Differential Gene Expression (part 2)

RNA-seq data analysis

Paulo Czarnewski
Julie Lorent

What is a GLM?

METHOD | Open Access | Published: 05 December 2014

Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2

Michael I Love, Wolfgang Huber & Simon Anders

Genome Biology **15**, Article number: 550 (2014) | Cite this article

Results and discussion

Model and normalization

The starting point of a *DESeq2* analysis is a count matrix K with one row for each gene i and one column for each sample j. The matrix entries K_{ij} indicate the number of sequencing reads that have been unambiguously mapped to a gene in a sample. Note that although we refer in this paper to counts of reads in genes, the methods presented here can be applied as well to other kinds of HTS count data. For each gene, we fit a generalized linear model (GLM) [12] as follows.

We model read counts K_{ij} as following a negative binomial distribution (sometimes also called a gamma-Poisson distribution) with mean μ_{ij} and dispersion α_i. The mean is taken as a quantity q_{ij}, proportional to the concentration of cDNA fragments from the gene in the sample, scaled by a normalization factor s_{ij}, i.e., $\mu_{ij} = s_{ij} q_{ij}$. For many applications, the same constant s_j can be used for all genes in a sample, which then accounts for differences in sequencing depth between samples. To estimate these *size factors*, the *DESeq2* package offers the median-of-ratios method already used in *DESeq* [4]. However, it can be advantageous to calculate gene-specific normalization factors s_{ij} to account for further sources of technical biases such as differing dependence on GC content, gene length or the like, using published methods [13],[14], and these can be supplied instead.

We use GLMs with a logarithmic link, $\log_2 q_{ij} = \sum_r x_{jr} \beta_{ir}$, with design matrix elements x_{jr} and

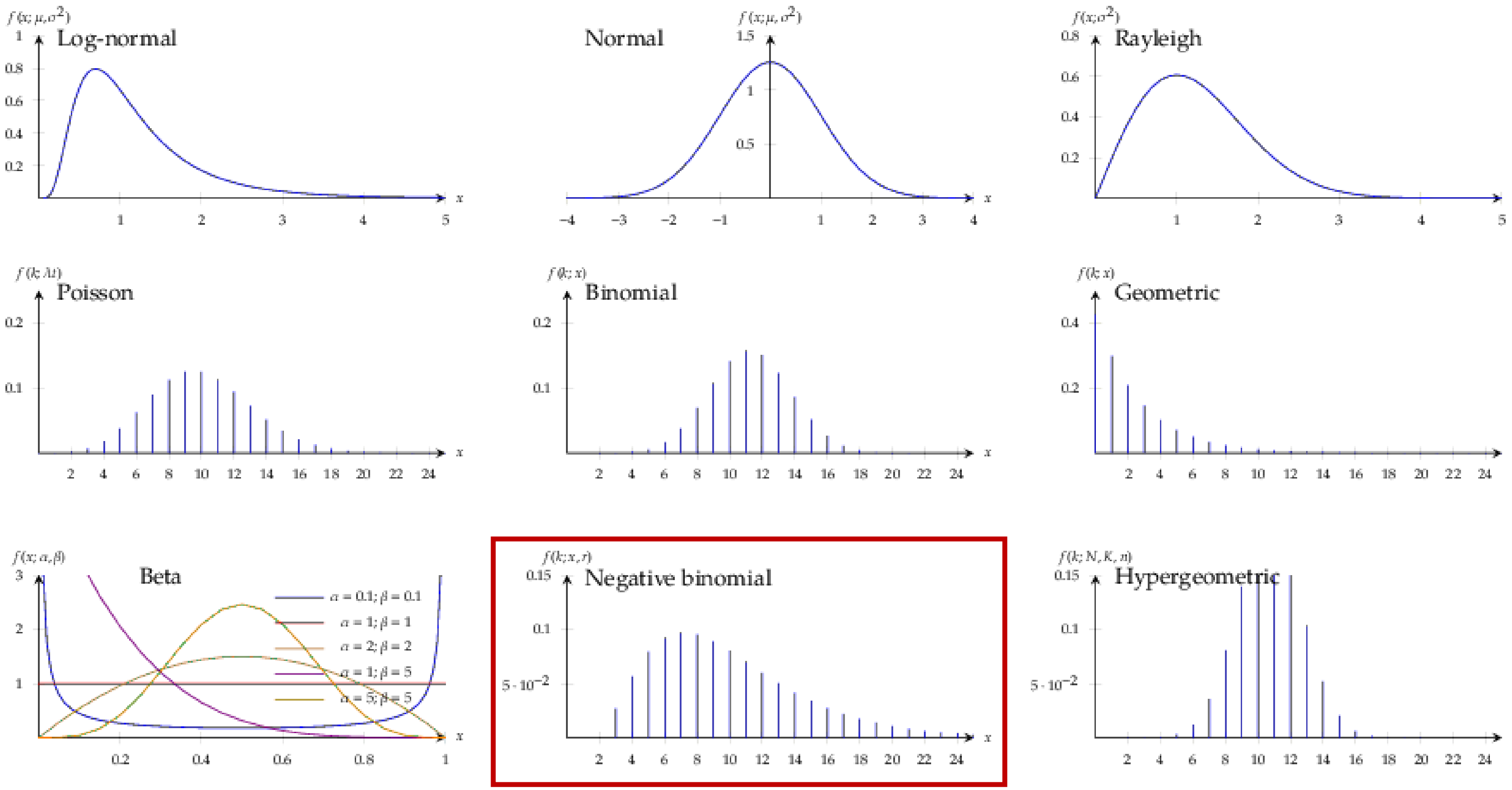
What is a GLM?

- GLMs extend linear model framework to allow outcome variables to be modeled via a link function
- Similar tools (for estimation, tests, diagnostic) as linear models after applying a link function
- They are most frequently used to model binary, categorical or count data
- **Flexible** method

Distribution	Support of distribution	Typical uses	Link name	Link function, $\mathbf{X}\boldsymbol{\beta} = g(\mu)$	Mean function
Normal	real: $(-\infty, +\infty)$	Linear-response data	Identity	$\mathbf{X}\boldsymbol{\beta} = \mu$	$\mu = \mathbf{X}\boldsymbol{\beta}$
Exponential	real: $(0, +\infty)$	Exponential-response data, scale parameters	Negative inverse	$\mathbf{X}\boldsymbol{\beta} = -\mu^{-1}$	$\mu = -(\mathbf{X}\boldsymbol{\beta})^{-1}$
Gamma					
Inverse Gaussian	real: $(0, +\infty)$		Inverse squared	$\mathbf{X}\boldsymbol{\beta} = \mu^{-2}$	$\mu = (\mathbf{X}\boldsymbol{\beta})^{-1/2}$
Poisson	integer: $0, 1, 2, \ldots$	count of occurrences in fixed amount of time/space	Log	$\mathbf{X}\boldsymbol{\beta} = \ln(\mu)$	$\mu = \exp(\mathbf{X}\boldsymbol{\beta})$
Bernoulli	integer: $\{0, 1\}$	outcome of single yes/no occurrence	Logit	$\mathbf{X}\boldsymbol{\beta} = \ln\left(\frac{\mu}{1-\mu}\right)$	$\mu = \frac{\exp(\mathbf{X}\boldsymbol{\beta})}{1+\exp(\mathbf{X}\boldsymbol{\beta})} = \frac{1}{1+\exp(-\mathbf{X}\boldsymbol{\beta})}$
Binomial	integer: $0, 1, \ldots, N$	count of # of "yes" occurrences out of N yes/no occurrences		$\mathbf{X}\boldsymbol{\beta} = \ln\left(\frac{\mu}{n-\mu}\right)$	
Categorical	integer: $[0, K)$	outcome of single K-way occurrence		$\mathbf{X}\boldsymbol{\beta} = \ln\left(\frac{\mu}{1-\mu}\right)$	
	K-vector of integer: $[0, 1]$, where exactly one element in the vector has the value 1				
Multinomial	*K*-vector of integer: $[0, N]$	count of occurrences of different types (1 .. *K*) out of *N* total *K*-way occurrences			

https://en.wikipedia.org/wiki/Generalized_linear_model

GLM Distributions



DESeq2 and **EdgeR** are improved negative-binomial GLMs

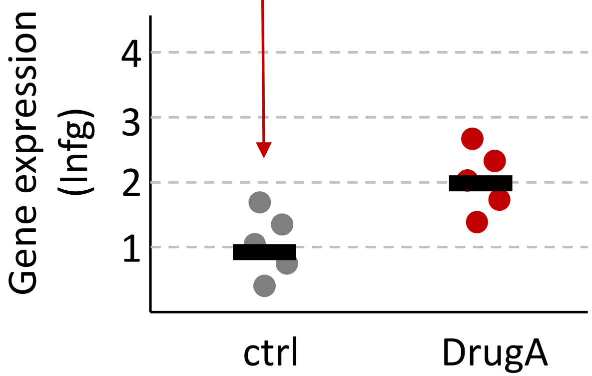
https://en.wikipedia.org/wiki/Generalized_linear_model

What if I have more than 2 groups?

What if I have 2 groups?

```
metadata$Drug <- factor( metadata$Drug ,
                 levels = c( "ctrl" , "DrugA" ) )
d <- DESeqDataSetFromMatrix(countData=cf, colData=metadata, design=~Drug)
```

intercept group
1st level of factor



Comparison between groups

```
y ~ Drug
```

gene	DrugA
⋮	⋮
Infg	1
⋮	⋮

effect sizes / FC / logFC

```
y ~ 0 + Drug
```

gene	ctrl	DrugA
⋮	⋮	⋮
Infg	1	2
⋮	⋮	⋮

Also testing if base expression is different than zero (not common)

What if I have 3 groups?

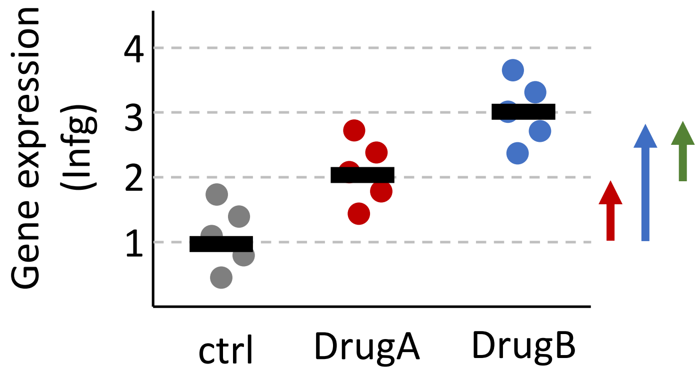
```
metadata$Drug <- factor( metadata$Drug ,
                 levels = c( "ctrl" , "DrugA" , "DrugB" ) )
d <- DESeqDataSetFromMatrix(countData=cf, colData=metadata, design=~Drug)
```

```
results(d, contrast=c("Drug","DrugA","ctrl"))
```

```
results(d, contrast=c("Drug","DrugB","ctrl"))
```

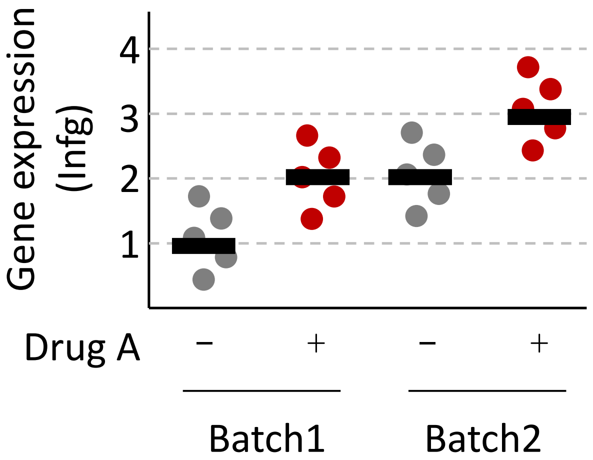
```
results(d, contrast=c("Drug","DrugB","DrugA"))
```

<u>intercept</u> group
1st level of factor



What if I have a batch effect?

```
metadata$Drug <- factor( metadata$Drug ,
                  levels = c( “ctrl” , ”DrugA” ) )
d <- DESeqDataSetFromMatrix(countData=cf, colData=metadata, design=~ Batch + Drug)
```




```
y ~ Batch + DrugA
```

```
results(d, contrast=c("Drug","DrugA","ctrl"))
```

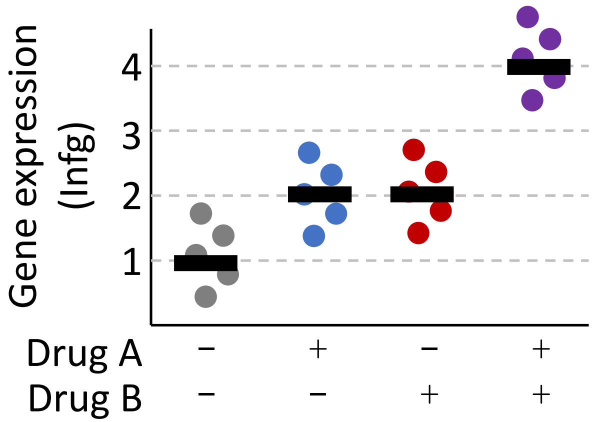
What if I have 2 variable groups and want to test for interactions?

```
metadata$DrugA <- factor( metadata$DrugA ,
                  levels = c( "ctrl" , "DrugA" ) )

metadata$DrugB <- factor( metadata$DrugB ,
                  levels = c( "ctrl" , "DrugB" ) )

d <- DESeqDataSetFromMatrix(countData=cf, colData=metadata, design ~DrugA +
DrugB + DrugA:DrugB)
d <- DESeq(d)
```

```
y ~ DrugA + DrugB + DrugA:DrugB
```

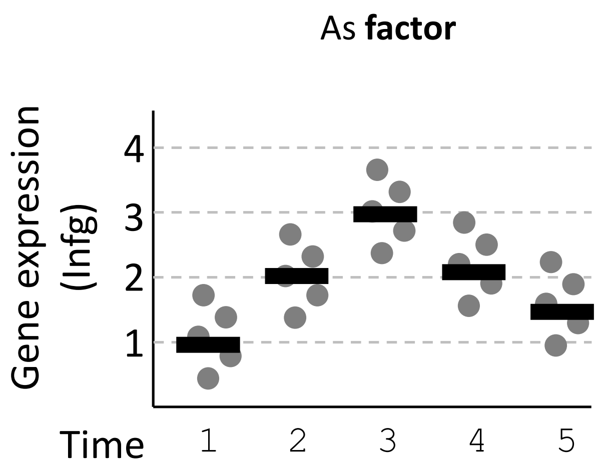



```
resultsNames(d)

"Intercept"             "DrugA_DrugA_vs_ctrl"
"DrugB_DrugB_vs_ctrl"   "DrugA.DrugB"
```

What if I have time series (or other continuous)?

```
y ~ Time
```

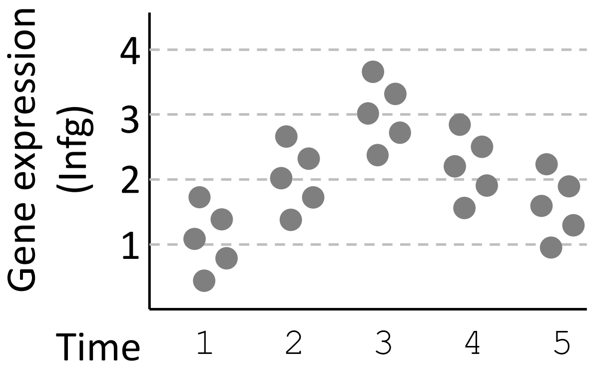


IMPORTANT: Other continuous covariates (such as **patient age** , **exposure time**, etc) should be used as **numeric** if they don't represent grouping variables.

What if I have time series (or other continuous)?

```
y ~ Time
```

As **numeric**



IMPORTANT: Other continuous covariates (such as **patient age** , **exposure time**, etc) should be used as **numeric** if they don't represent grouping variables.

What if I have time series and a treatment?

```
y ~ Time * Treatment
```

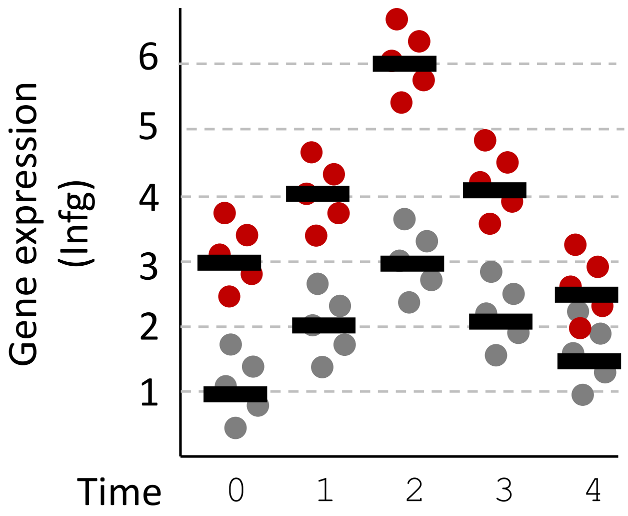
=

```
y ~ Time + Treatment + Time:Treatment
```

↑ Overall DGE at any time point

↑ DGE between conditions

↑ DGE between conditions specific to each time point



Thank you. Questions?