Multiple testing for genetics applications: the next step

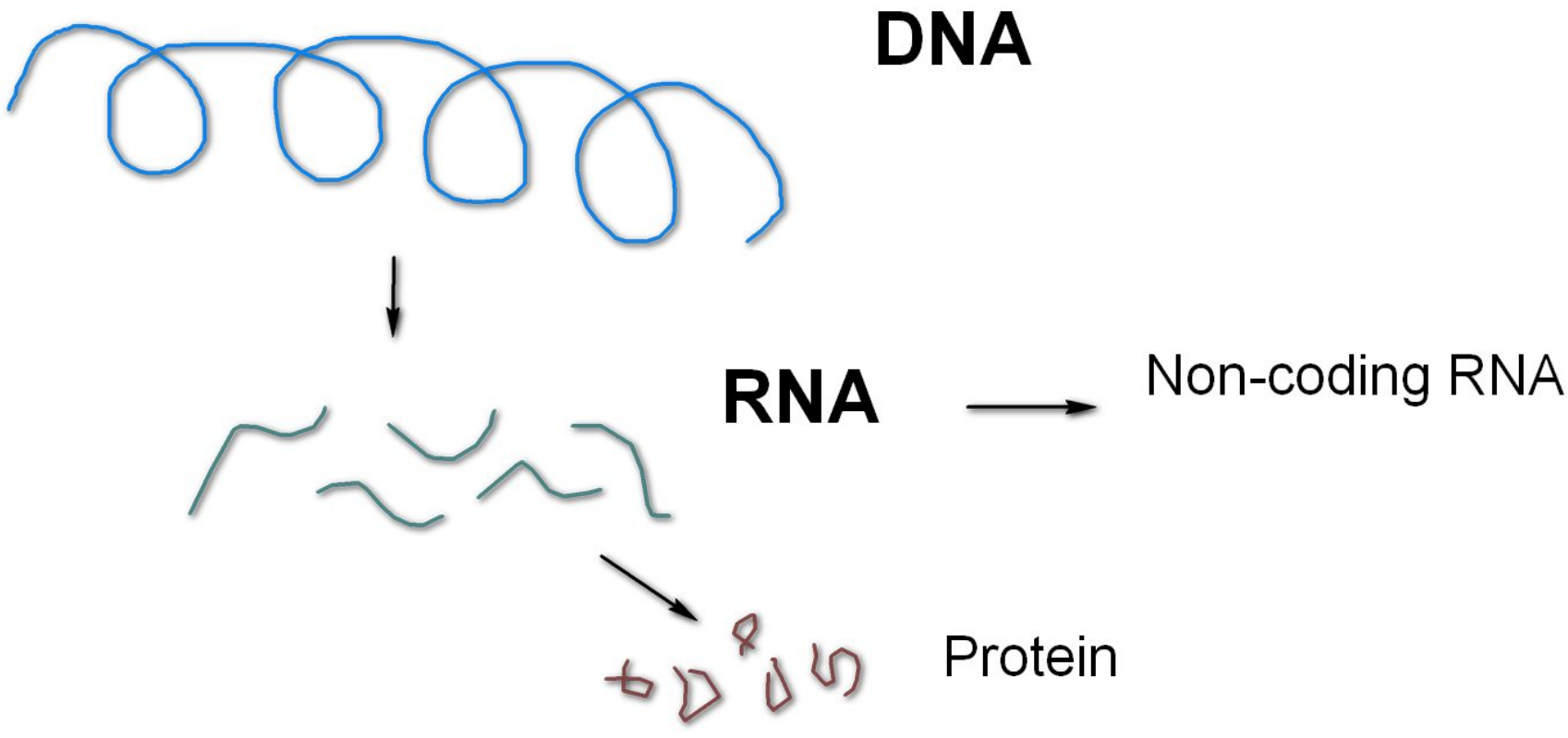
Oleg Makhnin
New Mexico Tech

September 22, 2017

Bioinformatics applications (microarrays, RNA-seq) require simultaneous testing of multiple quantitative traits. For example, in gene expression studies, two or more groups are compared, and we wish to identify which genes are expressed differently in these groups.

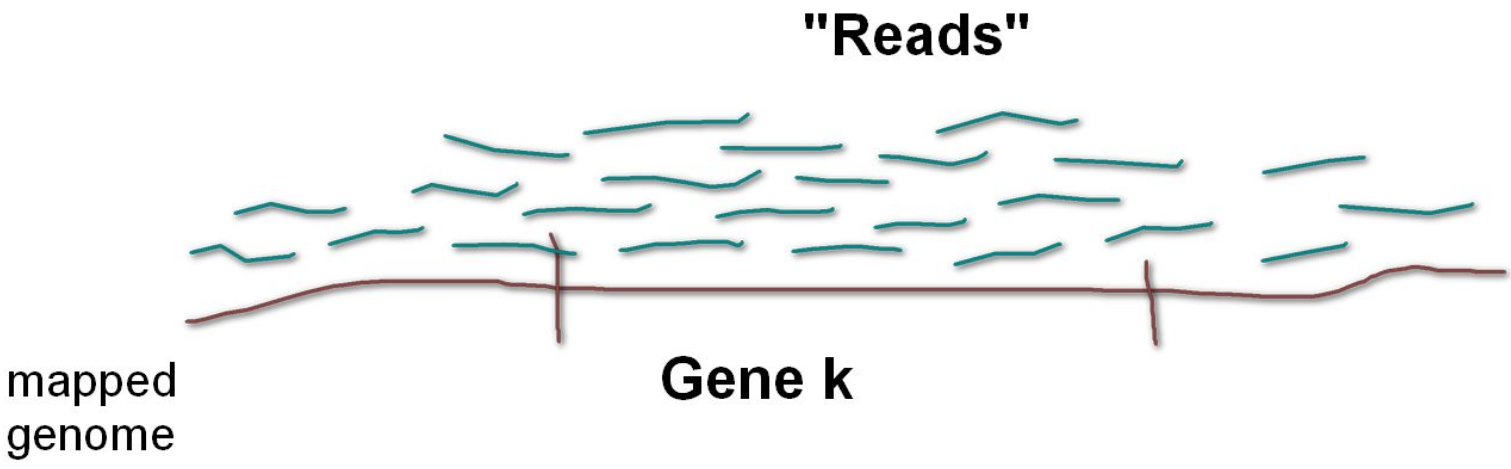
Based on a Bayesian framework developed earlier, I propose extensions for testing with correlated data, and for multi-level testing. Additionally, I present some opportunities for students to get involved.

Dogma



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 diff.exp. in RNA-seq are based on Negative Binomial distribution, pooling information across genes using empirical Bayesian methods.

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

Multiple testing

https://xkcd.com/882/

If the significance cutoff for the p-value is $\alpha = 0.05$ then 1 out of 20 results will be False Positive. If we are testing 20,000 genes then how many results will be FP?

An easy way to deal with it: Bonferroni correction. If there are n tests to run, use a p-value cutoff of α/n for each single test, then we will get only α probability of a False Positive.

$$\alpha\,(\texttt{False Positive Rate}) \approx \frac{\texttt{number of False Positives}}{\texttt{total number of genes}}$$

very inefficient!

Multiple testing ....

$$\alpha \text{ (False Positive Rate)} \approx \frac{\texttt{number of False Positives}}{\texttt{total number of genes}}$$

But this is too strict. Researchers would not mind a false positive every now and then. Therefore, use False Discovery Rate (FDR), or q-value, instead of p-value. 5% FDR will mean 1 out of 20 genes I found is expected to be False Positive.

$$FDR = \frac{\texttt{number of False Positives}}{\text{total number of } \textbf{positives}}$$

P- and Q-values

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

α= 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?"

Model

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

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

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.

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

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.

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 a long Monte Carlo sample of all parameters involved, use the samples to find estimates.

- Important to get computationally feasible and efficient algorithms!

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.

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

Bayesian Markov Chain Monte Carlo computation through Gibbs sampler produces samples of all unknown variables and parameters.
Our method allows for easy estimation of q-values:

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

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

Computation

Our method fits the framework of Bayesian model-based inference. The model is fitted using MCMC with Gibbs sampler with conjugate priors. 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 $p_k^{DE} < q_0$.

This test may be overly conservative since actual p_k^{DE} may be much lower than q_0 threshold.

Adjustment: pick q_0 but let $\widehat{FDR} = mean\{p_k^{DE} : p_k^{DE} < q_0\}$

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.

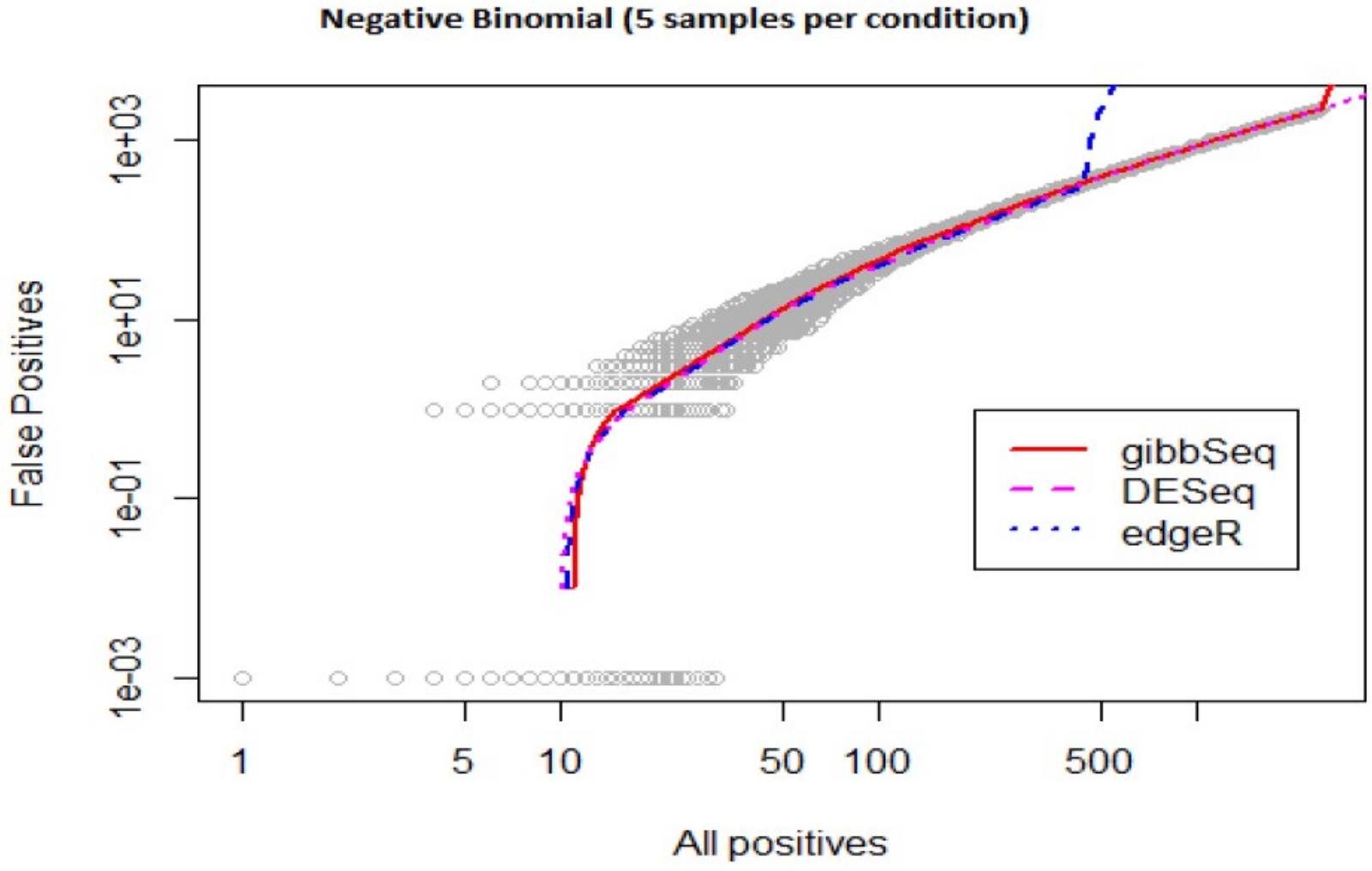
Number of genes $N = 1000$ (too small, but helps obtain more independent runs faster), percentage of positives $p = 0.1$.

Run times: Gibbs sampler employed in `gibbSeq` is expensive, it takes minutes on a standard desktop computer even for $N = 1000$. `DEseq` and `edgeR` run faster.

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

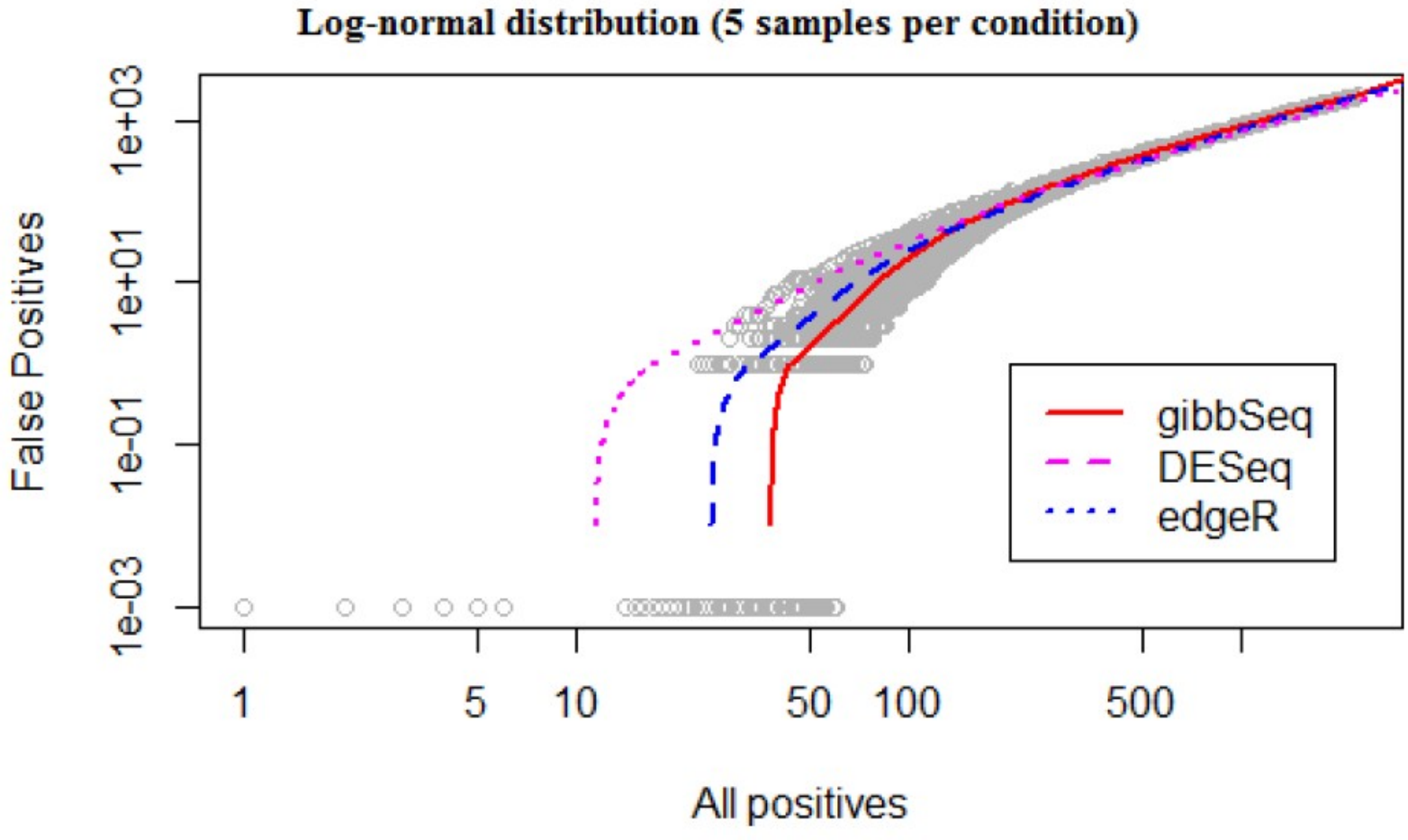
ROC curves

All methods behave similarly when the data are Neg.Bin. distributed



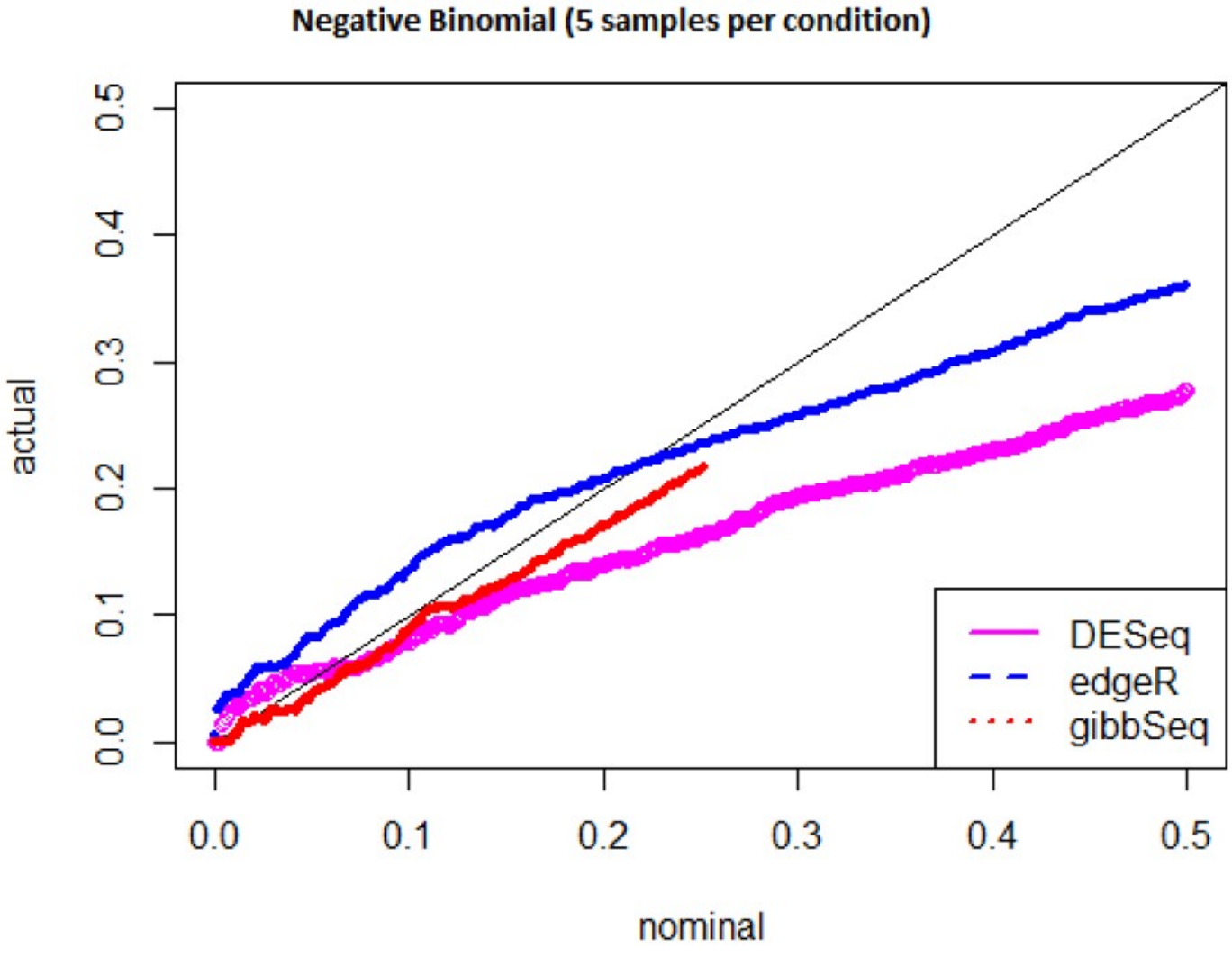
ROC curves

gibbSeq performs better when the data are Lognormally distributed.



checking FDR

Neg. Bin. distribution: `gibbSeq` estimates actual FDR fairly well. `DEseq` gives mixed results. `edgeR` underestimates FDR.

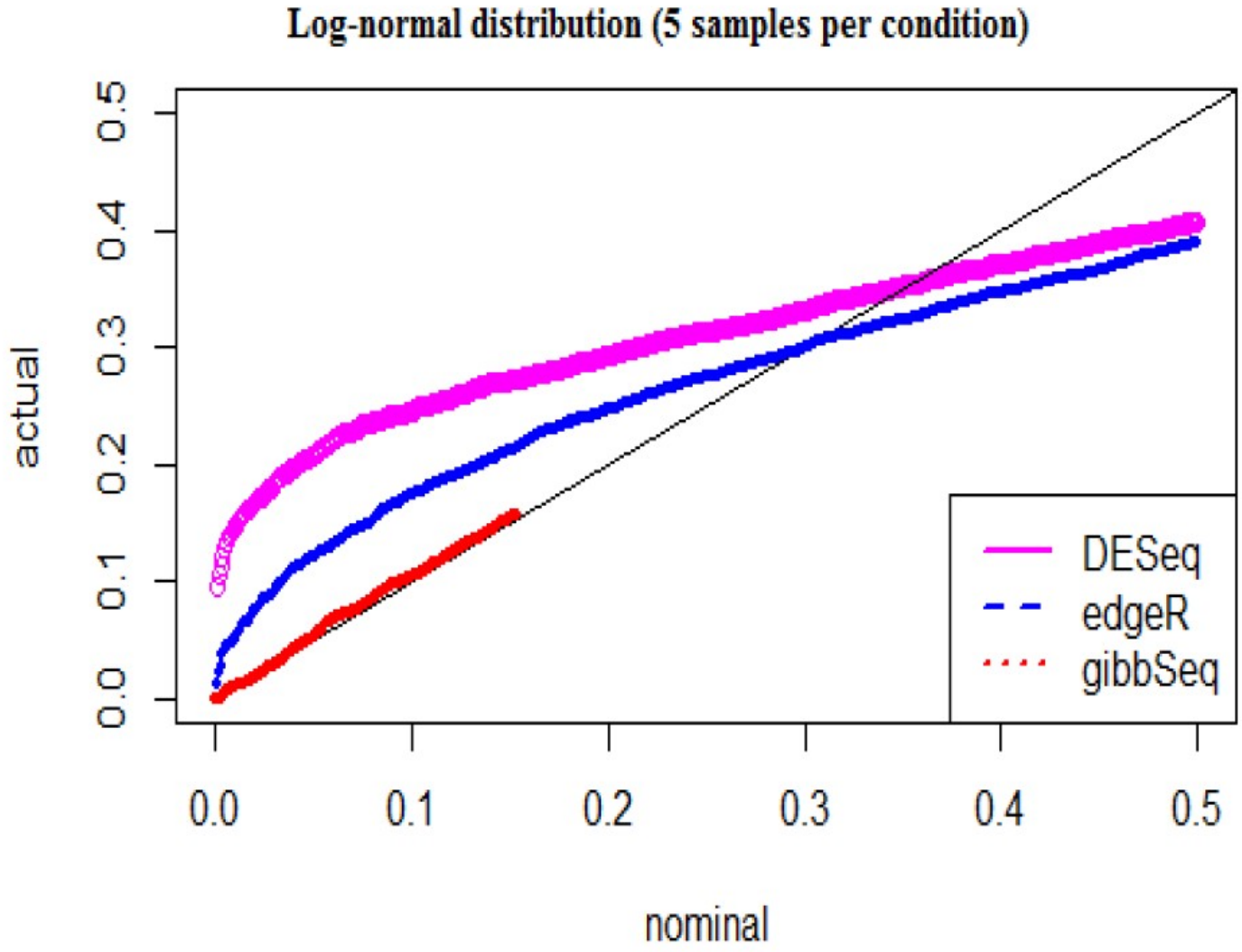


checking FDR

Lognormal distribution:

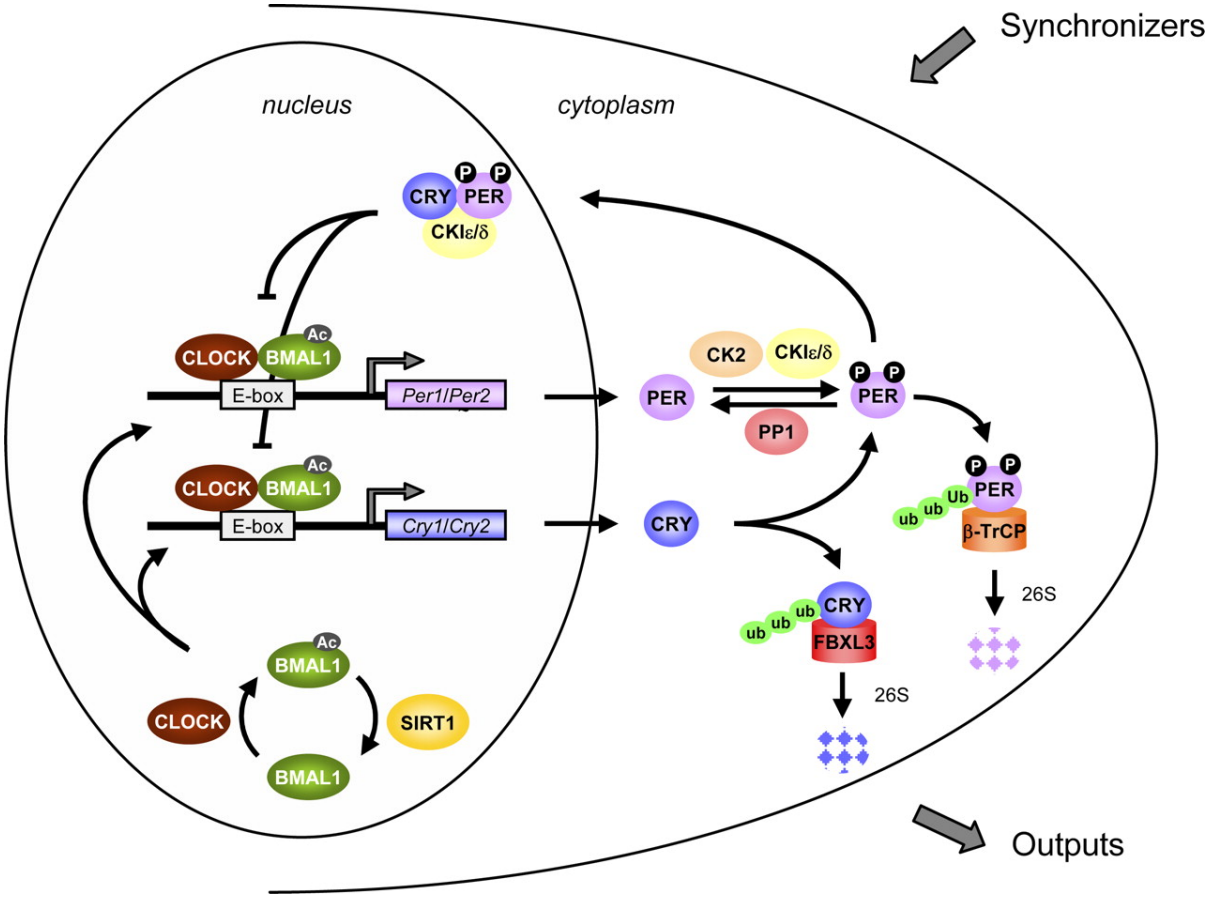
`gibbSeq` estimates actual FDR almost perfectly.

`DEseq` and `edgeR` both underestimate FDR.



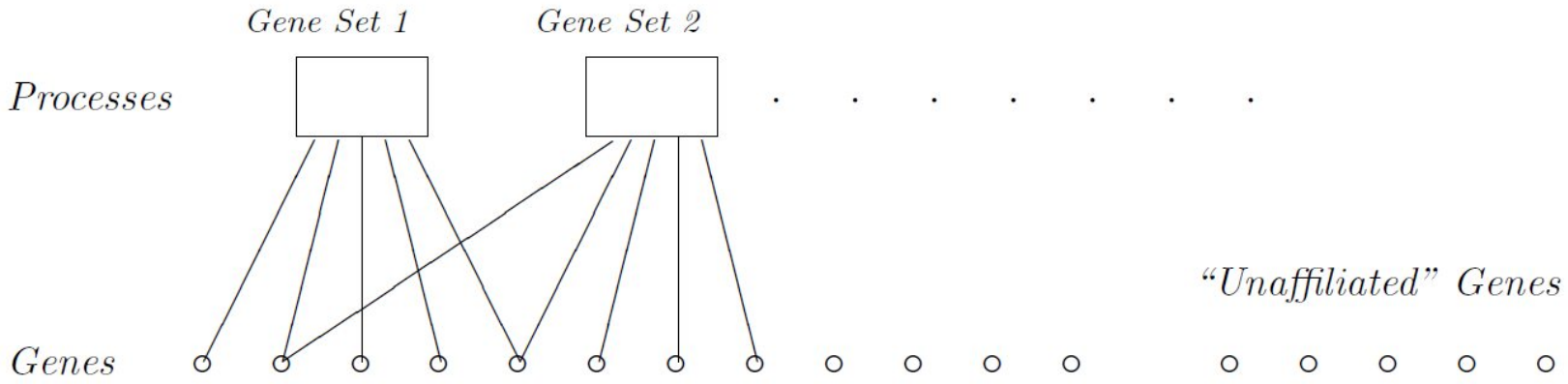
Next steps

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.

More work is needed to evolve the method: help wanted!

Skills needed: R programming, Bayesian inference.

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 2010, 11:422

QUESTIONS?

THANK YOU!

see www.nmt.edu/~olegm/talks/Gibb2