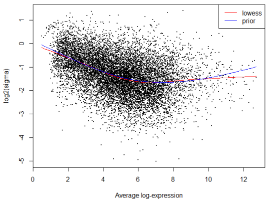
No voom



voom



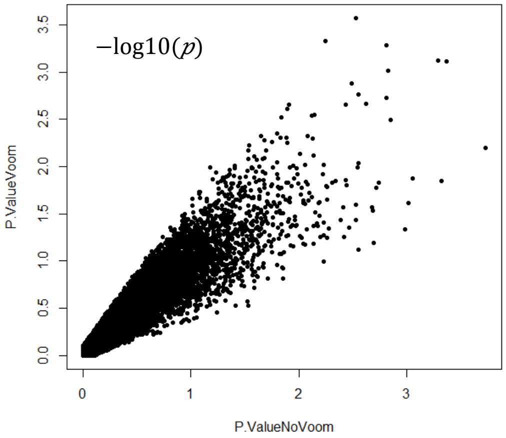
No voom, eBayes(..., trend=TRUE)



eBayes(..., trend=TRUE) should make the prior sensitive to the estimated m-v relationship.

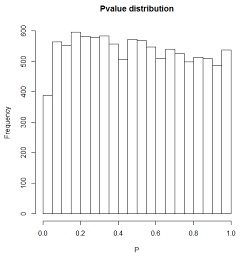
See the voom paper for details; voom worked better

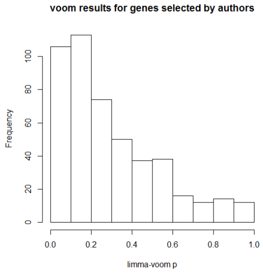
P-values with and without voom

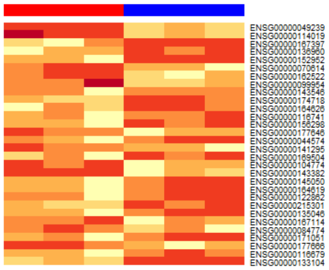


Gilad data set, limma. Your mileage may vary.

Gilad results for limma-voom







Using a model specific for counts

- Implementation: EdgeR, DESeq, baySeq, others
- Some groups used a Poisson model, but field moved to using negative binomial in a generalized linear model framework

- Originally approaches developed with SAGE in mind: small sample sizes, low "library size"

  (>1 million tags would be very unusual. 50-100k typical).
- More recently influenced by RNA-seq data.

EdgeR and DESeq

- Use negative binomial distribution.
- In addition, both try to address the mean-variance trend in special ways. How they do this is the main difference.
  - Both use NB + GLMs (and offer simpler method if you have a one-way layout)
  - Both use m-v trends to help moderate dispersion estimates.
- At best generate estimates of variance for each gene; voom does this for each observation.
- Caution: peer-reviewed explanations may be out of date, look at user manuals!

Negative binomial distribution

- A gamma mixture of Poisson distributions
  - Count sampling distribution = Poisson
  - Biological sampling means from gamma
    - i.e., distribution of replicates
- No other particular reason to use it – it's (somewhat) convenient.
- "Overdispersed Poisson"

- Has an extra parameter to estimate compared to Poisson: the dispersion.
- Key problem: Estimating the dispersion from small data sets is tricky.

Modeling using negative binomial dist.

$$\sigma^2{}_i = \mu_i(1 + \mu_i\phi_i)$$

where ϕ_i is the dispersion for gene i. With ϕ =0, get Poisson.

Could estimate directly from the data for gene *i*, but hard to trust data from small samples

Another option is to make ϕ a parametric function of the mean (e.g. quadratic). But popular methods use more flexible approach:

edgeR: ϕ is gene-specific but moderated towards a trend.

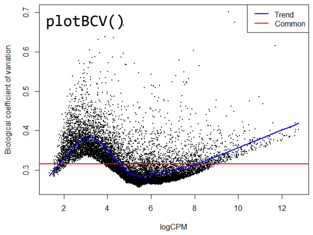
`estimateGLMTrendedDisp` – fits the trend (bin and fit spline) followed by

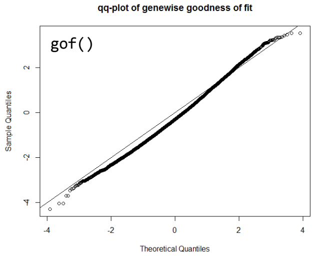
`estimateGLMTagwiseDisp` – squeezes towards the trend

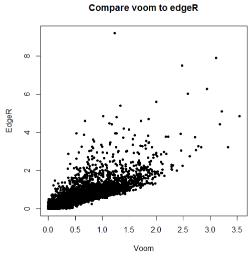
Early versions of edgeR used a common estimate and then squeezing

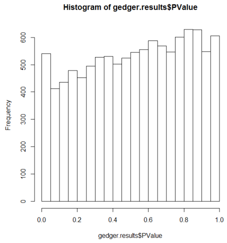
DESeq: ϕ also gene-specific: use trended value unless direct estimate is higher for gene (but offers many other options including ones that make it more like edgeR; see help for `estimateDispersions`)

EdgeR on the gilad data set

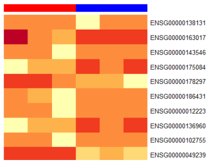




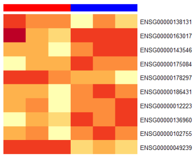




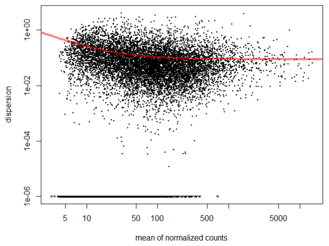
Top tags using `cpm()`

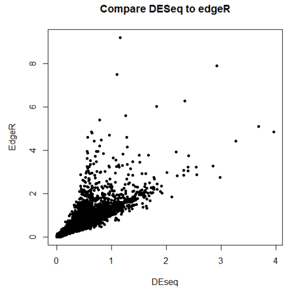


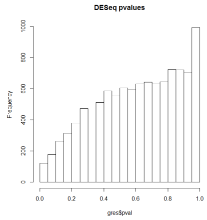
Top tags using `predFC()`



DESeq on the gilad data set







```
nbinomTest
```

To access GLMs, we would use

```
fullfit<-fitNbinomGLMs(gcds, count ~ gender)
nullfit<-fitNbinomGLMs(gcds, count ~ 1)
nbinomGLMTest(fullfit, nullfit)
```

Gives similar results here.

Summary of the differences between edgeR and DESeq

- Dispersion estimation
  - "edgeR uses moderated dispersion (towards trend)"
  - "DESeq use maximum of fitted trend and gene-wise" (conservative) – DESEq2 tries to fix this
  - "edgeR is somewhat sensitive to outliers, but DESeq suffers somewhat in power" – edgeR-robust tries to fix outlier sensitivity.
- Normalization
  - TMM -weighted trimmed mean of M-value
  - DESeq – sample-wise median ratio

Also, GLM features of DESeq are more limited than edgeR. Only provides p-values and some fit statistics; no 'toptable' and no easy facilities for accessing specific contrasts. So for complex designs edgeR is easier.

Quotes from http://www.fgcz.ch/education/StatMethodsExpression/
03_Count_data_analysis.pdf

Voom does well in simulations

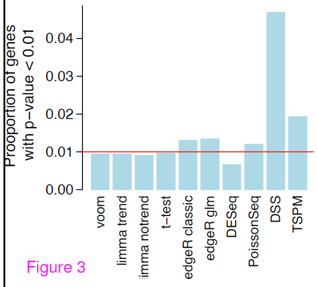


Figure 3

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.

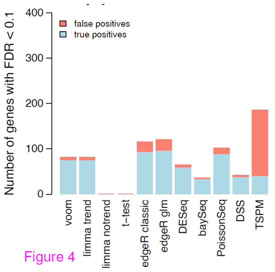


Figure 4

Figure 4 Power to detect true differential expression. Bars show the total number of genes that are detected as statistically significant (FDR < 0.1) (a) with equal library sizes and (b) with unequal library sizes. The blue segments show the number of true positives while the red segments show false positives. 200 genes are genuinely DE. Results are averaged over 100 simulations. Height of the blue bars shows empirical power. The ratio of the red to blue segments shows empirical FDR.

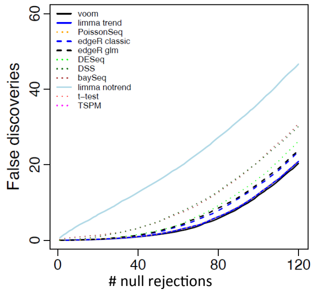


Figure 5 False discovery rates. The number of false discoveries is plotted for each method versus the number of genes selected as DE. Results are averaged over 100 simulations **(a)** with equal library sizes and **(b)** with unequal library sizes. Voom has the lowest FDR at any cutoff in either scenario.

Negative binomial model used to generate the data: should be optimal for edgeR and DESeq

Genome Biology 2014, **15**:R29

Voom has good FDRs on “spike-in” data

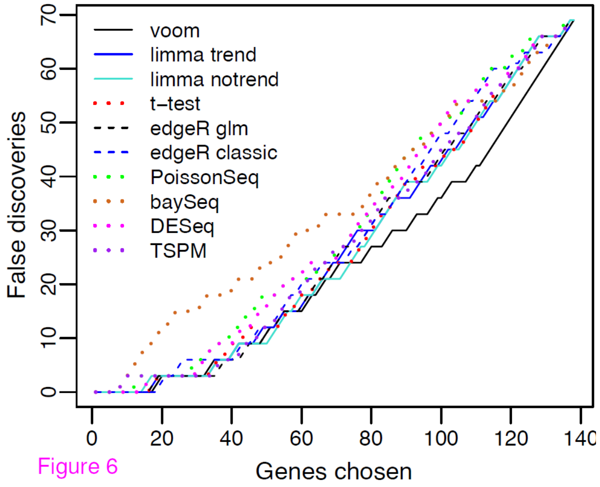


Figure 6

Spike-ins models 1.5-4-fold changes, ~6 million reads.

Only the spike-ins analyzed

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.

edgeR's latest iteration

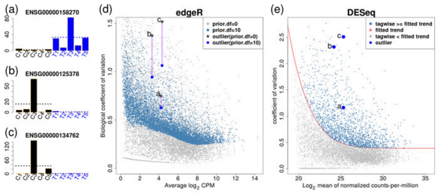
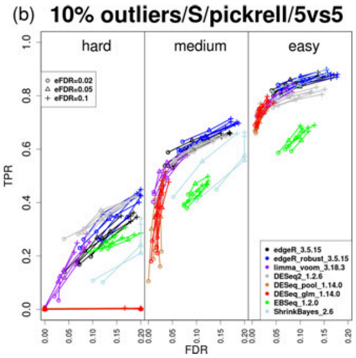


Figure 1. From Pickrell (10) data, 10 randomly selected samples from individuals are divided into two groups of 5, forming an artificial 'null' scenario. (**a**), (**b**) and (**c**) show barplots of log-counts-per-million (CPMs) of three genes from the top 10 DE genes with one or two extremely large observations. Dashed lines represent group-wise average log-CPMs. (**d**) and (**e**) plot genewise biological coefficient of variation (BCV) against gene abundance (in $\log_2$ counts per million) for edgeR and DESeq. In panel (d), gray dots show unmoderated biological BCV estimates ($\sqrt{\phi_i}\sqrt{\phi_i}$) (prior degrees of freedom = 0). Steel blue dots show moderated biological BCV with prior degree 10 (default setting for edgeR). Three outlier genes on (a), (b) and (c) are labeled by large blue dots. For (e), DESeq uses the maximum (steel blue dots) of a fitted dispersion-mean trend (red line) or the individual feature-wise (tagwise) dispersion estimate. Three outlier genes are also pointed out by large blue dots.



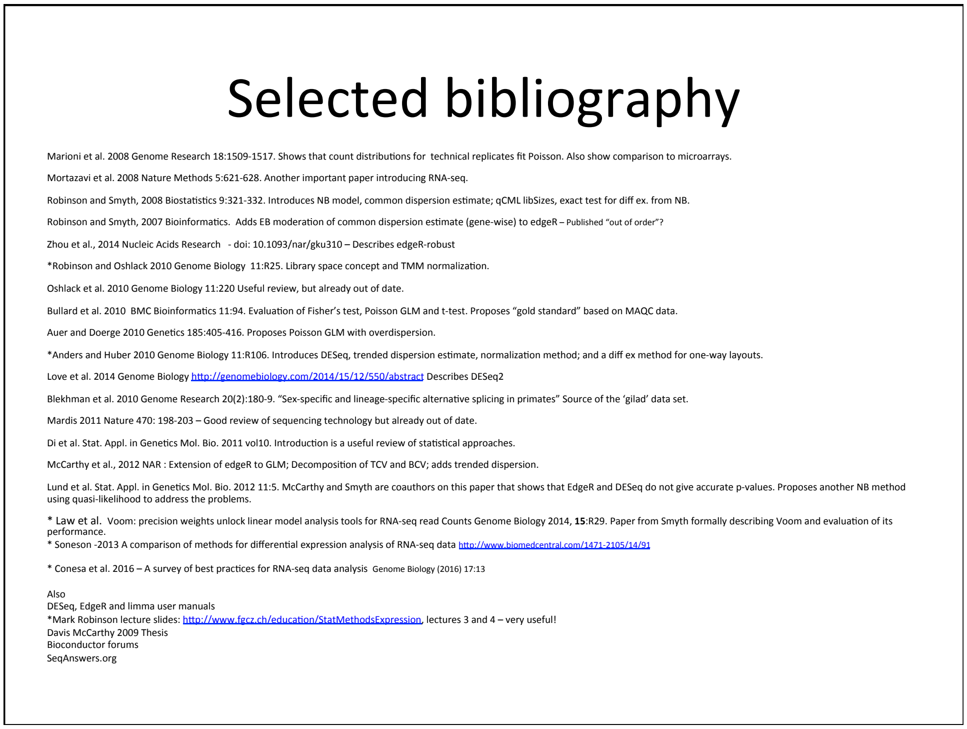
"In all cases, limma-voom controls FDR well and maintains power"

Nucleic Acids Research, 2014, Vol. 42, No. 11

How do we choose a method?

- There is no great gold standard to use. Simulations somewhat unsatisfying, spike-ins not completely realistic
- EdgeR and DESeq are very similar in design; latest versions might fix many of the issues they had last year.
- **Limma-voom** has emerged as a sound choice
  - Performs as well or better than NB (see paper for explanations why)
  - Flexible, fast
  - Familiar to limma users
  - Might not do as well well sample size is very small – but nobody should be doing N=2 experiments.

* None of the methods discussed deal specially with splice variants: See cuffdiff and DEXSeq, limma::diffSplice or quantify at transcript (or exon) level and proceed as usual.



Selected bibliography

Marioni et al. 2008 Genome Research 18:1509-1517. Shows that count distributions for technical replicates fit Poisson. Also show comparison to microarrays.

Mortazavi et al. 2008 Nature Methods 5:621-628. Another important paper introducing RNA-seq.

Robinson and Smyth, 2008 Biostatistics 9:321-332. Introduces NB model, common dispersion estimate; qCML libSizes, exact test for diff ex. from NB.

Robinson and Smyth, 2007 Bioinformatics. Adds EB moderation of common dispersion estimate (gene-wise) to edgeR – Published "out of order"?

Zhou et al., 2014 Nucleic Acids Research   - doi: 10.1093/nar/gku310 – Describes edgeR-robust

*Robinson and Oshlack 2010 Genome Biology 11:R25. Library space concept and TMM normalization.

Oshlack et al. 2010 Genome Biology 11:220 Useful review, but already out of date.

Bullard et al. 2010 BMC Bioinformatics 11:94. Evaluation of Fisher's test, Poisson GLM and t-test. Proposes "gold standard" based on MAQC data.

Auer and Doerge 2010 Genetics 185:405-416. Proposes Poisson GLM with overdispersion.

*Anders and Huber 2010 Genome Biology 11:R106. Introduces DESeq, trended dispersion estimate, normalization method; and a diff ex method for one-way layouts.

Love et al. 2014 Genome Biology http://genomebiology.com/2014/15/12/550/abstract Describes DESeq2

Blekhman et al. 2010 Genome Research 20(2):180-9. "Sex-specific and lineage-specific alternative splicing in primates" Source of the 'gilad' data set.

Mardis 2011 Nature 470: 198-203 – Good review of sequencing technology but already out of date.

Di et al. Stat. Appl. in Genetics Mol. Bio. 2011 vol10. Introduction is a useful review of statistical approaches.

McCarthy et al., 2012 NAR : Extension of edgeR to GLM; Decomposition of TCV and BCV; adds trended dispersion.

Lund et al. Stat. Appl. in Genetics Mol. Bio. 2012 11:5. McCarthy and Smyth are coauthors on this paper that shows that EdgeR and DESeq do not give accurate p-values. Proposes another NB method using quasi-likelihood to address the problems.

* Law et al. Voom: precision weights unlock linear model analysis tools for RNA-seq read Counts Genome Biology 2014, **15**:R29. Paper from Smyth formally describing Voom and evaluation of its performance.
* Soneson -2013 A comparison of methods for differential expression analysis of RNA-seq data http://www.biomedcentral.com/1471-2105/14/91

* Conesa et al. 2016 – A survey of best practices for RNA-seq data analysis Genome Biology (2016) 17:13

Also
DESeq, EdgeR and limma user manuals
*Mark Robinson lecture slides: http://www.fgcz.ch/education/StatMethodsExpression, lectures 3 and 4 – very useful!
Davis McCarthy 2009 Thesis
Bioconductor forums
SeqAnswers.org