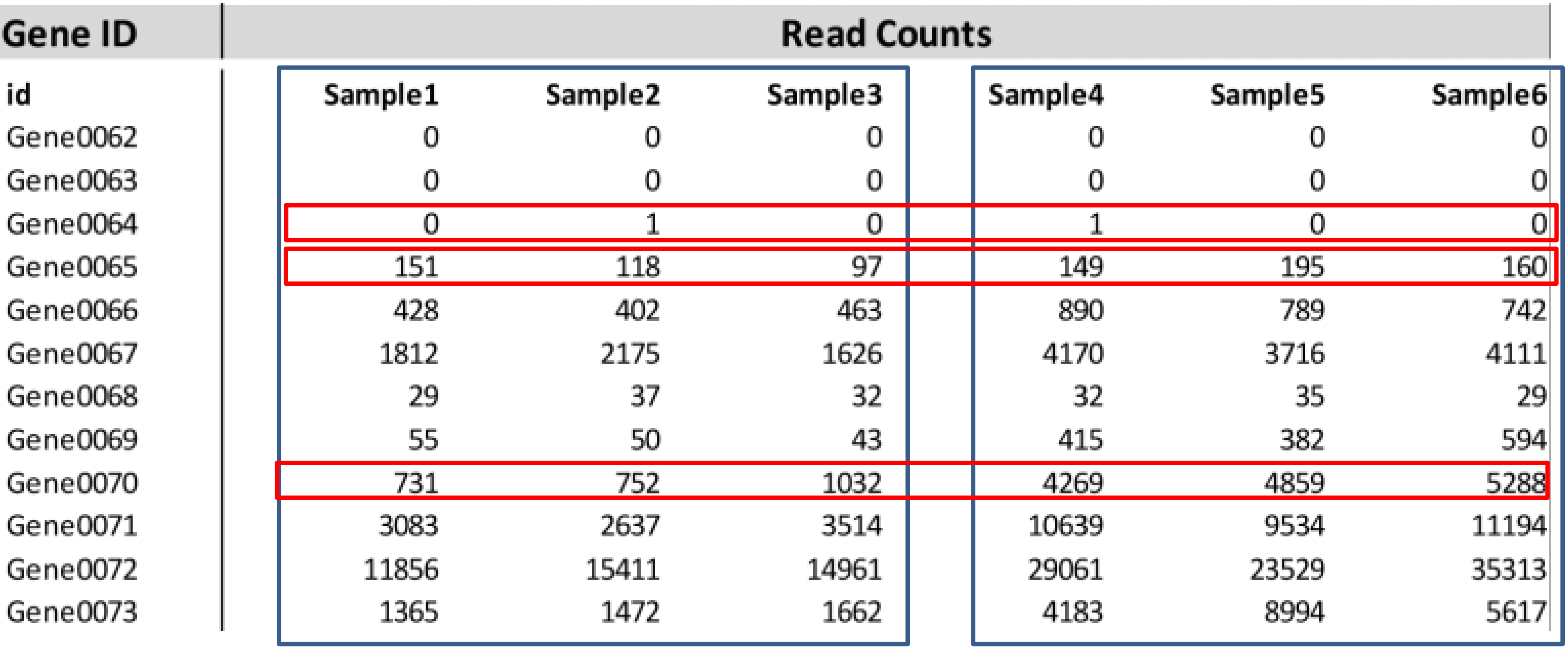
Statistical methods for RNA-seq analysis

Marcelo Falsarella Carazzolle
mcarazzo@lge.ibi.unicamp.br

Feb/2015

Read count matrix



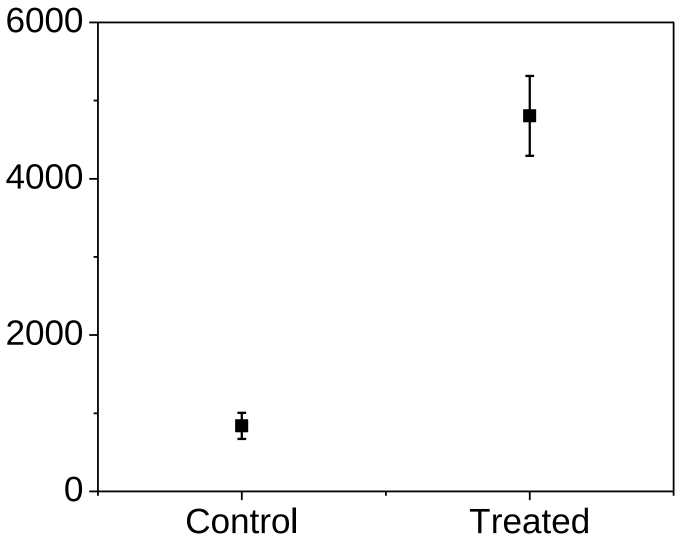
Gene ID	Read Counts					
id	Sample1	Sample2	Sample3	Sample4	Sample5	Sample6
Gene0062	0	0	0	0	0	0
Gene0063	0	0	0	0	0	0
Gene0064	0	1	0	1	0	0
Gene0065	151	118	97	149	195	160
Gene0066	428	402	463	890	789	742
Gene0067	1812	2175	1626	4170	3716	4111
Gene0068	29	37	32	32	35	29
Gene0069	55	50	43	415	382	594
Gene0070	731	752	1032	4269	4859	5288
Gene0071	3083	2637	3514	10639	9534	11194
Gene0072	11856	15411	14961	29061	23529	35313
Gene0073	1365	1472	1662	4183	8994	5617

…

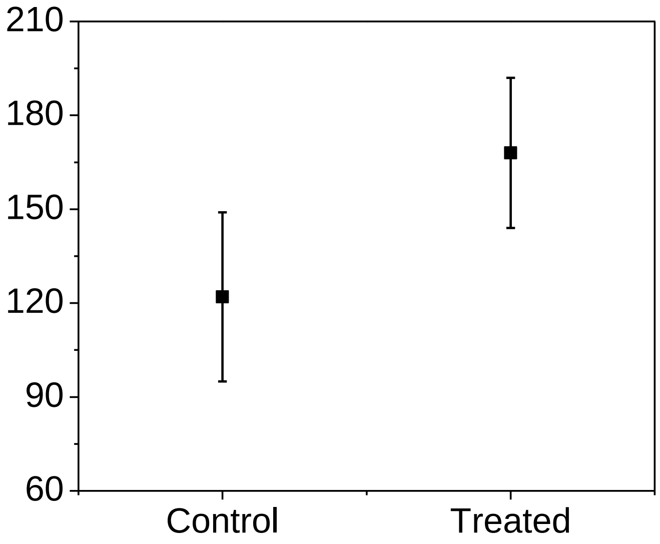
Biological triplicate of control

Biological triplicate of treatment

Gene25000



Treated/control ratio = 5.7x



Treated/control ratio = 1.4x
Are there differences ??
How to do analysis for 25,000 genes ?

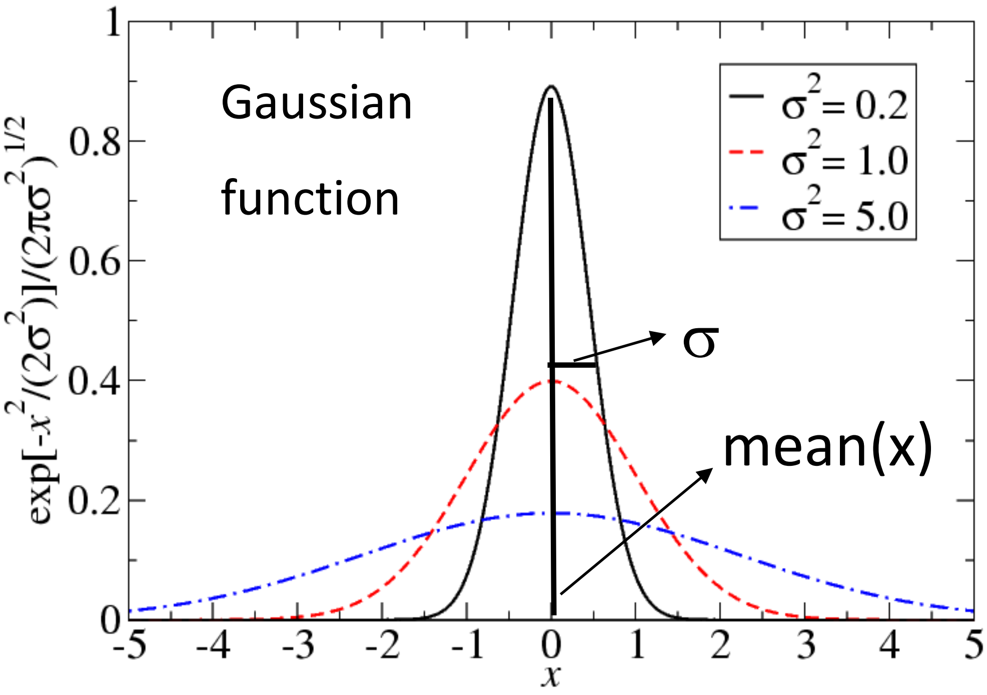
Statistical methods

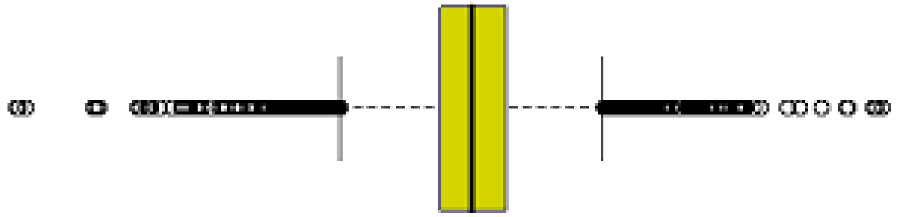
- Measured value = true valor ± error
- Error = experimental (~~systematic and randomic~~) + biological variation

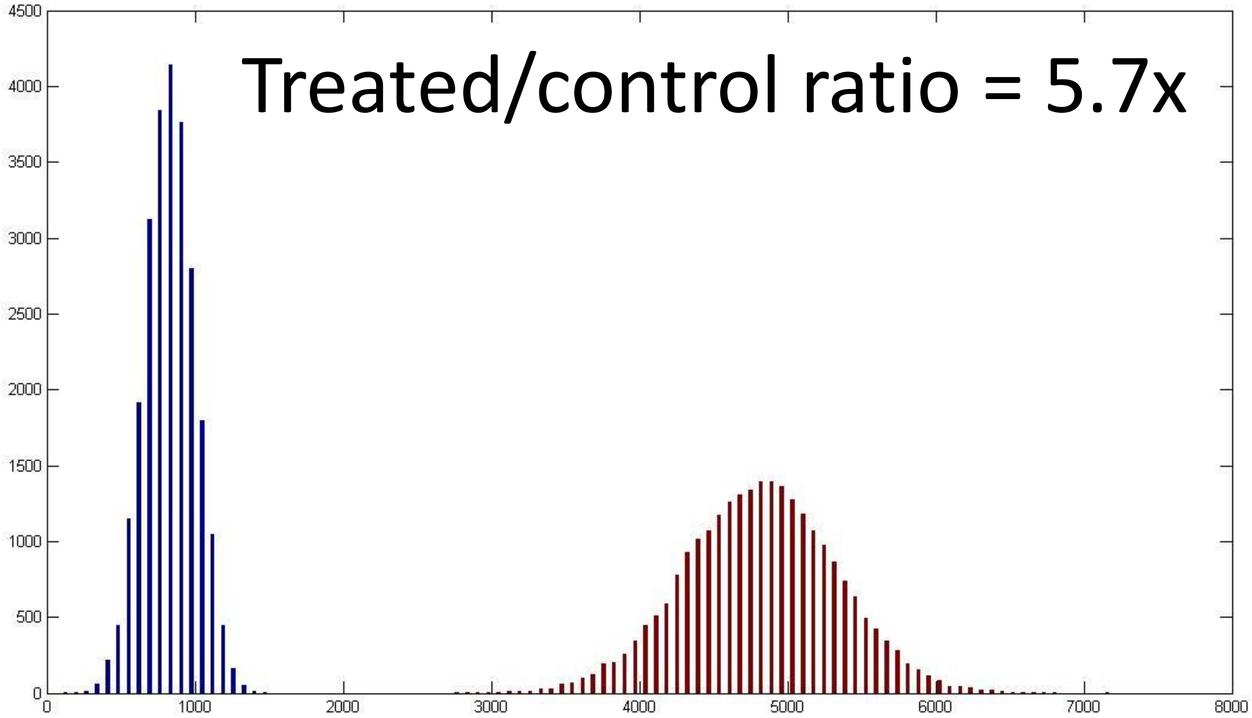
- Biological variation is described by gaussian distribution that can be estimated using experimental replicates

Average : $$\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{x_1 + x_2 + \cdots + x_N}{N}$$

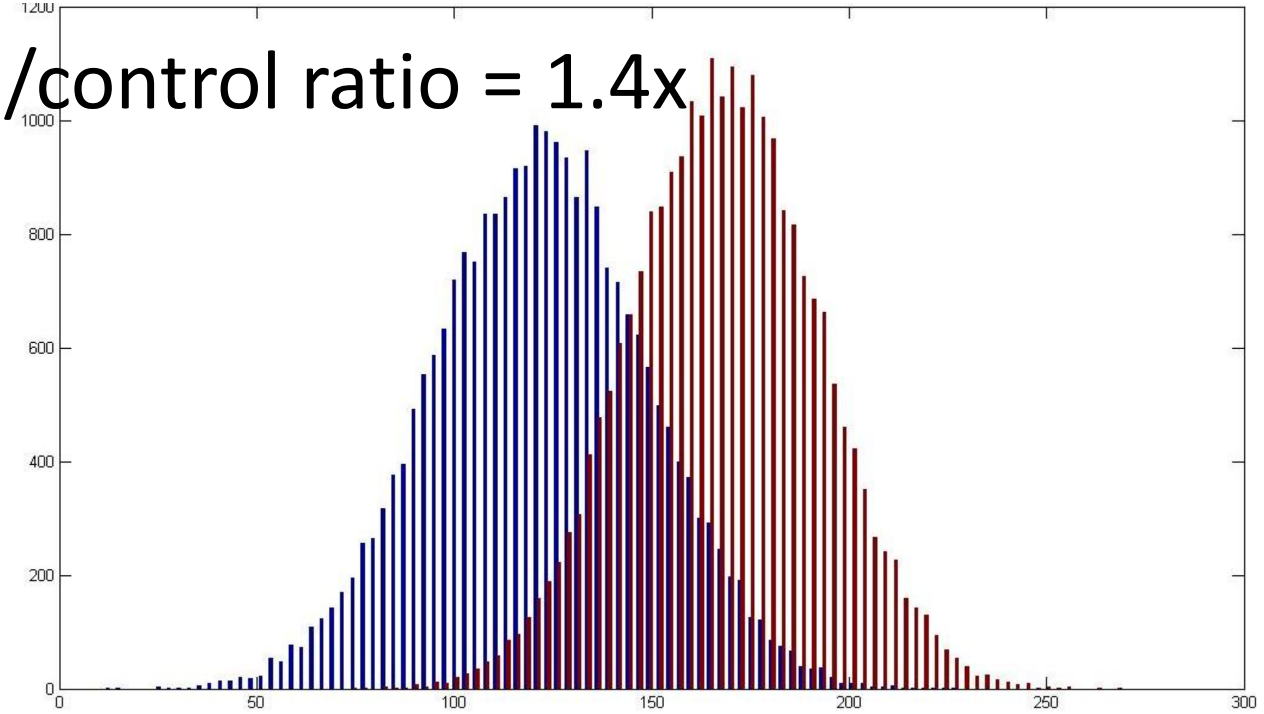
Stdev : $$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2}$$





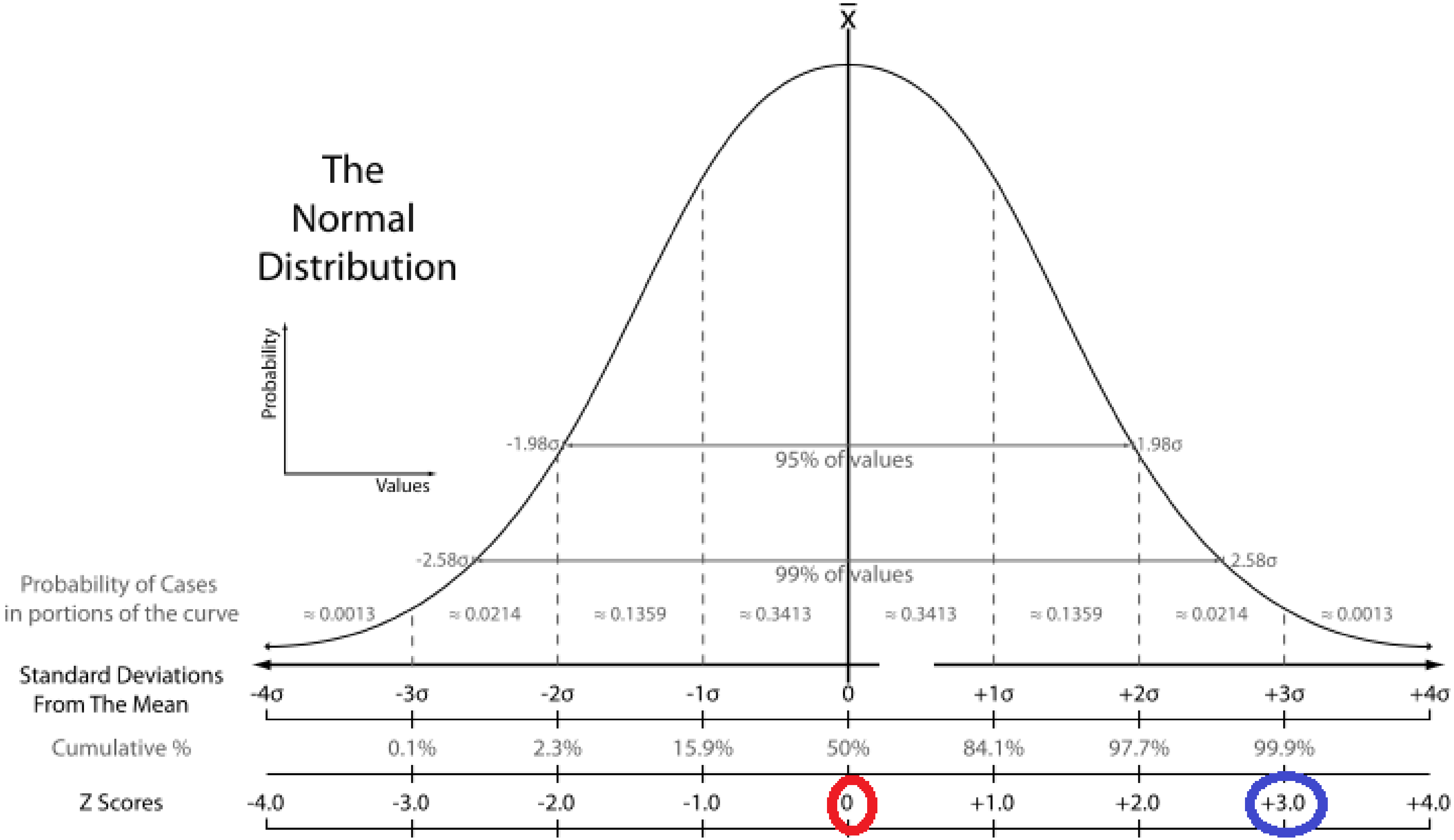


Treated/control ratio = 5.7x



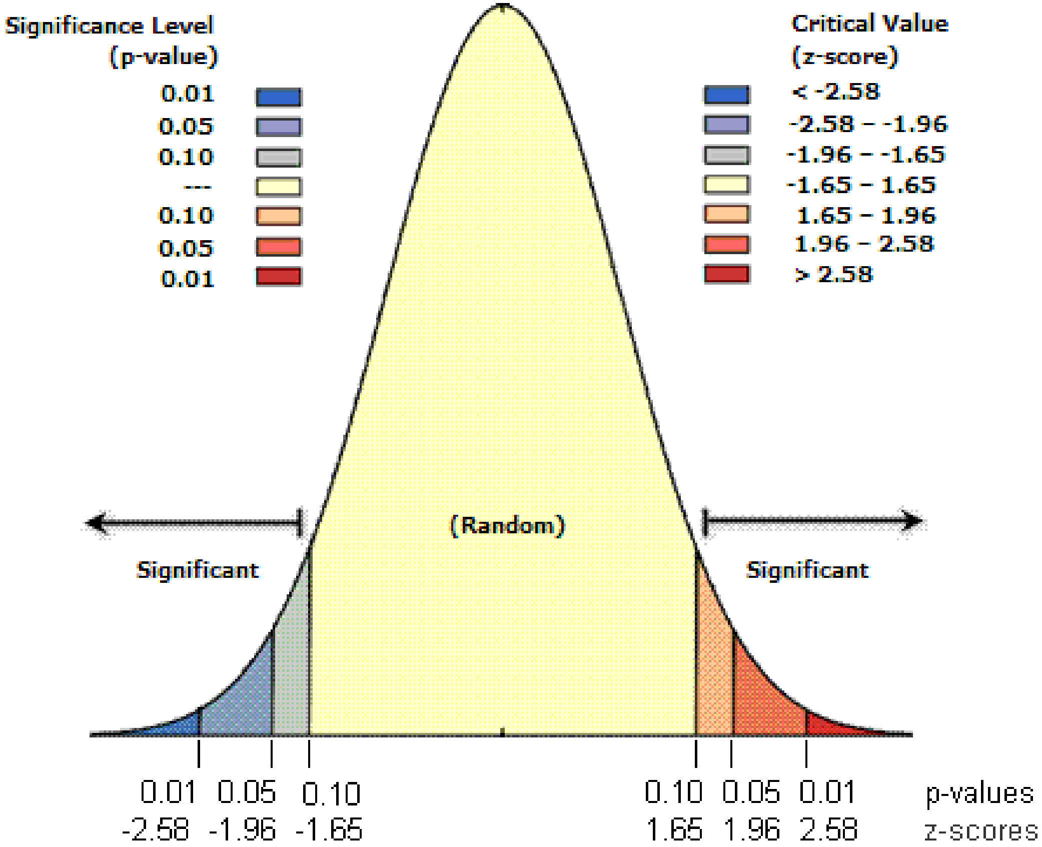
Treated/control ratio = 1.4x

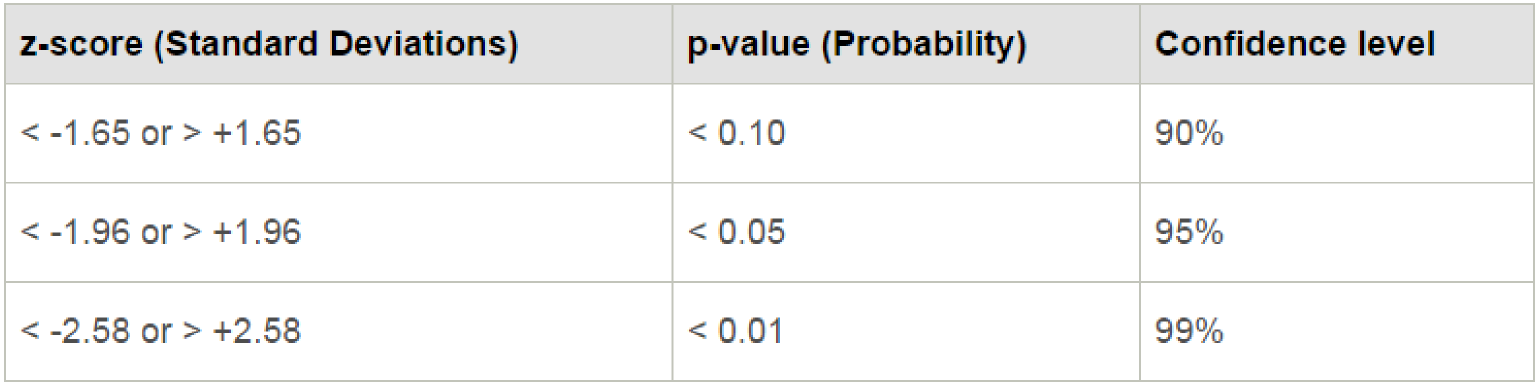
Confidence intervals



$$\textbf{Z Score=} \ \frac{x - \mu}{\sigma}$$

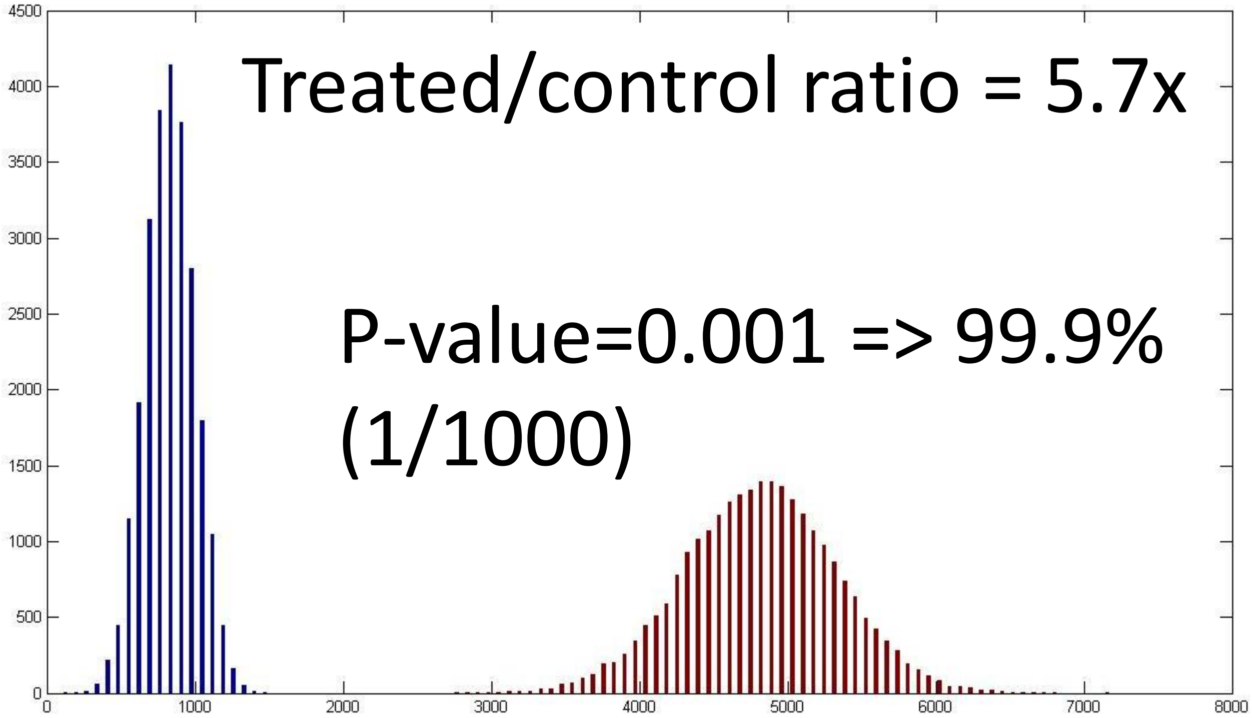
$$\textbf{Z Score=} \ \frac{\text{Raw score - Mean}}{\text{Standard deviation}}$$





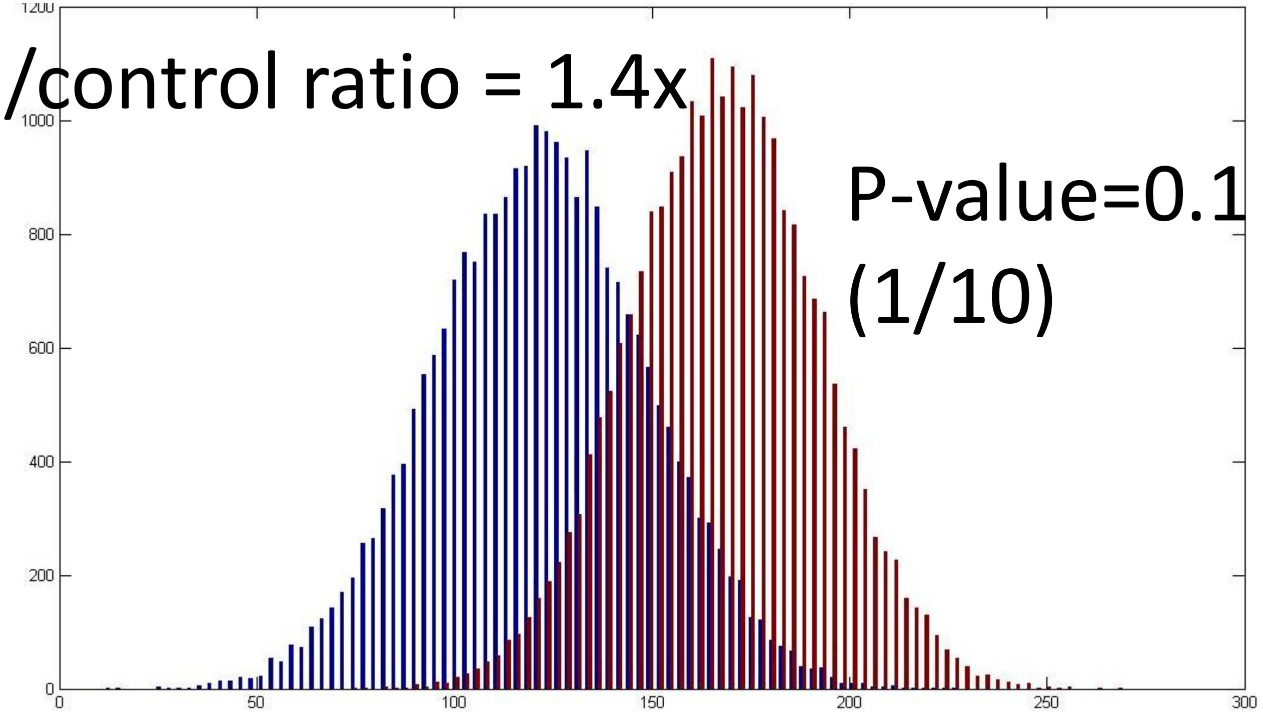
z-score (Standard Deviations)	p-value (Probability)	Confidence level
< -1.65 or > +1.65	< 0.10	90%
< -1.96 or > +1.96	< 0.05	95%
< -2.58 or > +2.58	< 0.01	99%

P-value is defined as a probability of rejection of the null hypothesis (there is no difference between these values)



Treated/control ratio = 5.7x

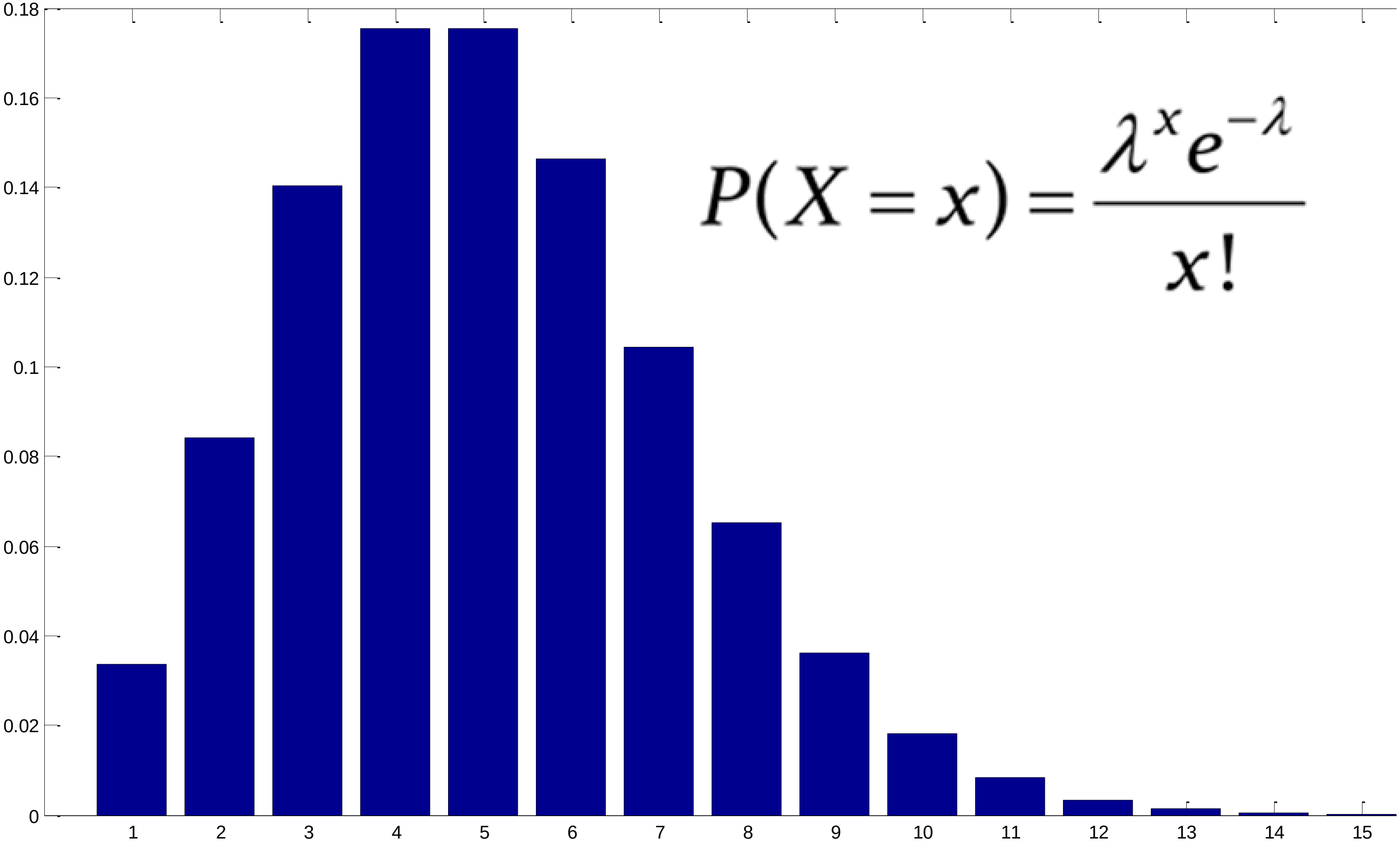
P-value=0.001 => 99.9% (1/1000)



Treated/control ratio = 1.4x

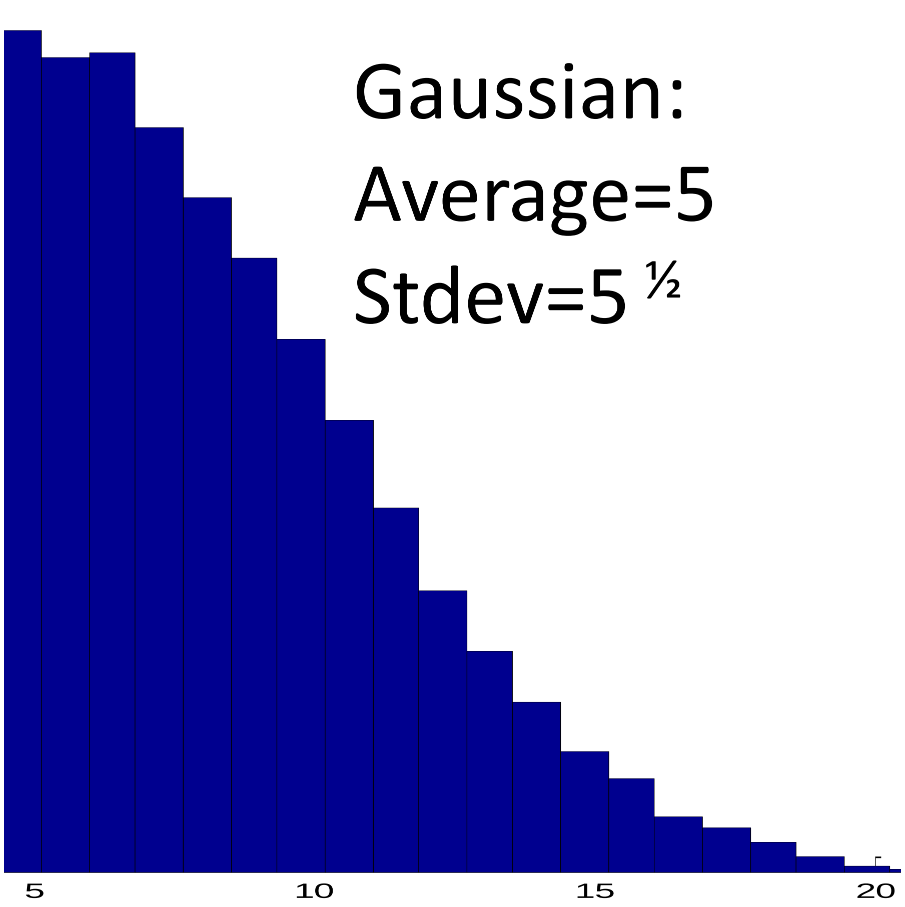
P-value=0.1 => 90.0% (1/10)

Poisson distribution



$$P(X=x)=\frac{\lambda^{x}e^{-\lambda}}{x!}$$

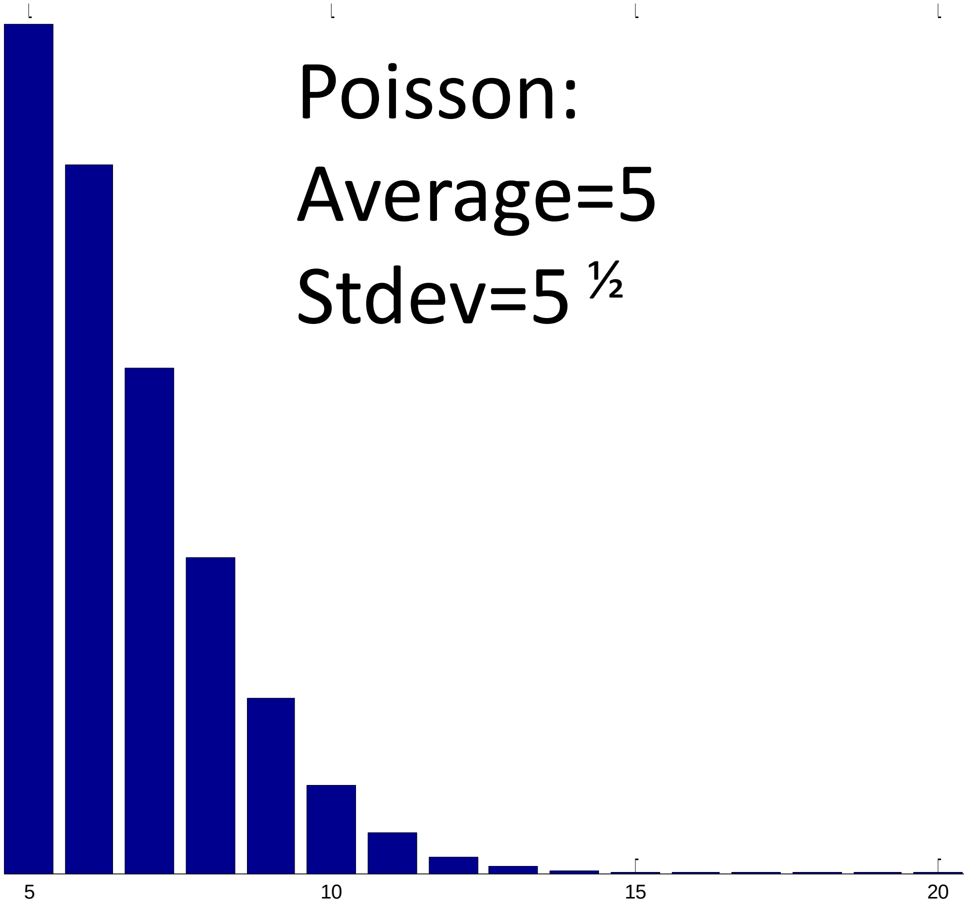
- Asymmetric distribution
- Applied for rare events (probability close to zero)
- Average = (stdev)^2



Gaussian:

Average=5

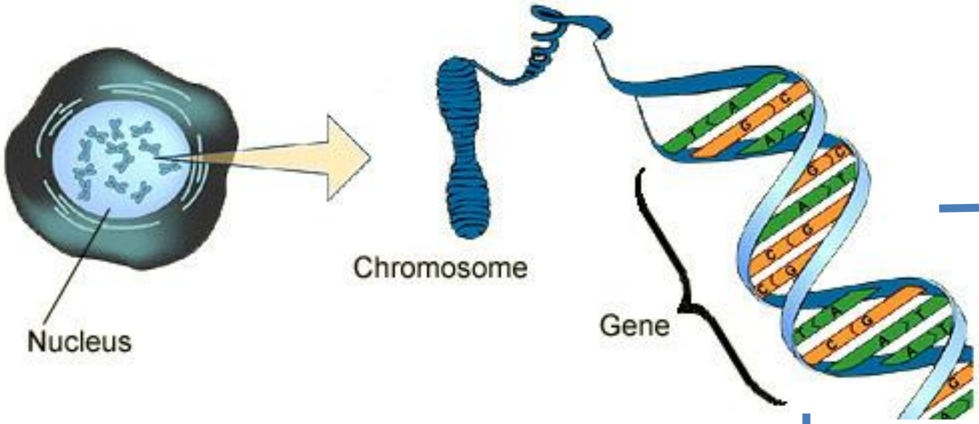
Stdev=$5^{½}$



Poisson:

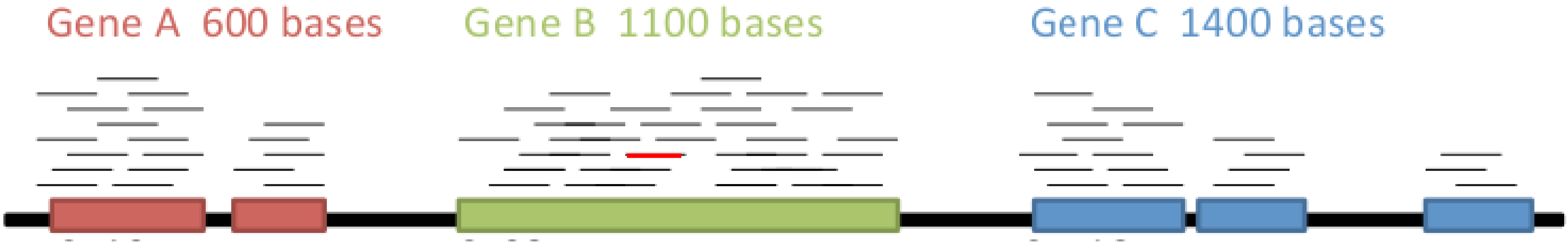
Average=5

Stdev=$5^{½}$



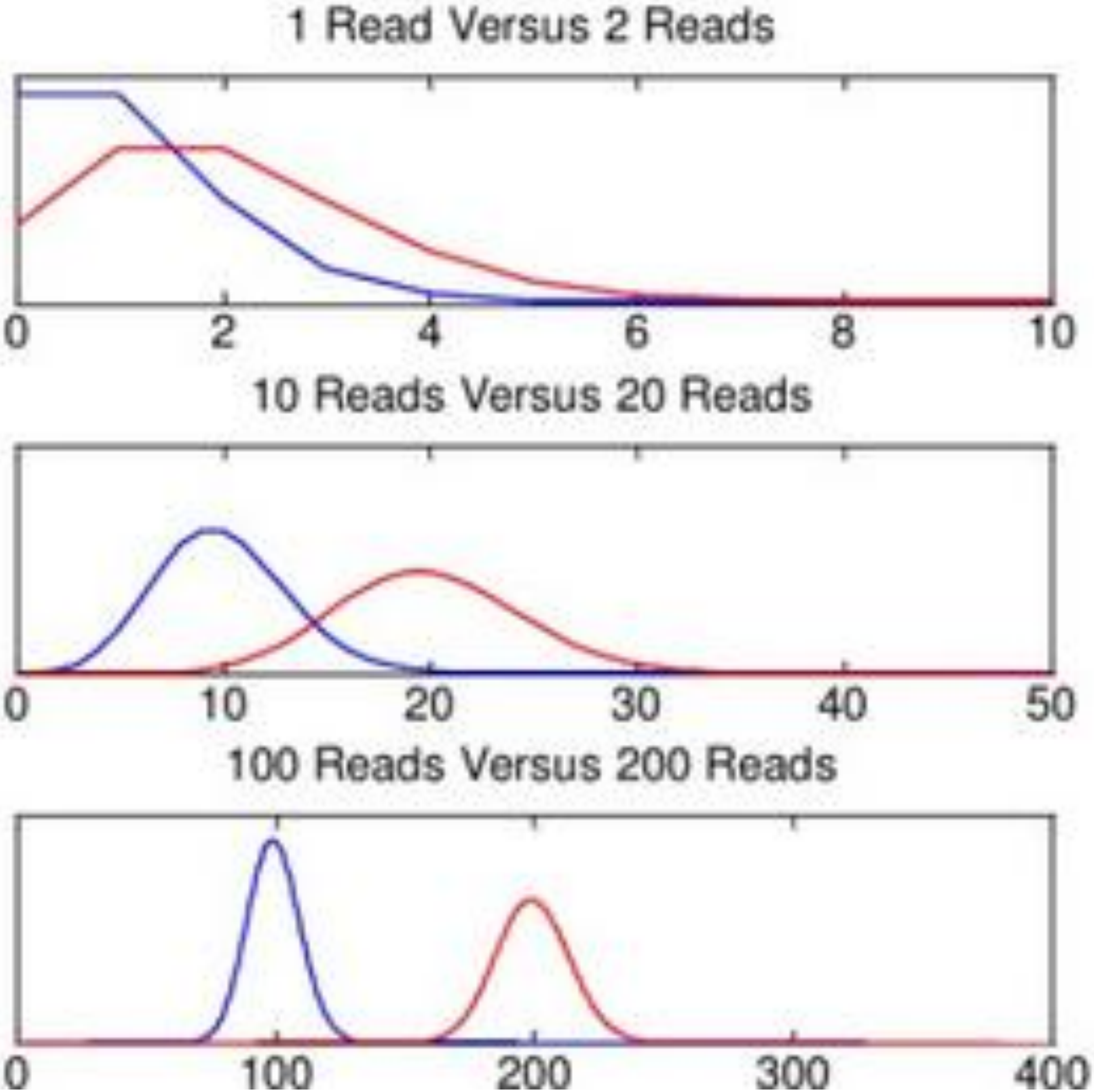
Several thousand genes

Several millions of sequenced reads



Q: What is the probability to get one specific read for one specific gene considering all expressed genes ??

R: Too small => read count follows a Poisson distribution

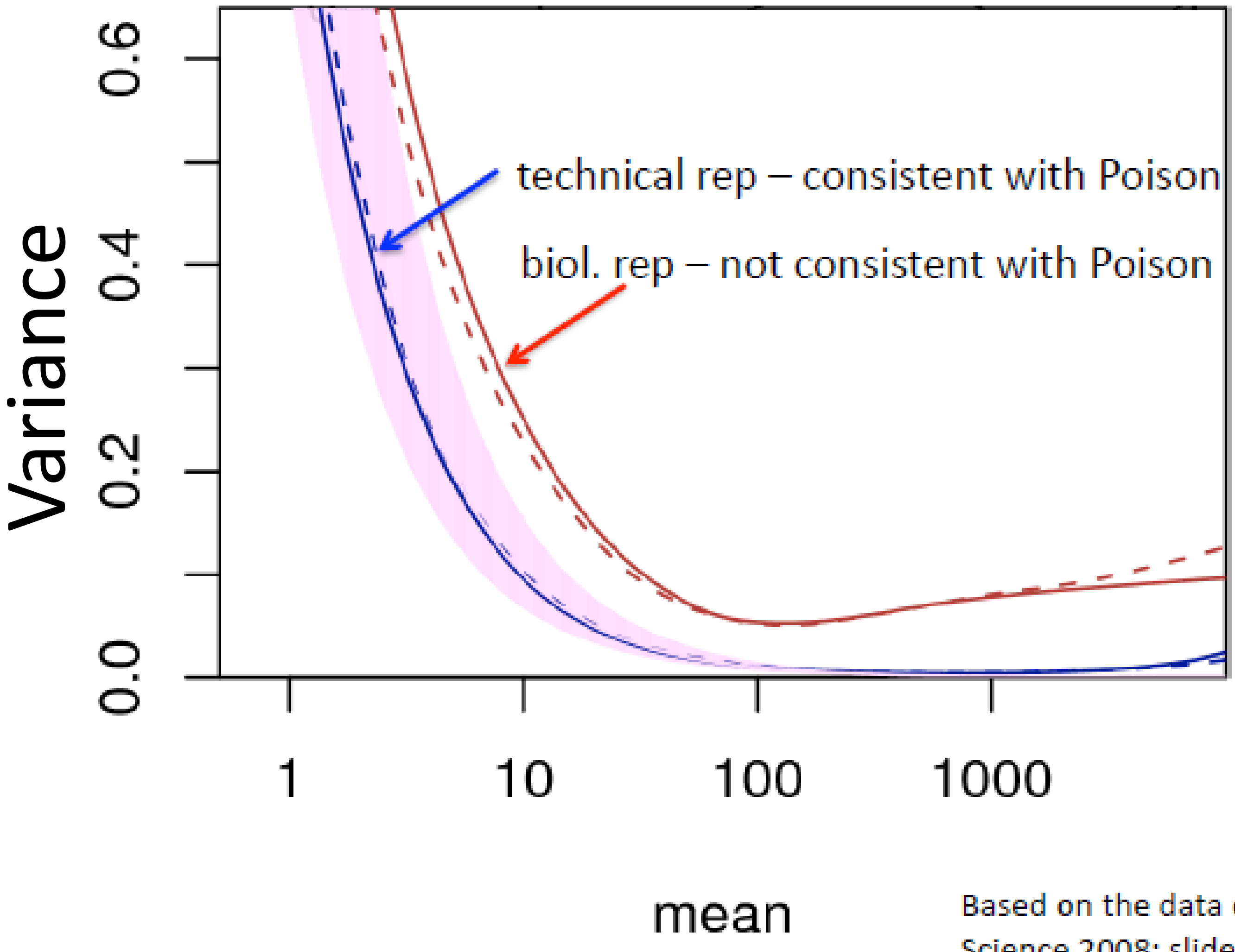


1 Read Versus 2 Reads

10 Reads Versus 20 Reads

100 Reads Versus 200 Reads

Need to account for extra variability

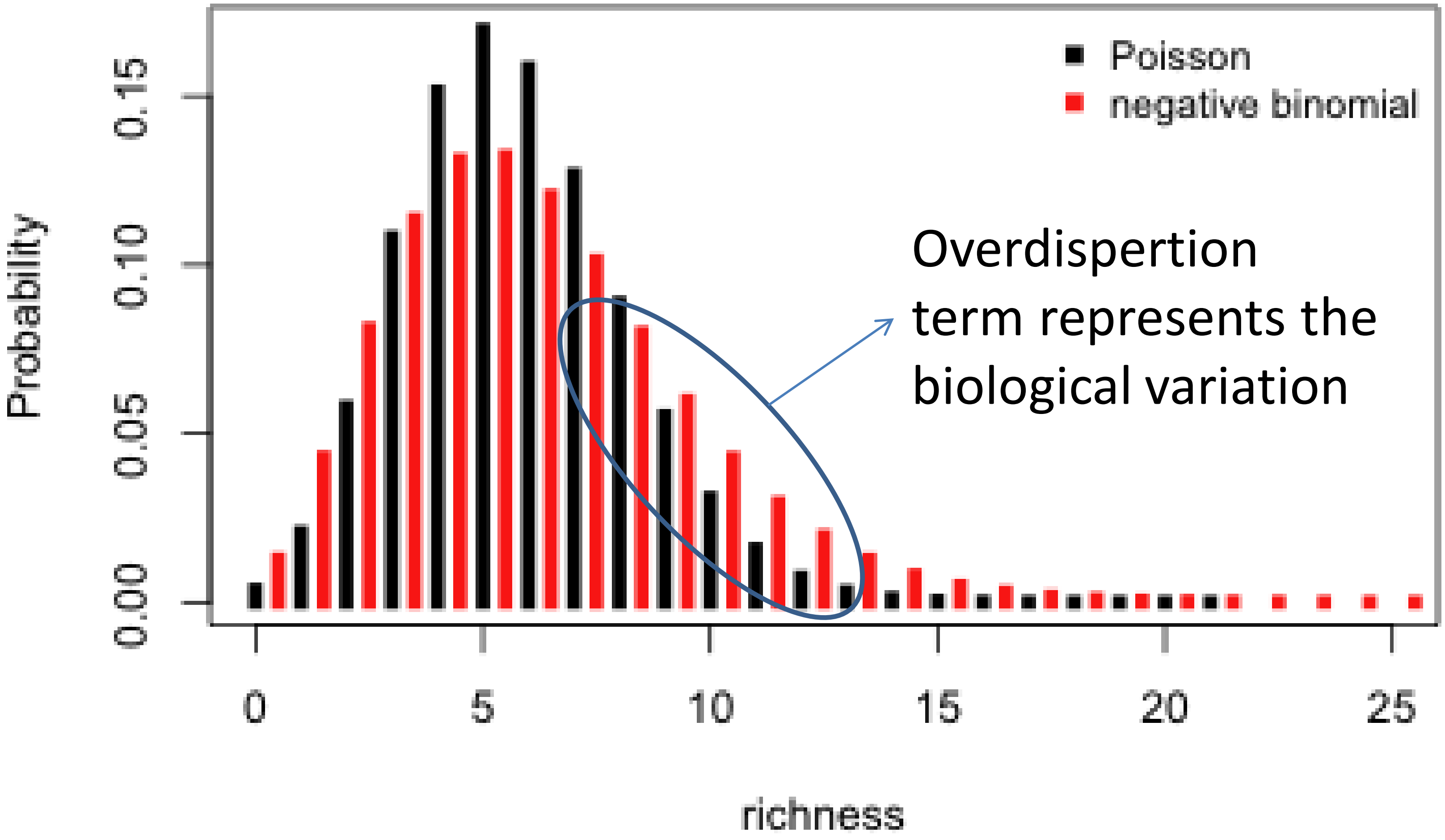


technical rep – consistent with Poison

biol. rep – not consistent with Poison

Based on the data of Nagalakshmi et al. Science 2008; slide adapted from Huber;

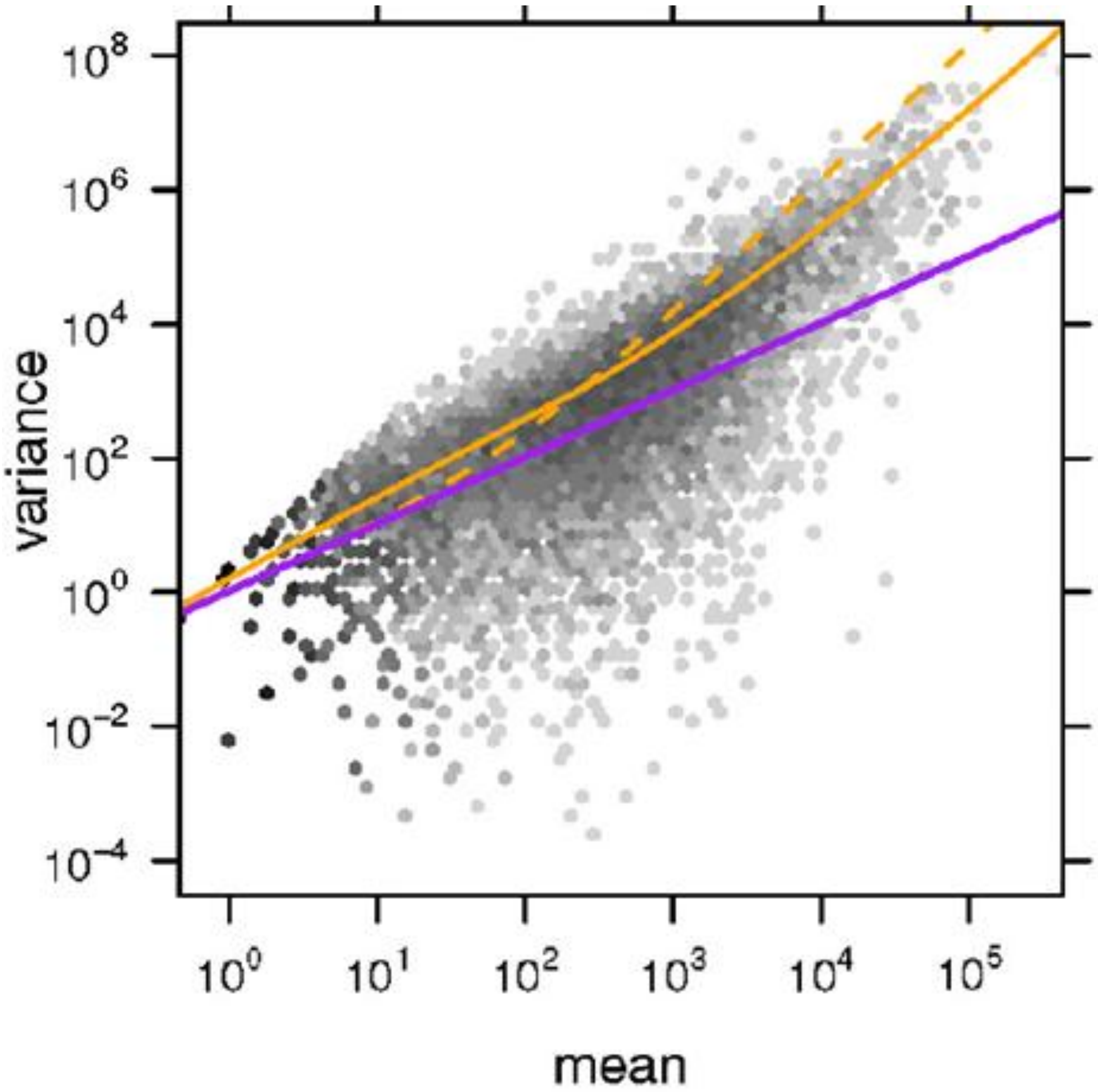
Negative binomial distribution



Overdispertion term represents the biological variation

(stdev)^2 = average + overdispertion_term

$$\sigma^2 = \mu + \frac{1}{r}\mu^2$$



-Negative binomial distribution is the most used probability distribution for RNA-seq

-The distribution is centered at read count average (biological replicates)

-The overdispertion term is obtained from fitting of variance vs mean

Genome Biology (2010), 11:R106

Adjusted p-value

-P-value is defined as a probability of rejection of the null hypothesis for each gene

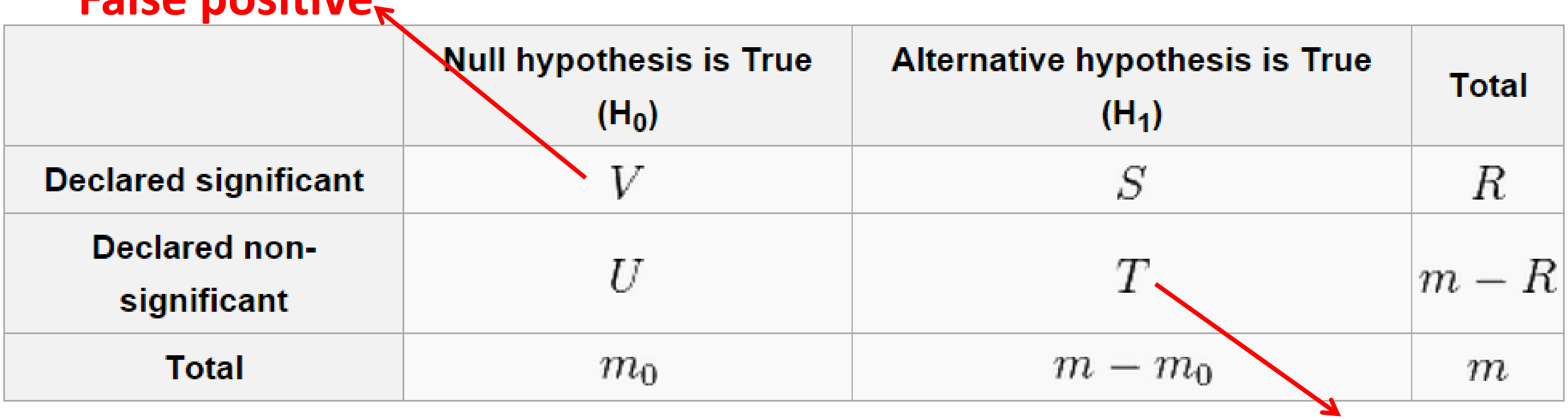
-Typically one RNA-seq experiment generates tens of thousands statistical tests

-Considering a set of 20,000 genes and p-value <= 0.05 (5/100) => 5/100 * 20,000 = 1000 false positives (differentialy expressed genes classified incorrectly)

-Adjusted p-value methodologies are necessary to decrease the false positive rates

PNAS (2003), vol. 100: 9440-9445

False positive



	Null hypothesis is True (H_0)	Alternative hypothesis is True (H_1)	Total
Declared significant	V	S	R
Declared non-significant	U	T	$m - R$
Total	m_0	$m - m_0$	m

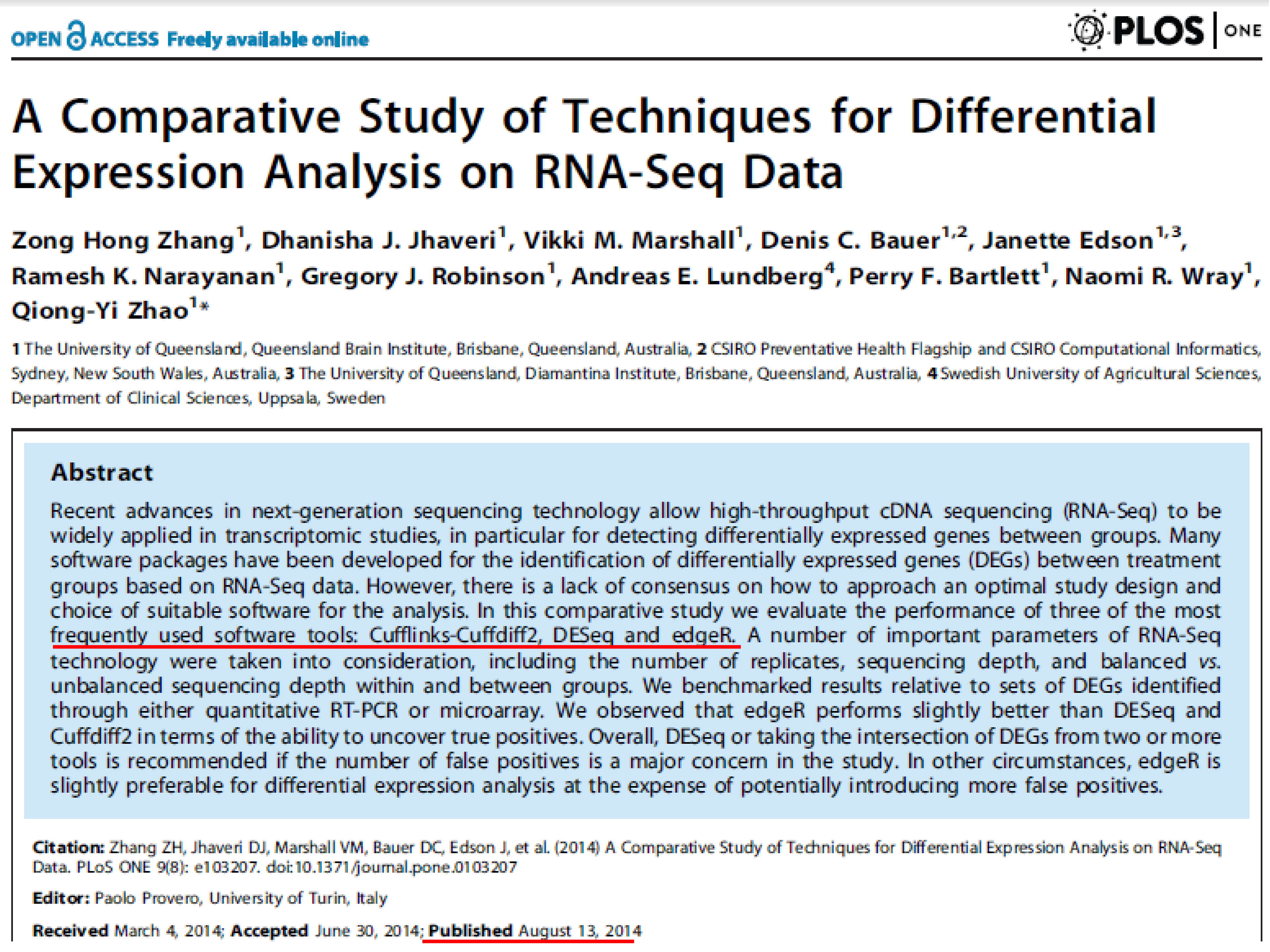
False negative

$$FDR = Q_e = \mathrm{E}[Q] = \mathrm{E}\left[\frac{V}{V+S}\right] = \mathrm{E}\left[\frac{V}{R}\right],$$

→ The proportion of false positive features among all of called significant

P-value of 0.05 implies that 5% of all tests will result in false positives. An FDR adjusted p-value (or q-value) of 0.05 implies that 5% of significant tests will result in false positives

Nat Biotechnol. 2009 December ; 27(12): 1135–1137.



OPEN ACCESS Freely available online

PLOS | ONE

A Comparative Study of Techniques for Differential Expression Analysis on RNA-Seq Data

Zong Hong Zhang[1], Dhanisha J. Jhaveri[1], Vikki M. Marshall[1], Denis C. Bauer[1,2], Janette Edson[1,3], Ramesh K. Narayanan[1], Gregory J. Robinson[1], Andreas E. Lundberg[4], Perry F. Bartlett[1], Naomi R. Wray[1], Qiong-Yi Zhao[1]*

1 The University of Queensland, Queensland Brain Institute, Brisbane, Queensland, Australia, 2 CSIRO Preventative Health Flagship and CSIRO Computational Informatics, Sydney, New South Wales, Australia, 3 The University of Queensland, Diamantina Institute, Brisbane, Queensland, Australia, 4 Swedish University of Agricultural Sciences, Department of Clinical Sciences, Uppsala, Sweden

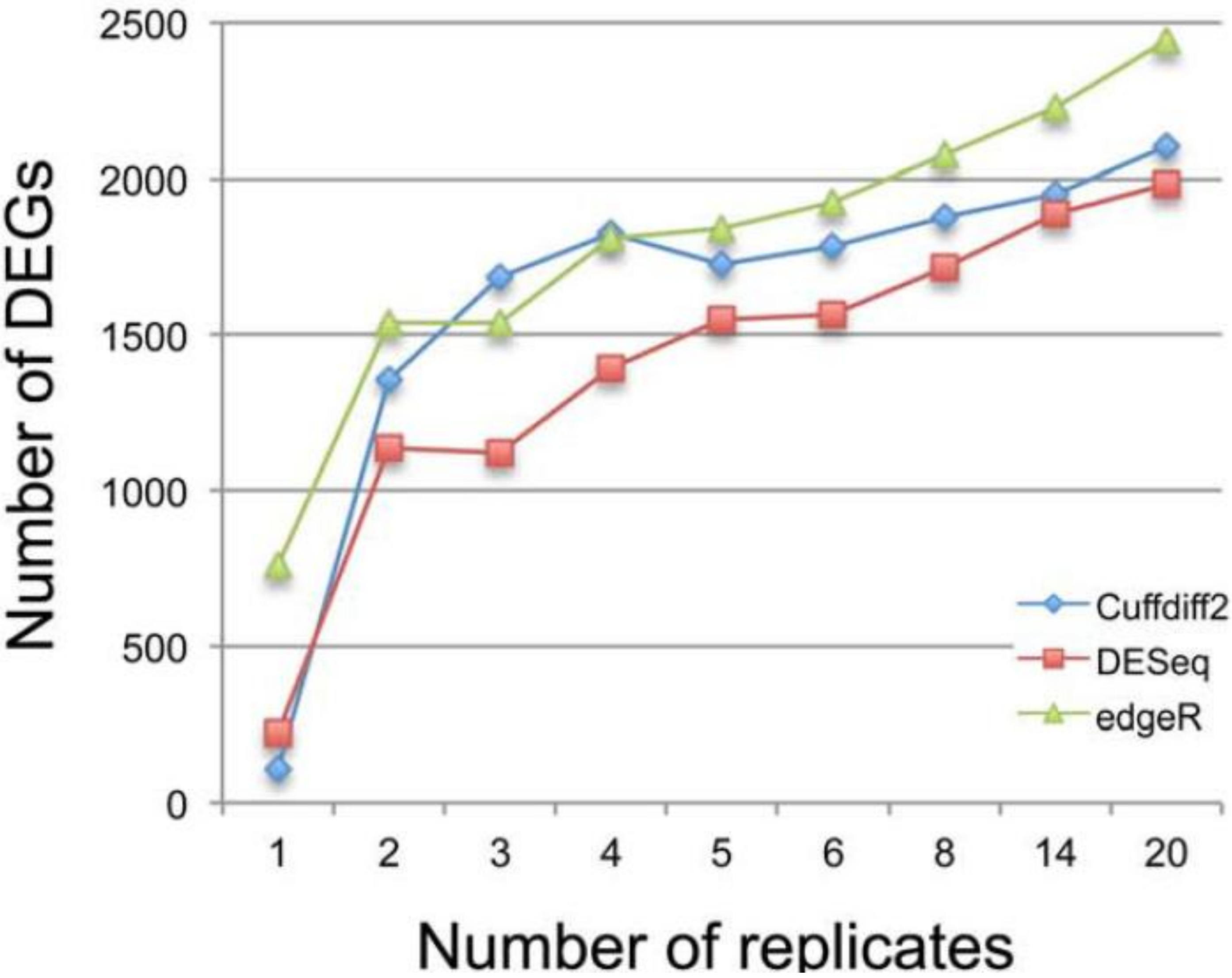
Abstract

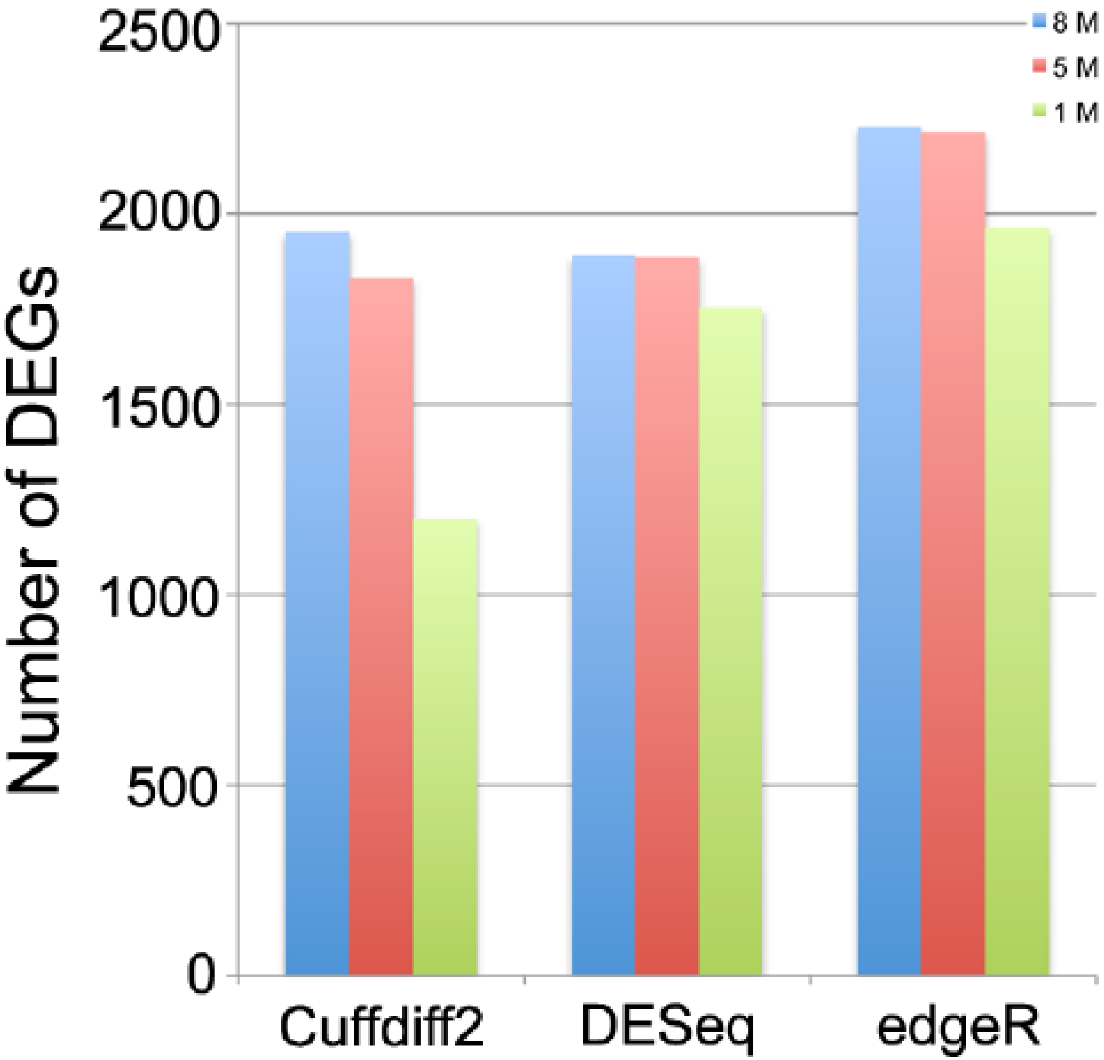
Recent advances in next-generation sequencing technology allow high-throughput cDNA sequencing (RNA-Seq) to be widely applied in transcriptomic studies, in particular for detecting differentially expressed genes between groups. Many software packages have been developed for the identification of differentially expressed genes (DEGs) between treatment groups based on RNA-Seq data. However, there is a lack of consensus on how to approach an optimal study design and choice of suitable software for the analysis. In this comparative study we evaluate the performance of three of the most frequently used software tools: Cufflinks-Cuffdiff2, DESeq and edgeR. A number of important parameters of RNA-Seq technology were taken into consideration, including the number of replicates, sequencing depth, and balanced *vs.* unbalanced sequencing depth within and between groups. We benchmarked results relative to sets of DEGs identified through either quantitative RT-PCR or microarray. We observed that edgeR performs slightly better than DESeq and Cuffdiff2 in terms of the ability to uncover true positives. Overall, DESeq or taking the intersection of DEGs from two or more tools is recommended if the number of false positives is a major concern in the study. In other circumstances, edgeR is slightly preferable for differential expression analysis at the expense of potentially introducing more false positives.

Citation: Zhang ZH, Jhaveri DJ, Marshall VM, Bauer DC, Edson J, et al. (2014) A Comparative Study of Techniques for Differential Expression Analysis on RNA-Seq Data. PLoS ONE 9(8): e103207. doi:10.1371/journal.pone.0103207

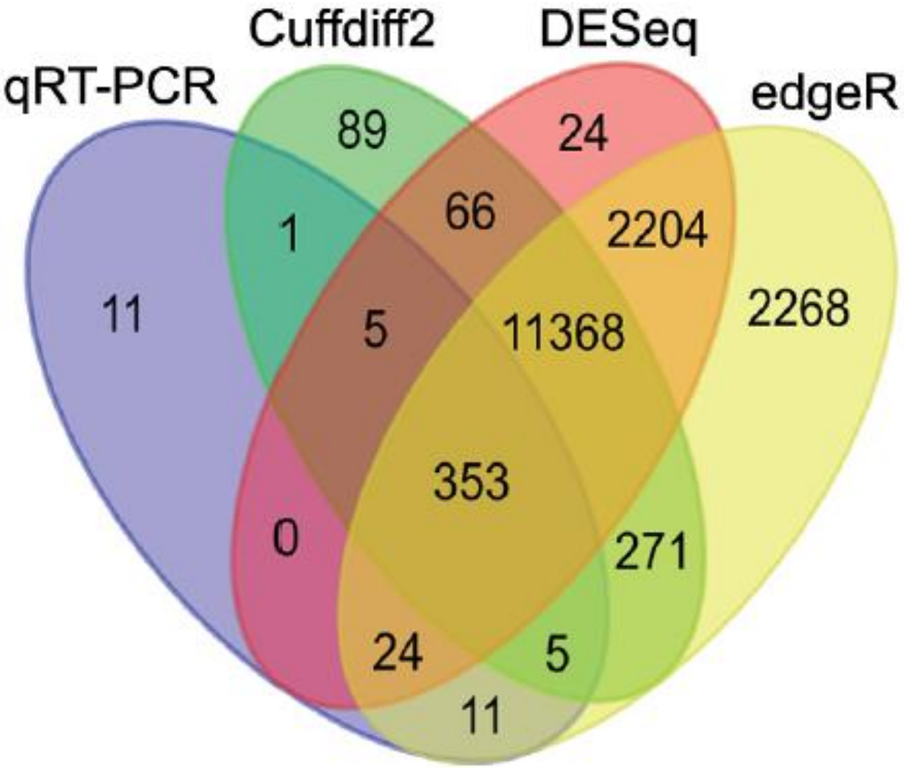
Editor: Paolo Provero, University of Turin, Italy

Received March 4, 2014; **Accepted** June 30, 2014; **Published** August 13, 2014


Number of DEGs
2500
2000
1500
1000
500
0
1
2
3
4
5
6
8
14
20
Number of replicates
Cuffdiff2
DESeq
edgeR



8 M
5 M
1 M
Number of DEGs
2500
2000
1500
1000
500
0
Cuffdiff2
DESeq
edgeR



-All three tools perform much better where there are biological replicates

-Cuffdiff2 is very sensitive to sequencing depth (> 20M for mouse is recomended)

-DESeq is more sensitive to unbalanced sequencing depth (EdgerR worked very well in this situation)

-EdgeR can always detect more DEGs than other two tools, but introduce more false positives

- DESeq or the intersection of two tools is recomended to reduce the number of false positive

END