# Multiple testing for genomics applications with Negative Binomial distribution

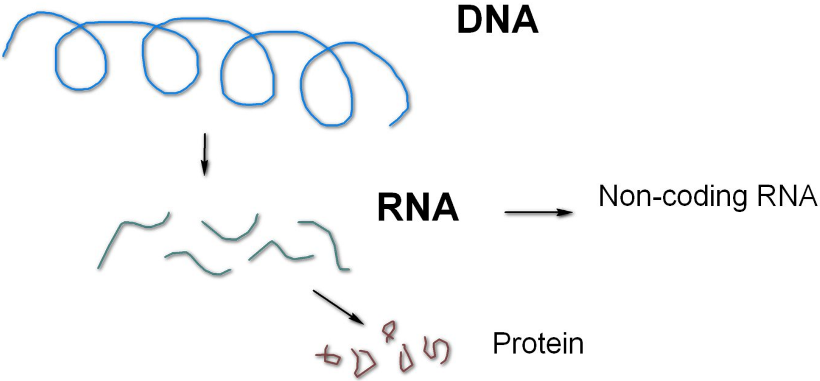
Oleg Makhnin
New Mexico Tech

September 6, 2019

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.

In this work, I develop a full Bayesian testing procedure based on hidden variables. Instead of previously used Lognormal distribution, the Negative Binomial distribution will be used.

# Central Dogma



DNA

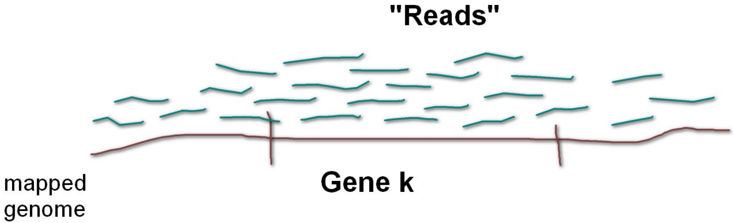
RNA

Non-coding RNA

Protein

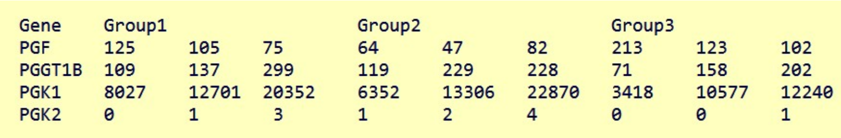
## RNA-seq

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 |

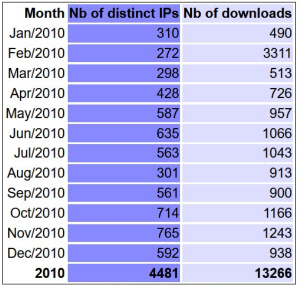
.......

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 (NB) distribution.

Most popular current methods for diff.exp. in RNA-seq are based on NB distribution, pooling information across genes using empirical Bayesian methods.

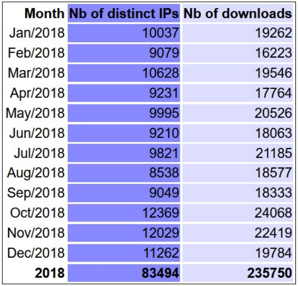
`R/Bioconductor` packages `edgeR, baySeq, DEseq` ...

# edgeR and DEseq2 history



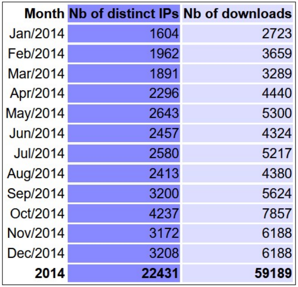
| Month | Nb of distinct IPs | Nb of downloads |
|---|---|---|
| Jan/2010 | 310 | 490 |
| Feb/2010 | 272 | 3311 |
| Mar/2010 | 298 | 513 |
| Apr/2010 | 428 | 726 |
| May/2010 | 587 | 957 |
| Jun/2010 | 635 | 1066 |
| Jul/2010 | 563 | 1043 |
| Aug/2010 | 301 | 913 |
| Sep/2010 | 561 | 900 |
| Oct/2010 | 714 | 1166 |
| Nov/2010 | 765 | 1243 |
| Dec/2010 | 592 | 938 |
| **2010** | **4481** | **13266** |

edgeR_2010_stats.tab



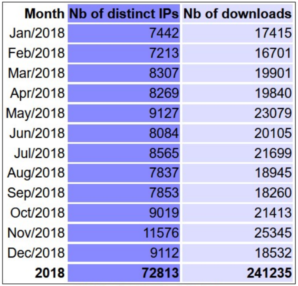
| Month | Nb of distinct IPs | Nb of downloads |
|---|---|---|
| Jan/2018 | 10037 | 19262 |
| Feb/2018 | 9079 | 16223 |
| Mar/2018 | 10628 | 19546 |
| Apr/2018 | 9231 | 17764 |
| May/2018 | 9995 | 20526 |
| Jun/2018 | 9210 | 18063 |
| Jul/2018 | 9821 | 21185 |
| Aug/2018 | 8538 | 18577 |
| Sep/2018 | 9049 | 18333 |
| Oct/2018 | 12369 | 24068 |
| Nov/2018 | 12029 | 22419 |
| Dec/2018 | 11262 | 19784 |
| **2018** | **83494** | **235750** |

edgeR_2018_stats.tab



| Month | Nb of distinct IPs | Nb of downloads |
|---|---|---|
| Jan/2014 | 1604 | 2723 |
| Feb/2014 | 1962 | 3659 |
| Mar/2014 | 1891 | 3289 |
| Apr/2014 | 2296 | 4440 |
| May/2014 | 2643 | 5300 |
| Jun/2014 | 2457 | 4324 |
| Jul/2014 | 2580 | 5217 |
| Aug/2014 | 2413 | 4380 |
| Sep/2014 | 3200 | 5624 |
| Oct/2014 | 4237 | 7857 |
| Nov/2014 | 3172 | 6188 |
| Dec/2014 | 3208 | 6188 |
| **2014** | **22431** | **59189** |

DESeq2_2014_stats.tab



| Month | Nb of distinct IPs | Nb of downloads |
|---|---|---|
| Jan/2018 | 7442 | 17415 |
| Feb/2018 | 7213 | 16701 |
| Mar/2018 | 8307 | 19901 |
| Apr/2018 | 8269 | 19840 |
| May/2018 | 9127 | 23079 |
| Jun/2018 | 8084 | 20105 |
| Jul/2018 | 8565 | 21699 |
| Aug/2018 | 7837 | 18945 |
| Sep/2018 | 7853 | 18260 |
| Oct/2018 | 9019 | 21413 |
| Nov/2018 | 11576 | 25345 |
| Dec/2018 | 9112 | 18532 |
| **2018** | **72813** | **241235** |

DESeq2_2018_stats.tab

# Negative Binomial (NB) distribution

$$P(X = u) = \frac{\Gamma(u+r)}{\Gamma(r)u!} (1-p)^u \, p^r \tag{1}$$

In the limit, as $r \to \infty$, we will get Poisson distribution.

NB distribution has some nice properties: for example, if $X_1, X_2, ..., X_n$ are independent NB with parameters $(p, r)$ then $Y = X_1 + X_2 + ... + X_n$ is also NB with parameters $(p, nr)$.

We will use a different parameterization though: $(m, r)$ with $m = \frac{(1-p)r}{p}$. It will lead to a less correlated Gibbs Sampler (below).

$m$ is the mean parameter, and $r$ is the dispersion parameter.

# 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 $\alpha/n$ for each single test, then we will get only $\alpha$ probability of a False Positive.

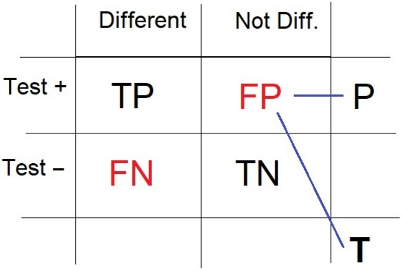
$$\alpha\;(\texttt{False Positive Rate}) \approx \frac{\texttt{number of False Positives}}{\texttt{total number of genes}}$$

very inefficient!

# .....Multiple testing

False Positive Rate

$$\alpha \approx \frac{FP}{T}$$



| | Different | Not Diff. | |
|---|---|---|---|
| Test + | TP | FP | P |
| Test – | FN | TN | |
| | | | T |

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.

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

## Model

Two groups $X$ and $Y$, for Gene $k$,

$$\begin{aligned} X_{k,i} &\sim NB(m_k, r_k) &\quad \text{for } i = 1, .., n_A \\ Y_{k,j} &\sim NB(m_k D_k, r_k) &\quad \text{for } j = 1, .., n_B \end{aligned} \tag{2}$$

Borrow info across Genes to better estimate dispersion $r_k$.
$D_k > 0$ is the ratio of *"differential expression"* for Gene $k$ between groups.
The prior densities of $D_k$ are

$$\begin{aligned} \pi(D_k) &= (\gamma - 1) D_k^{-\gamma}, &\quad D_k > 1 \\ \pi(D_k) &= (\gamma - 1) D_k^{\gamma - 2}, &\quad 0 < D_k < 1 \end{aligned} \tag{3}$$

where $\gamma > 1$ for integrability ("proper prior")

# Hidden variable method

Idea: introduce *hidden variables* which indicate whether the change occurred.

$$h_k = \begin{cases} 1, & \text{if } D_k > 1 & \text{up} \\ 0, & \text{if } D_k = 1 & \text{no change} \\ -1, & \text{if } 0 < D_k < 1 & \text{down} \end{cases} \tag{4}$$

with prior probabilities $p_-$, $p_+$ and $p_0 = 1 - p_- - p_+$.
Bayesian Markov Chain Monte Carlo computation through Gibbs sampler produces samples of all unknown variables and parameters.

Our method allows an easy estimation of q-values:

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

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

where $h_{k,m}$ is the $m$th sample from the hidden variable $h_k$.

$M$ is the Monte Carlo sample size.

## Gibbs sampler

- Based on Full Conditional Posterior (FCP) densities:

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

For example, the FCP of $r_k$ is

$$\ln f(r_k \mid \ldots) = const + \sum_i \ln \Gamma(X_{k,i} + r_k) + n_A r_k \ln r_k -$$

$$-(n_A r_k + \sum_i X_{k,i}) \ln(r_k + m_k) + similar\ terms\ with\ Y_{k,j}$$

This is not any known density, but Metropolis algorithm allows sampling, no need for proportionality constant.

## ...Gibbs sampler

• 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.

• The method hangs on our ability to integrate out $D_k$ to get the FCP of $h_k$.

• Important to get computationally feasible and efficient algorithms!

# Computation

Our method fits the framework of Bayesian model-based inference. The model is fitted using MCMC with the Gibbs sampler. The sample proportion of $h_k$ is used to estimate the posterior probabilities $p_k^+$ and $p_k^+$. Then let $p_k^{DE} = \max\{p_k^+, p_k^-\}$.

To obtain a test with estimated genomewise FDR (false discovery rate) below $q_0$, declare

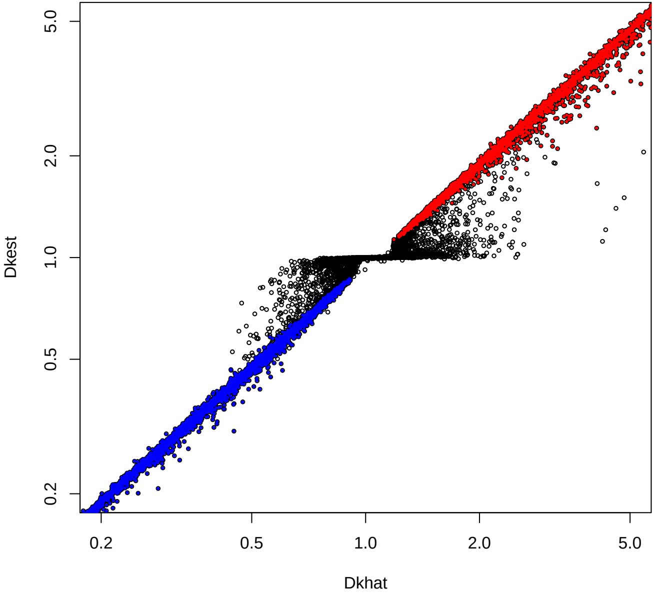
$$\text{Gene } k \text{ differentially expressed if} \quad 1 - p_k^{DE} < q_0.$$

No multiple testing correction is necessary!

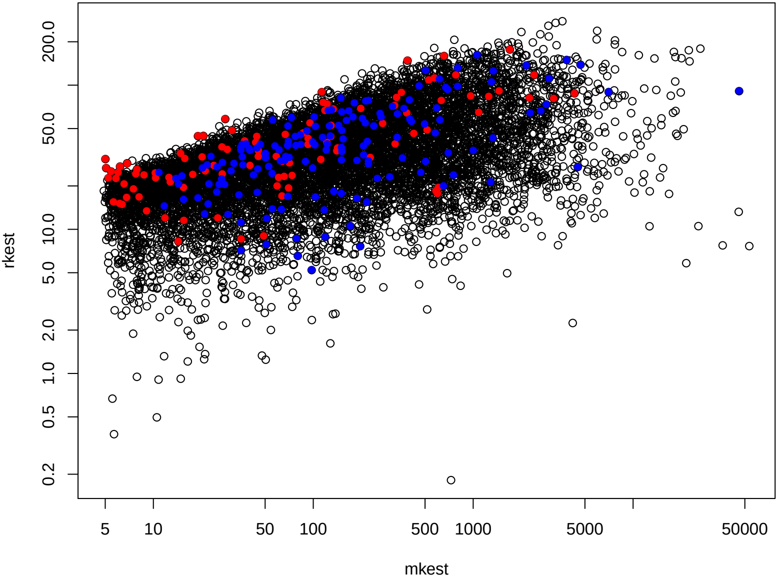
This test may be overly conservative since the actual $1 - p_k^{DE}$ may be much lower than $q_0$ threshold (i.e. $FDR < q_0$)

**Adjustment**: let $\widehat{FDR} = mean\{1 - p_k^{DE} : 1 - p_k^{DE} < q_0\}$

# Examples



# Examples


200.0
50.0
10.0
5.0
2.0
1.0
0.5
0.2
rkest
5
10
50
100
500
1000
5000
50000
mkest

## 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).

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 Cited by 11976

- Love MI, Huber W, Anders S (2014). `Moderated estimation of fold change and dispersion for RNA-seq data with` DESEQ2. Genome Biology, 15, 550 Cited by 11092

QUESTIONS?

THANK YOU!

see

euler.nmt.edu/~olegm/talks/GibbNB