gibbSeq: a fully Bayesian multiple testing method for differential gene expression

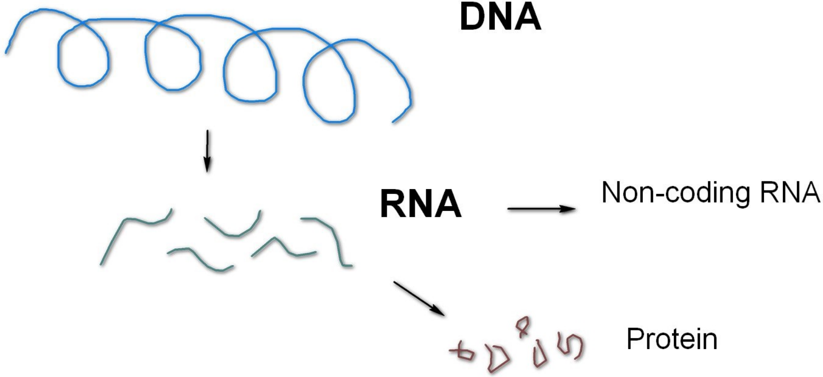
Oleg Makhnin
New Mexico Tech

June 13, 2018

gibbSeq is a fully Bayesian method for multiple testing based on hidden variables. It is primarily developed for the gene expression data, for example, RNA-seq. The method is based on lognormal distribution approximation for the RNA-seq read counts. It directly estimates the FDR (false discovery rate) for the tests. The method also allows for direct testing of differential expression of gene sets, and may help account for lack of independence among the counts.

In simulation studies, it performs really well compared to currently popular empirical Bayes methods (edgeR, DESeq), when the data are indeed lognormal; the results are more mixed but competitive when the simulated data are Negative Binomial.

“The Central Dogma”



DNA

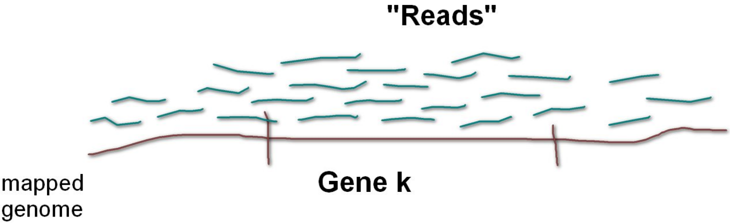
RNA

Non-coding RNA

Protein

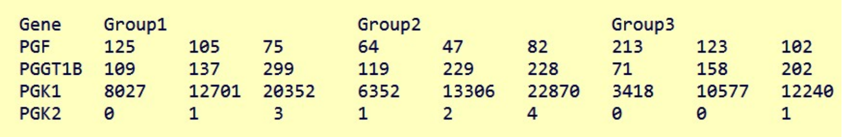
RNA-seq

Data: counts of fragments of RNA ("reads") mapped to each Gene



Quantifies *Gene expression*, i.e. a measure of activation of each Gene.

Data



Gene	Group1			Group2			Group3		
PGF	125	105	75	64	47	82	213	123	102
PGGT1B	109	137	299	119	229	228	71	158	202
PGK1	8027	12701	20352	6352	13306	22870	3418	10577	12240
PGK2	0	1	3	1	2	4	0	0	1

....

RNA-Seq data are the counts of RNA fragments that are mapped to a particular gene. As count data, they are usually modeled as Poisson, or, to account for extra variation, Negative Binomial distribution.

Most popular current methods for differenital expression (comparing counts in 2 or more groups) for RNA-seq are based on Negative Binomial distribution, pooling information across genes using empirical Bayesian methods.

`R/Bioconductor` packages `edgeR`, `baySeq`, `DESeq` ...

p- and q-values

The usual approach: declare the change (in gene *k*) statistically significant when *p*-value $< \alpha$, for a given threshold α.

$\alpha=$ false positive rate (FPR). Due to multiple testing, a correction is required.

q-value (Storey and Tibshirani, 2003) is the opposite of p-value:

$$\text{p-value} \approx \text{P(Test positive} \mid \text{no change)} = \text{False Positive Rate}$$

$$\text{q-value} = \text{P(no change} \mid \text{Test positive)} = \text{False Discovery Rate}$$

q-value may be more desirable to practitioners: ”What fraction of genes I have ‘discovered’ are bogus?”

Classical statistics

Good old two-sample t-test:

$$t = \frac{\overline{X}_1 - \overline{X}_2}{s_p \sqrt{1/n_1 + 1/n_2}}$$

The main difficulty is estimating s_p for each Gene. Sample sizes are very small (usually $n_i \leq 5$)!

Thus, the t-test is very inefficient.

Bayesian idea: pool the variance estimation across different Genes.

Model

$N_{k,i}$: read count for Gene k, sample i. Two experimental conditions A, B.

$$\begin{aligned} \log N^A_{k,i} &= \mu_k + \varepsilon^A_{k,i}, \qquad i = 1, \ldots, n_A \\ \log N^B_{k,i} &= \mu_k + D_k + \varepsilon^B_{k,i}, \qquad i = 1, \ldots, n_B \end{aligned} \tag{1}$$

where μ_k is the baseline mean for the Gene k, and D_k is the amount of *"differential expression"* for Gene k between A and B.

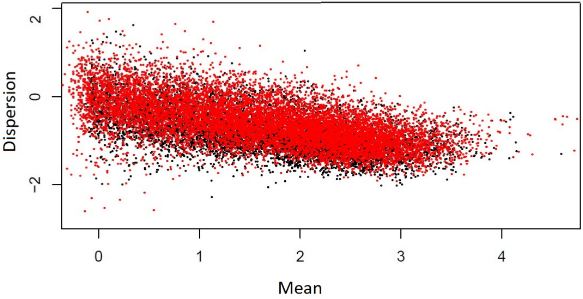
$$\begin{array}{cc} D_k \sim \texttt{Normal}(0, \tau^2) & \text{with probability } p_{DE} \\ D_k = 0 & \text{with probability } 1 - p_{DE} \end{array} \tag{2}$$

The errors $\varepsilon_{k,i}$ are Normal with mean 0 and the Gene-specific variance σ^2_k. Borrow info across Genes to better estimate σ^2_k.

Estimating variances

We may use $\sigma_k^2 \sim \mathit{Inv.Sc.}\chi^2(\mathtt{df}_{\mathtt{prior}}, \sigma_0^2)$ where $\mathtt{df}_{\mathtt{prior}}, \sigma_0^2$ are estimated from the data.

Also important: variance (or dispersion parameter) depends on the mean expression level.



after Soneson and Delorenzi, 2013

Gibbs sampler

- Based on Full Conditional Posterior (FCP) densities:

$$f(\texttt{parameter}\ j \mid \text{all other parameters})$$

For example, the FCP of σ_k^2 is *inverse scaled Chi-square* with parameters $\texttt{df} = \texttt{df}_{\texttt{prior}} + n_A + n_B$ and

$$\texttt{scale} = \frac{\texttt{df}_{\texttt{prior}} \cdot \sigma_0^2 + \sum(\log N_{k,i}^A - \mu_k)^2 + \sum(\log N_{k,i}^B - \mu_k - D_k)^2}{\texttt{df}}$$

- Draw samples from all the parameters based on their FCPs, obtain long Markov Chain Monte Carlo (MCMC) samples of all parameters involved, use the samples to find estimates.

Hidden variable method

Traditionally, after p-values are computed, they are converted into q-values (FDR) with, e.g., Benjamini-Hochberg procedure.

We can obtain them naturally while running Gibbs sampler.
Idea: introduce *hidden variables* which indicate whether the change occurred.

$$
\begin{aligned}
h_k = 1, &\quad \text{if change in Gene } k \\
h_k = 0, &\quad \text{if no change in Gene } k
\end{aligned}
$$

Our method allows for easy estimation of q-values:

$$q_k = 1 - p_k^{DE} \approx 1 - \#\{h_{k,m} = 1, m = 1, ..., M\}/M$$

where $h_{k,m}$ is the mth sample from the hidden variable h_k.
M is the MCMC sample size.

Computation

The average value of h_k is used as an estimate of the posterior probability p_k^{DE} that the Gene k is differentially expressed. No multiple testing correction is necessary!

To obtain a test with estimated genomewise FDR (false discovery rate) below q_0, declare Gene k differentially expressed iff $q_k = 1 - p_k^{DE} < q_0$.

This test may be overly conservative since actual q_k could be much lower than q_0 threshold.

Adjustment: pick q_0 but let $\widehat{FDR} = mean\{q_k : q_k < q_0\}$

Simulation studies

Several simulation scenarios were run, some are based on Negative Binomial scenario in [Hardcastle and Kelly, 2010], some on our lognormal model. Parameters are designed to mimic real RNA-seq datasets.

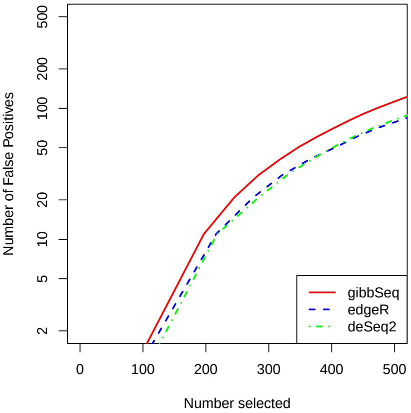
Number of genes $N = 5000$ (too small, but helps obtain more independent runs faster), sample sizes $n_A = n_B = 2, 3, 5$ percentage of positives $p_{DE} = 0.1, 0.3, 0.5$.

Run times: Gibbs sampler employed in `gibbSeq` is expensive, it takes a few minutes on a standard desktop computer even for $N = 5000$. `DEseq` and `edgeR` run faster.

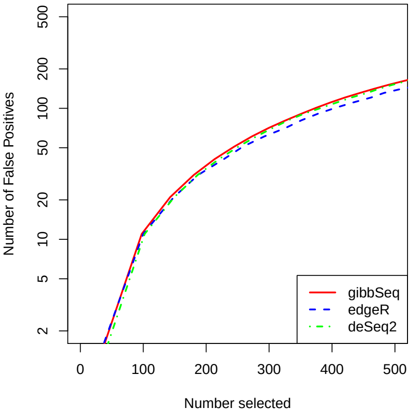
We compare ROC curves (with varying FDR thresholds), and how precisely FDR is estimated by the methods.

Negative Binomial ROC:

$n = 5, p_{DE} = 0.5$

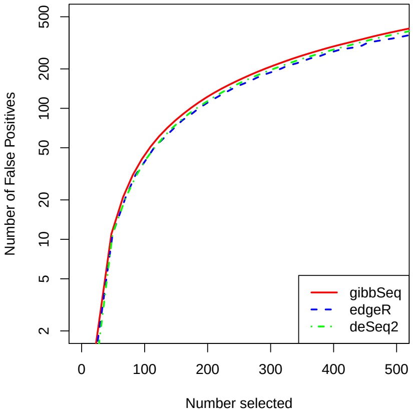


$n = 2, p_{DE} = 0.5$

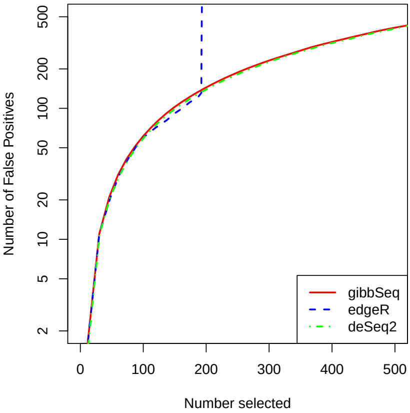


Negative Binomial ROC:

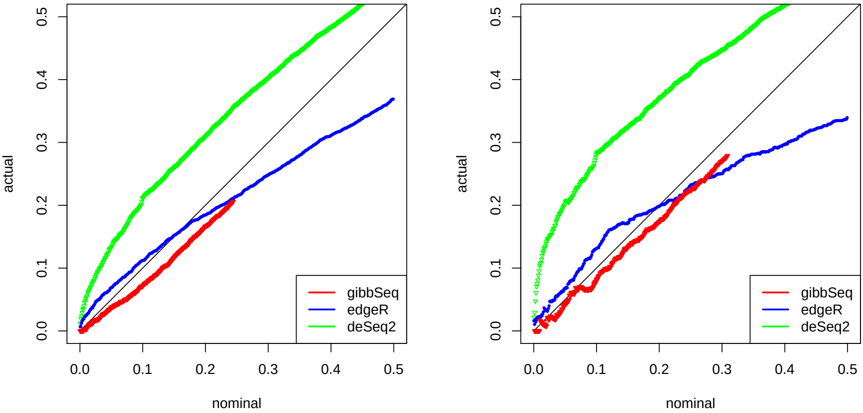
$n = 5, p_{DE} = 0.1$



$n = 3, p_{DE} = 0.1$



Negative Binomial FDR:

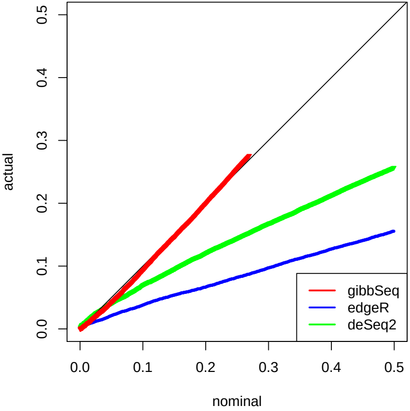


$n = 5, p_{DE} = 0.1$

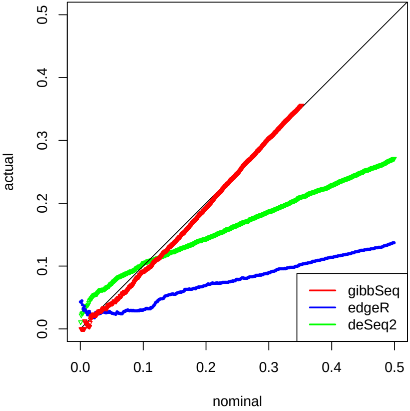
$n = 3, p_{DE} = 0.1$

Negative Binomial FDR:

$n = 5, p_{DE} = 0.5$

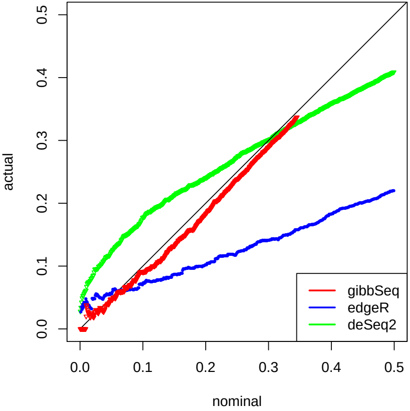


$n = 2, p_{DE} = 0.5$

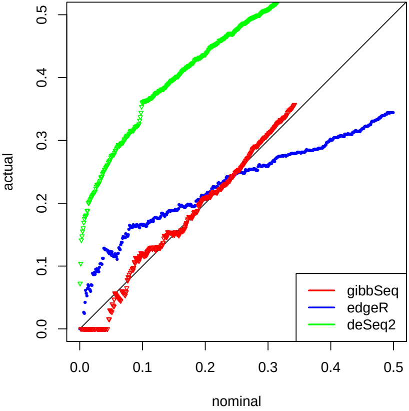


Negative Binomial FDR:

$n = 2, p_{DE} = 0.3$

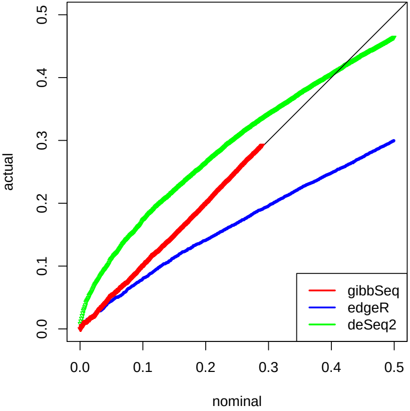


$n = 2, p_{DE} = 0.1$

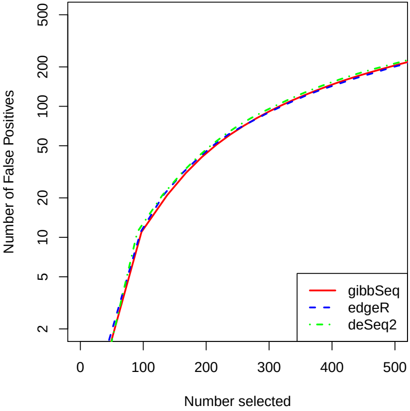


LogNormal:

$n = 3, p_{DE} = 0.3$ FDR

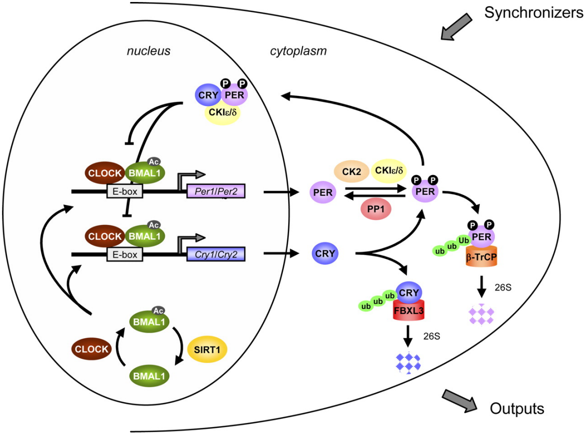


ROC



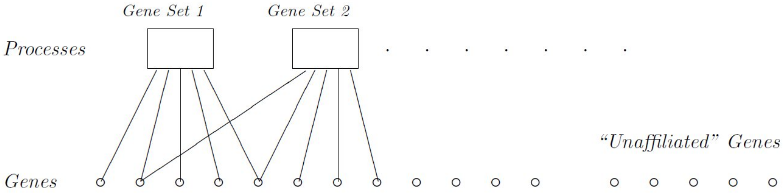
Next steps

1) **Correlated data.** Genes can be "co-expressed", for example, as a part of the same *biological pathway*.



Next steps

2) **Multi-level inference.** Researchers are interested not only in single genes, but in **gene sets** (e.g. which pathways are activated).



Main idea: together with p_{DE} for all Genes, estimate p^j_{DE} for each Gene Set j. Declare Gene Set j to be diff. expressed if $p^j_{DE} > p_{DE}$ consistently across samples.

Example: NRAP data [Peter Guerra, Rebecca Reiss].

Instead of Genes, we look at bacterial Species.

Kingdom: Bacteria

Phylum: Firmicutes

Class: Clostridia

Order: Clostridiales

Family: Lachnospiraceae

Genus: Anaerostipes

Species: Anaerostipes hadrus

Compare the abundance of species (and Genera, Families etc.) before and after remediation.

Conclusions

Our method is based on full Bayesian inference (MCMC) and is potentially more flexible in modeling gene expression. Also, it enables a straightforward calculation of false discovery rate (FDR).

Even in cases of small counts when normal approximation does not hold, our method can still outperform the established methods.

Bibliography

- Storey, Tibshirani (2003) *Statistical significance for genome-wide studies.* PNAS, 100: 9440-9445

- Robinson, McCarthy and Smyth (2010) EDGER*: a Bioconductor package for differential expression analysis of digital gene expression data.* Bioinformatics, 26(1): 139-140

- Hardcastle and Kelly (2010) BAYSEQ*: Empirical Bayesian methods for identifying differential expression in sequence count data.* BMC Bioinformatics, 11: 422

- Love, Huber and Anders (2014). *Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.* Genome Biology, 15: 550

Acknowledgements

- Daniel Acheampong (NMT)
- Rebecca Reiss (NMT)
- Sam Hokin (NCGR)
- SAMSI 2014-15 Program in Bioinformatics

QUESTIONS?

THANK YOU!

see

`www.nmt.edu/~olegm/talks/JRC2018`