Regression for counts

Jeff Leek

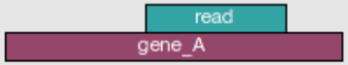
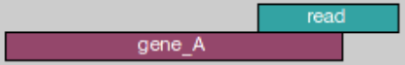
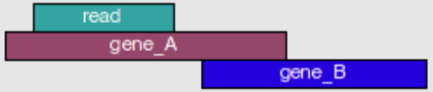
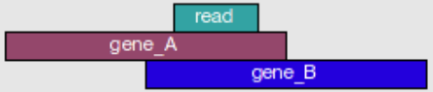
@jtleek

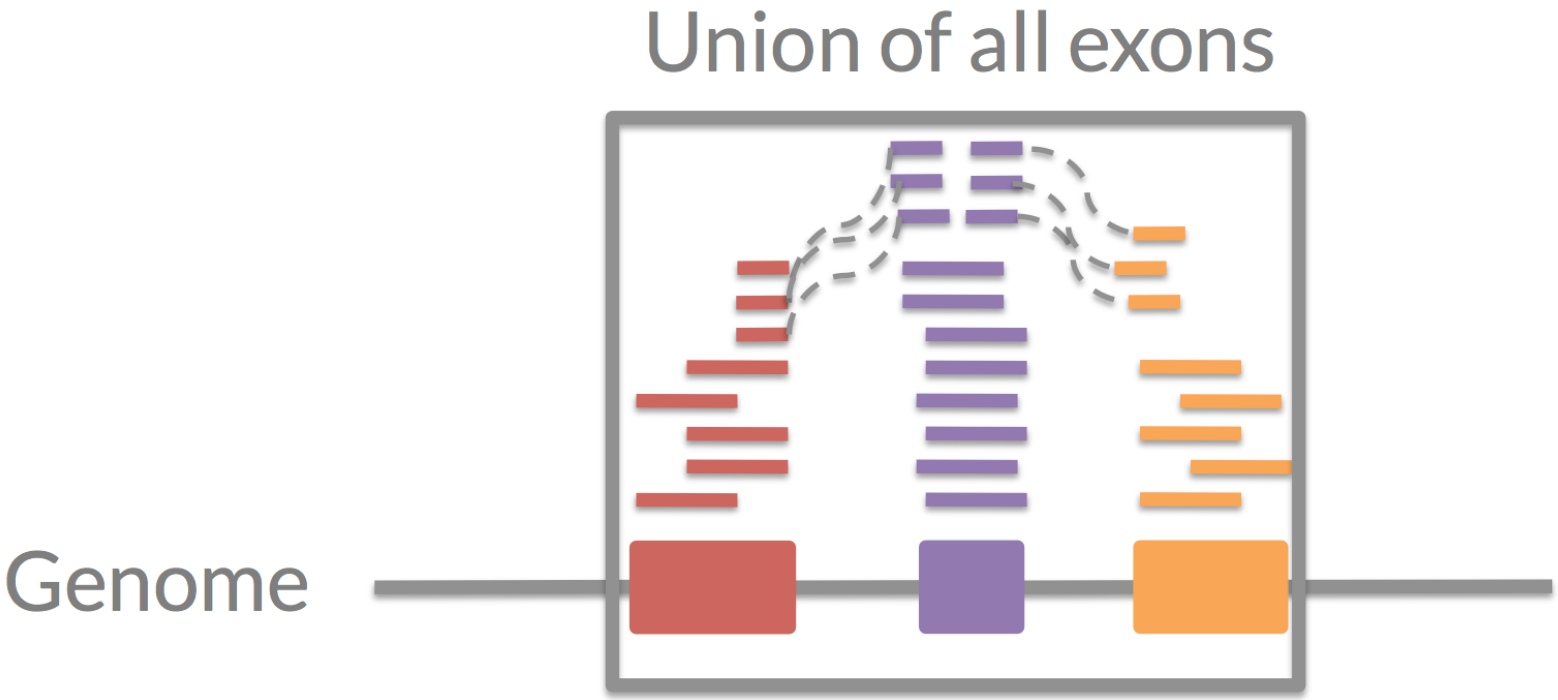
www.jtleek.com

Data aren’t always “Normal”

Sequencing data is often counts

http://www-huber.embl.de/users/anders/HTSeq/doc/count.html

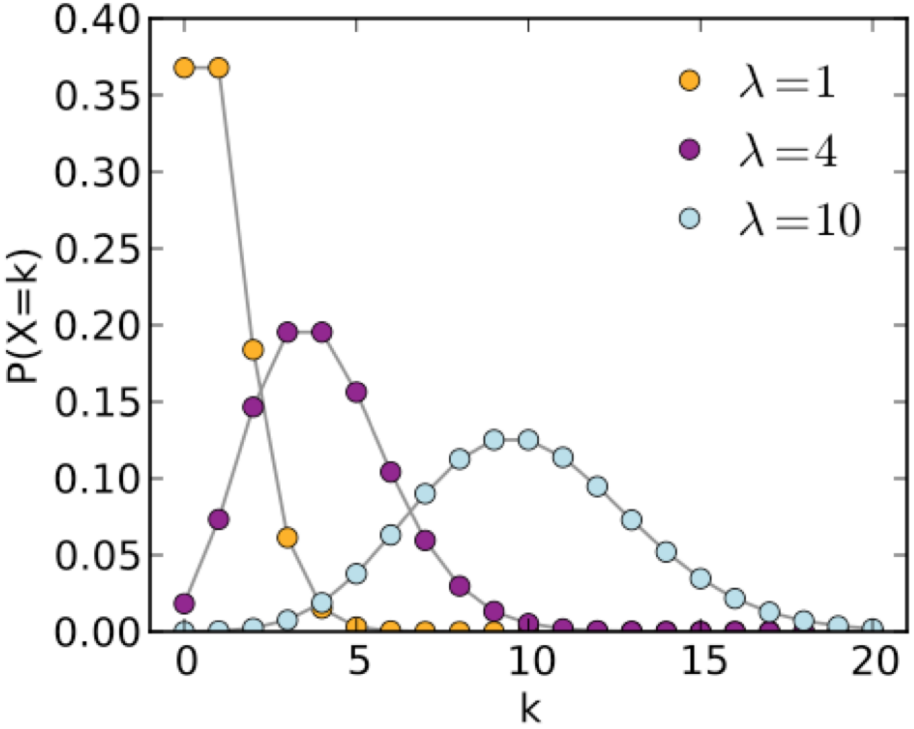
	union	intersection_strict	intersection_nonempty
	gene_A	gene_A	gene_A
	gene_A	no_feature	gene_A
	gene_A	no_feature	gene_A
	gene_A	gene_A	gene_A
	gene_A	gene_A	gene_A
	ambiguous	gene_A	gene_A
	ambiguous	ambiguous	ambiguous


Union of all exons
Genome

http://www.molgen.mpg.de/1242892/rnaseq

	sample1	sample2	sample3
gene1	0	0	0
gene2	0	12	1
gene3	1000	2000	100
gene4	10	20	2

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



$$f(k;\lambda) = \frac{e^{-\lambda}\lambda^k}{k!}$$

Mean & Variance: λ

http://www.genomebiology.com/content/11/8/r83

Normalized Counts
For Gene i, Sample j

Normalization Constant
For Sample j

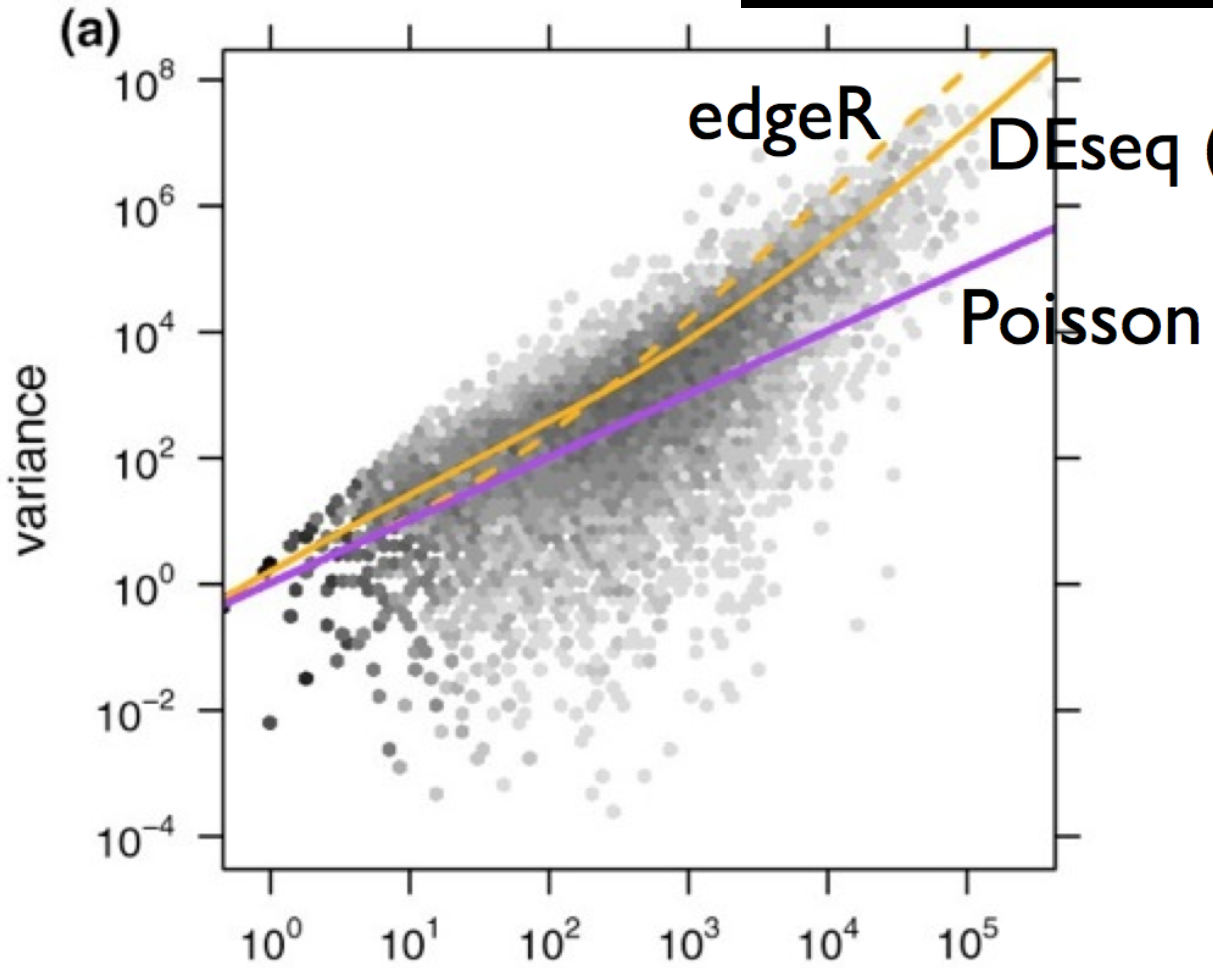
$$g(E[f(c_{ij}) \mid y_j]) = b_{i0} + \eta_i \log(q_j) + b_{i1} y_j$$

Link Function

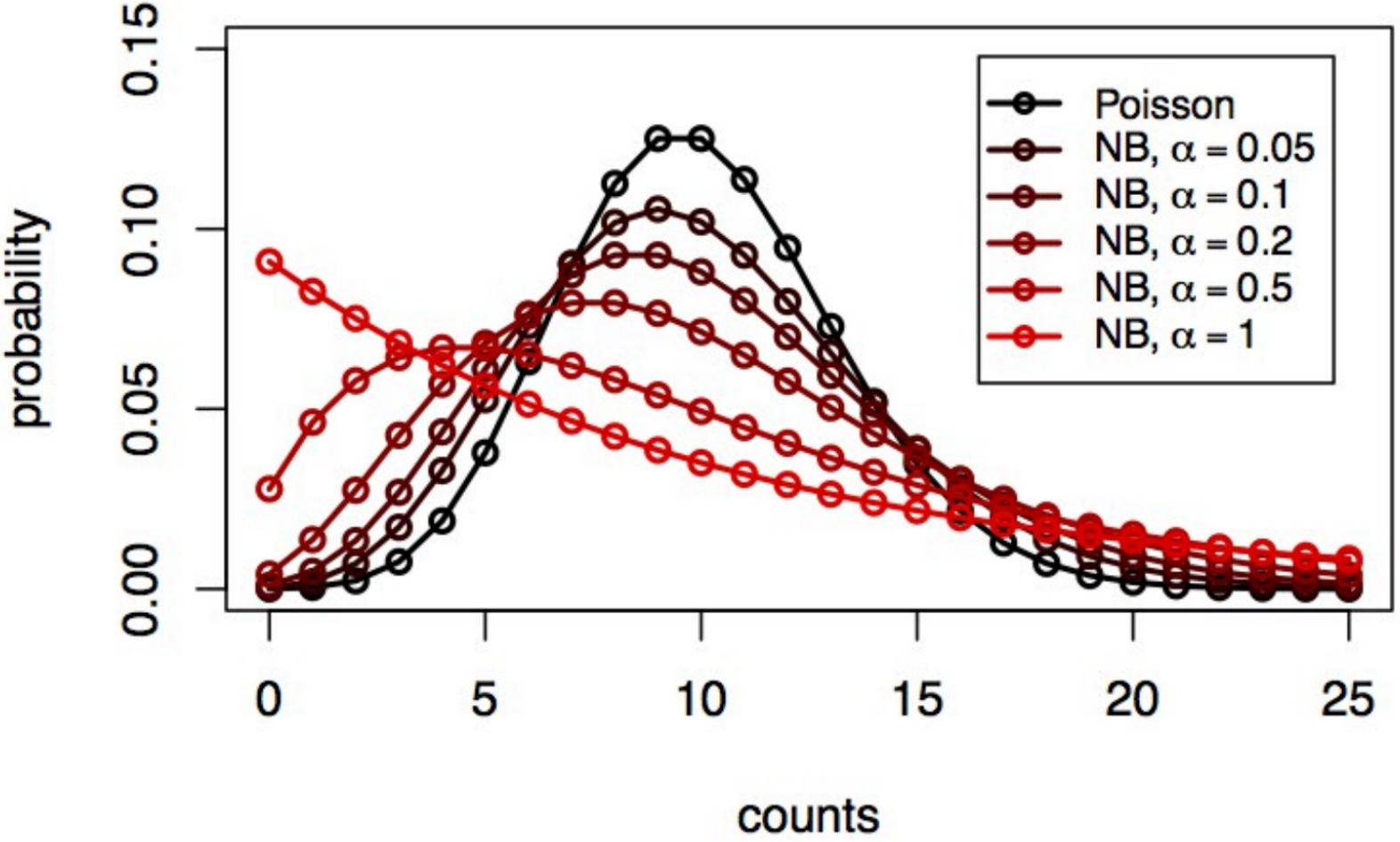
Group Indicator

Parameter We Test

http://www.genomebiology.com/content/11/10/R106



http://www.molgen.mpg.de/1242892/rnaseq



$$K_{ij} \sim \mathrm{NB}(\mu_{ij}, \alpha_i)$$

$$\mu_{ij} = s_j q_{ij}$$

$$\log_2(q_{ij}) = x_{j*}\vec{\beta}_i$$

K_{ij}	counts of reads for gene i, sample j
μ_{ij}	fitted mean
α_i	gene-specific dispersion
s_j	sample-specific size factor
q_{ij}	parameter proportional to the expected true concentration of fragments
x_{j*}	the j-th row of the design matrix X
$\vec{\beta}_i$	the log fold changes for gene i for each column of X

Notes and further reading

- Negative binomial/Poisson regression are “generalized linear models”
  - https://en.wikipedia.org/wiki/Generalized_linear_model
- A nice set of lecture notes
  - http://data.princeton.edu/wws509/notes/
- This is again a huge topic and we have only scratched the surface.