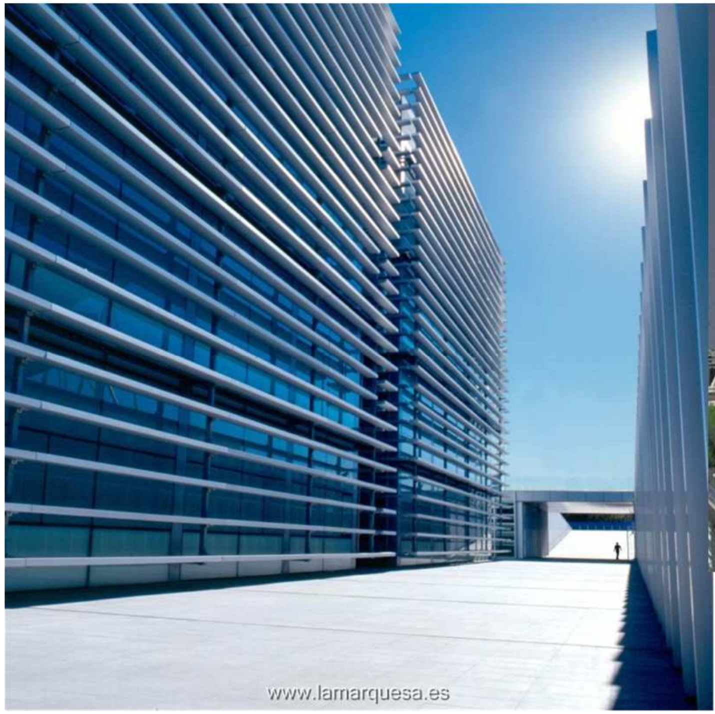
Differential expression with RNA-seq: a matter of depth

Ana Conesa
Genomics of Gene Expression Lab
Centro de Investigaciones Príncipe Felipe
Valencia
aconesa@cipf.es

PRINCIPE FELIPE
CENTRO DE INVESTIGACION

Prince Felipe Research Center



- Research Foundation in Biomedical Sciences
- Since 2004
- 300 scientists
  - Biomedicine
  - Regenerative Medicine
  - Quantitative Biology

Bioinformatics and Genomics Department

PRINCIPE FELIPE
CENTRO DE INVESTIGACION

INB

ciberer

MINISTERIO DE CIENCIA E INNOVACIÓN

F.E.D.E.R.

Proyecto Deméter

Medical Genome Project

ISCB AFFILIATED REGIONAL INSTITUTE

BioinformaticsandGenomics

The Department | Tools | Databases | Publications | Meetings and Courses | Services | Resources | login

Upcoming events

» VII International Course of Massive Data Analysis. Valencia, Spain
Mon, 21/03/2011 (All day)

» Course on Transcriptomic Data Analysis. Cambridge, UK
Wed, 28/09/2011 (All day)

show all events

Latest news

» Senior Bioinformaticist position at Medical Genome Project (Sevilla, Spain)
published on 19/03/2011 - 13:14

» Paintomics: a new web application for visual analysis of transcriptomcis and metabolomics data
published on 09/02/2011 - 13:22

» The MicroArray Quality Control (MAQC)-II study published in Nature Biotechnology
published on 26/09/2010 - 20:06

show all news

Tools usage

Tags in publications

Welcome

Biomedicine can only be understood in the context of genomics and with the concourse of bioinformatics. Our department aims to tackle biomedical problems from a system's biology perspective. Following this, the general objective we seek through the main lines of research is to relate the mutations (Pharmacogenomics and Comparative Genomics) to their effect at cellular and phenotypic level (Functional Genomics) trying to understand the mechanism of action (Structural Genomics).

Functional genomics

Genes operate within an intricate network of interactions that we have only recently started to envisage. Many higher-order levels of interaction are continuously being discovered. In this scenario we are interested in developing methods and tools which can help to understand large-scale experiments from a systems biology perspective.

Comparative genomics

We are interested in the analysis of patterns and processes occurred during the evolution of our genome, and in the application of the evolutionary thought in human health and disease.

- Adaptive Human Evolution
- Evolutionary Pharmacogenetics
- SNP's and Human Disease

Structural genomics

Our Unit aims to develop and apply computational methods for understanding the molecular mechanisms of cell regulation beyond proteins. In particular, we apply our

Search this site:

Search

Try our tools

DBAli

GEPAS

SARA server

BLAST2GO

Pupa Suite

snow

Sponsors

aecid

RNA-seq

OPEN ACCESS Freely available online PLoS one

The Prevalence and Regulation of Antisense Transcripts in *Schizosaccharomyces pombe*

Ting Ni[1,2,9], Kang Tu[1,2,9], Zhong Wang[3], Shen Song[1], Han Wu[1,2], Bin Xie[4], Kristin C. Scott[1]
Yuan Gao[4,5], Jun Zhu[1,2]*

Science. 2008 June 6; 320(5881): 1344–1349. doi:10.1126/science.1158441.

The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing

Ugrappa Nagalakshmi, Zhong Wang, Karl Waern, Chong Shou, Debasish Raha, Mark
‑ rstein, and Michael Snyder
partment of Molecular, Cellular, and Developmental Biology, Program in Computer Science,
partment of Molecular, Biophysics and Biochemistry, Yale University, New Haven, CT 06520

Stem cell transcriptome profiling via massive-scale mRNA sequencing

Nicole Cloonan[1,4], Alistair R R Forrest[1,3,4], Gabriel Kolle[1,4], Brooke B A Gardiner[1], Geo
Mellissa K Brown[1], Darrin F Taylor[1], Anita L Steptoe[1], Shivangi Wani[1], Graeme Bethel
Andrew C Perkins[1], Stephen J Bruce[1], Clarence C Lee[2], Swati S Ranade[2], Heather E Pec
Ke

PROTOCOL

RNA-Seq analysis to capture the transcriptome landscape of a single cell

Fuchou Tang[1], Catalin Barbacioru[2], Ellen Nordman[2], Bin Li[2], Nanlan Xu[2], Vladimir I Bashkirov[2], Kaiqin Lao[2] &
M Azim Surani[1]

LETTERS

Understanding mechanisms underlying human gene expression variation with RNA sequencing

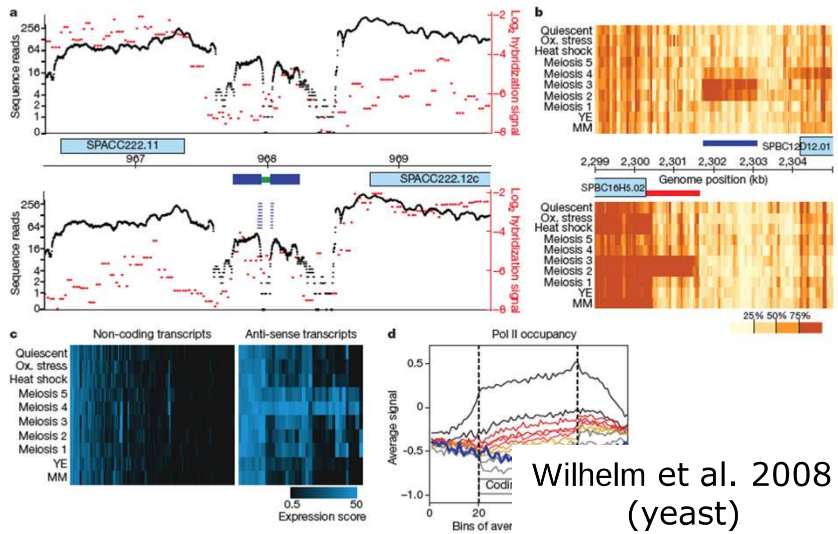
Joseph K. Pickrell[1], John C. Marioni[1], Athma A. Pai[1], Jacob F. De
Jean-Baptiste Veyrieras[1], Matthew Stephens[1,4], Yoav Gilad[1] & J

Nature. 2008 November 27; 456(7221): 470–476. doi:10.1038/nature07509.

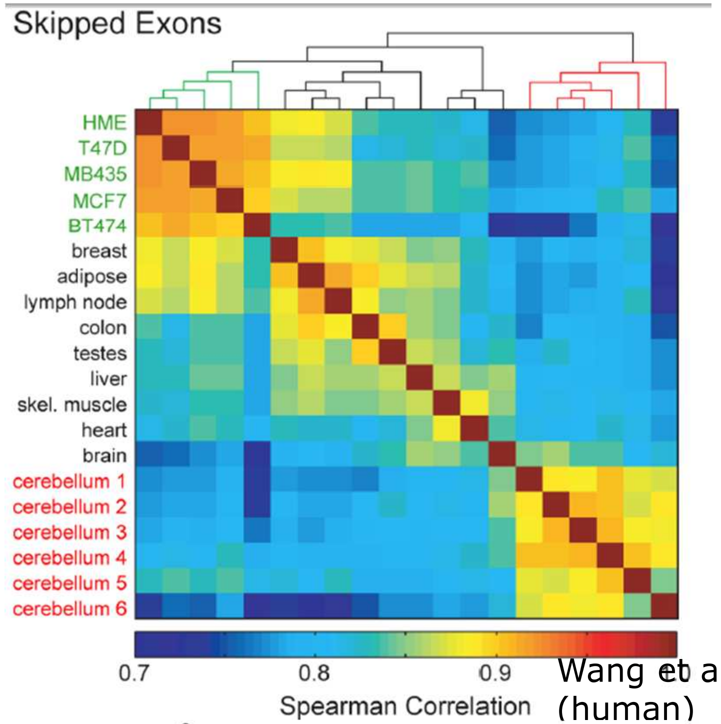
Alternative Isoform Regulation in Human Tissue Transcriptomes

Eric T. Wang[1,2,*], Rickard Sandberg[1,3,*], Shujun Luo[4], Irina Khrebtukova[4], Lu Zhang[4],
Christine Mayr[5], Stephen F. Kingsmore[6], Gary P. Schroth[4], and Christopher B. Burge[1,7]

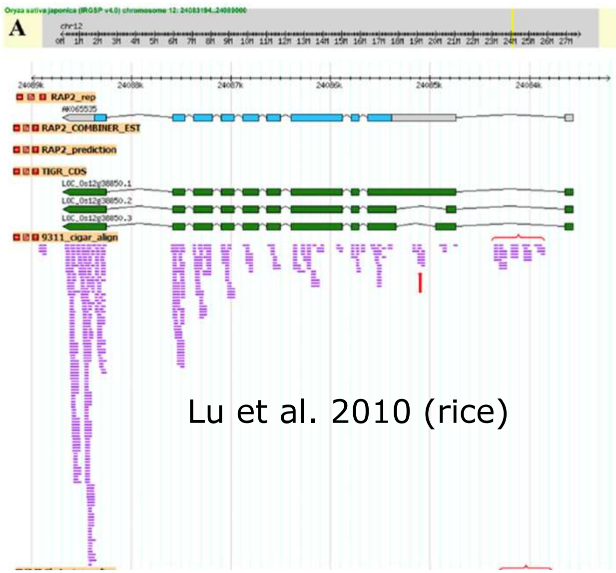
RNA-seq results



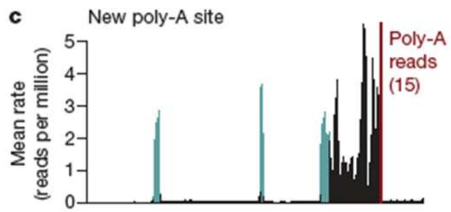
Wilhelm et al. 2008 (yeast)



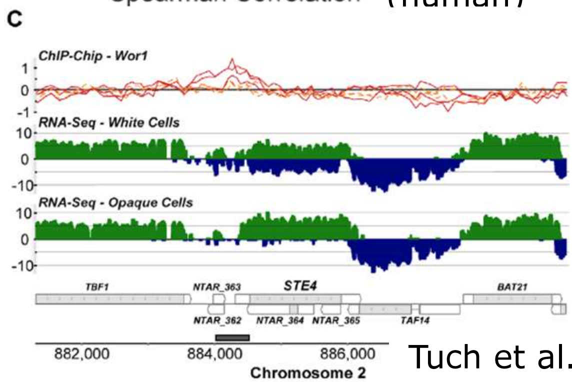
Wang et al. 2008 (human)



Lu et al. 2010 (rice)



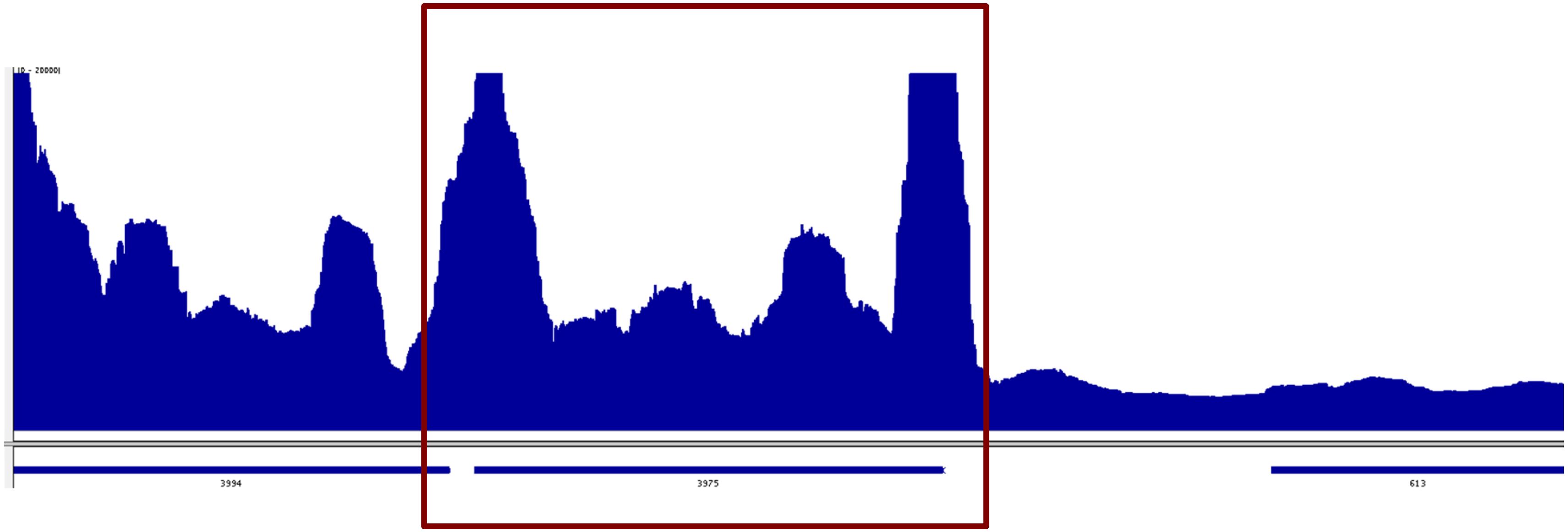
Pickrell et al. 2010 (human)



Tuch et al. 2010 (candida)

RNA-seq quantification

"The number of reads mapped to a gene
is a quantification of its expression"



IGV vew of algU expression in *Pseudomonas aeruginosa*

RNAseq quantification: RPKM

To estimate **expression** value of a transcripts the number of mapped count needs to be **normalized** by the **length** of the transcript and the total number of reads, or **library size.**

RPKM: Reads per Kilobase of exon model per Million reads

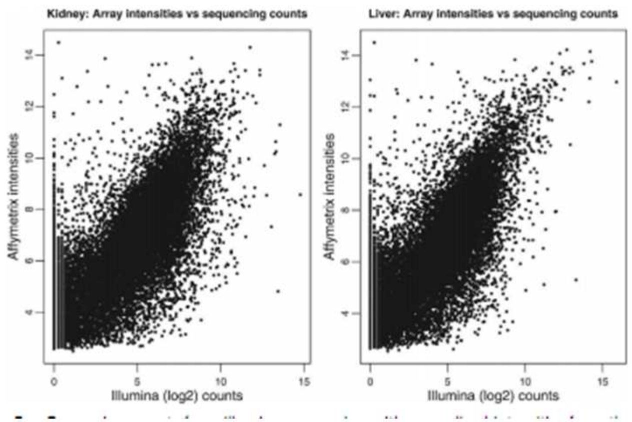
		20 M. reads	27 M. reads			
	Length	**Condition 1**	**Condition 2**	**RPMK1**	**RPKM2**	**Fold-change**
Gene1	1000 nts	700	500	35	18	**2**
Gene2	3000 nts	1000	1800	16	22	**1.5**

RNA-seq for differential expression

Methods

RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays

John C. Marioni,[1,6] Christopher E. Mason,[2,3,6] Shrikant M. Mane,[4] Matthew Stephens,[1,5,7] and Yoav Gilad[1,7]



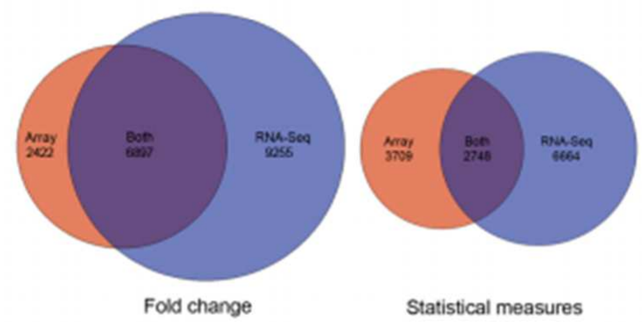
Bradford *et al. BMC Genomics* 2010, **11**:282
http://www.biomedcentral.com/1471-2164/11/282

BMC Genomics

RESEARCH ARTICLE **Open Access**

A comparison of massively parallel nucleotide sequencing with oligonucleotide microarrays for global transcription profiling

James R Bradford[1], Yvonne Hey[2], Tim Yates[1], Yaoyong Li[1], Stuart D Pepper[2] and Crispin J Miller*[1]



Applications of RNA-seq

Qualitative:

* Alternative splicing
* Antisense expression
* Extragenic expression
* Alternative 5’ and 3’ usage
* Detection of fusion transcripts

....

Quantitative:

* Differential expression
* Dynamic range of gene expression

....

Tophat/Cufflinks
Scripture
Alexa

edgeR
DESeq
baySeq
NOISeq

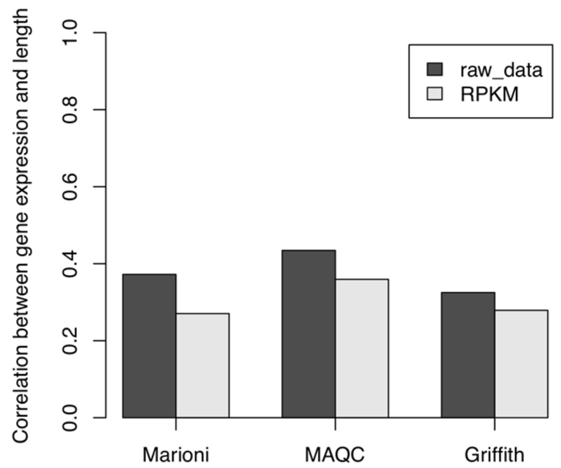
Advantages of RNA-seq?

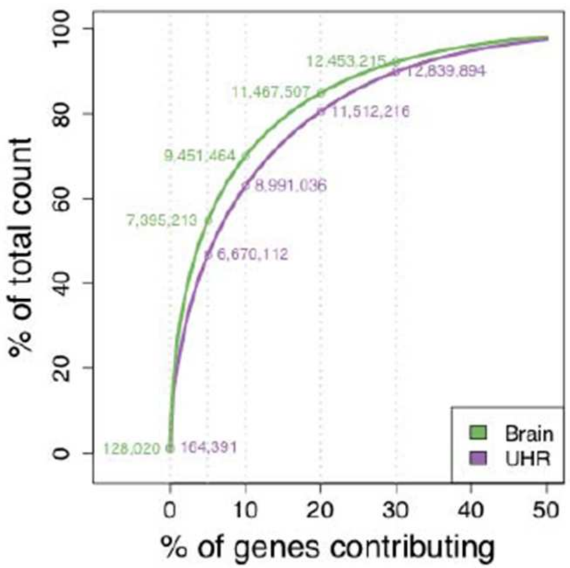
RNAseq	microarrays
* Non targeted transcript detection	* Restricted to probes on array
* No need of reference genome	* Needs genome knowledge
* Strand specificity	* Normally, not strand specific
* Find novels splicing sites	* Exon arrays difficult to use
* Larger dynamic range	* Smaller dynamic range
* Detects expression and SNVs	* Does not provide sequence info
* Detects rare transcripts	* Rare transcripts difficult
….	….

and…. are there any disadvantages?????

Surprises of RNA-seq data

Positive **correlation** between expression level and transcript length. Also with RPKM!!!





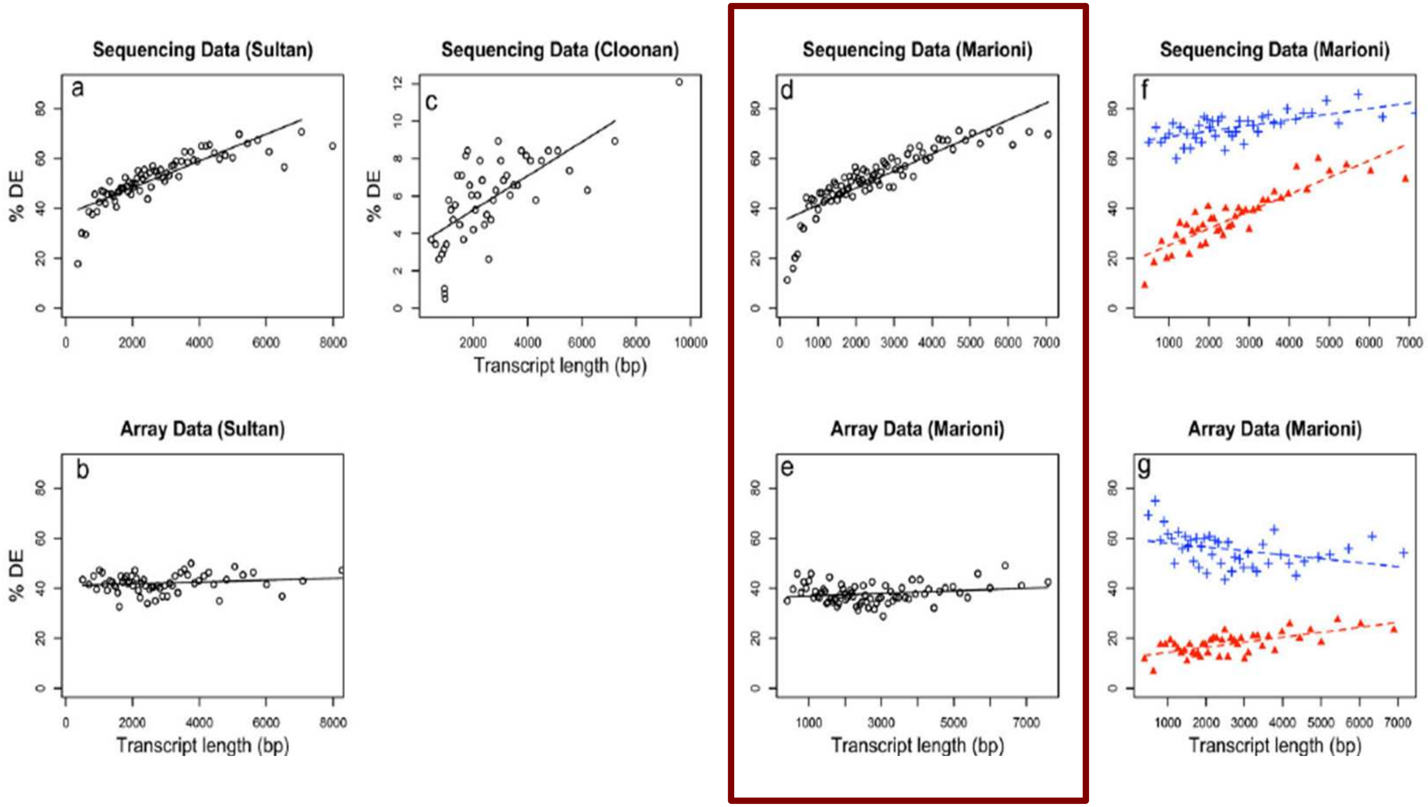
Equal transcript **distributions** between samples do not always hold

Surprises of RNA-seq data

With RNAseq, there is a relationship between the chance that a gene is declared **differentially expressed** and its length

RNAseq

microarrays



RNA-seq and sequencing depth

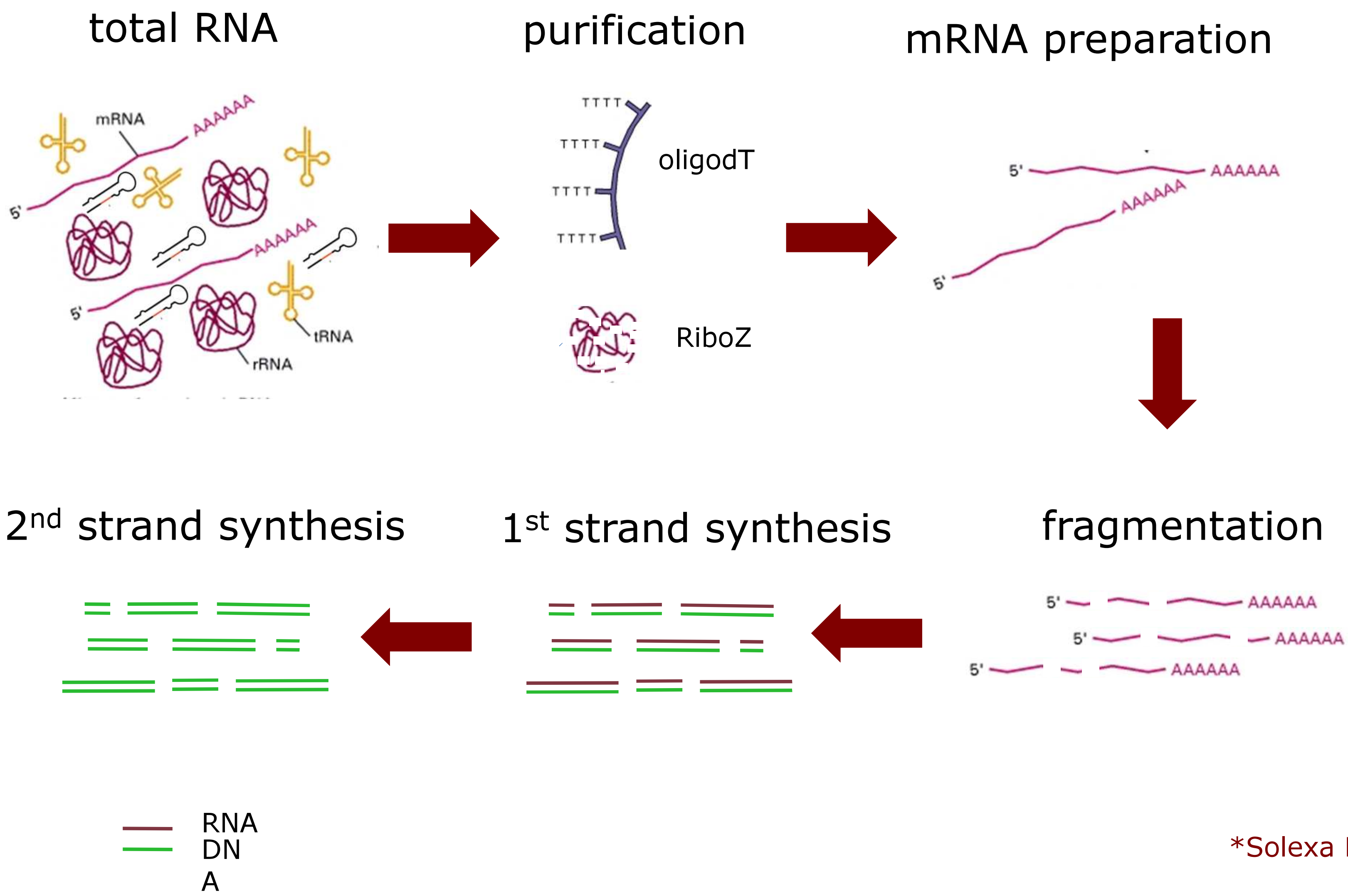
* Amount of reads sequenced in a RNAseq experiment

* More sequencing depth → Rare genes detected
  → Better estimation of expression

* How does SD affects gene detection and differential expression?

* How many reads do I have to generate to saturate the system?

RNA-seq protocol*



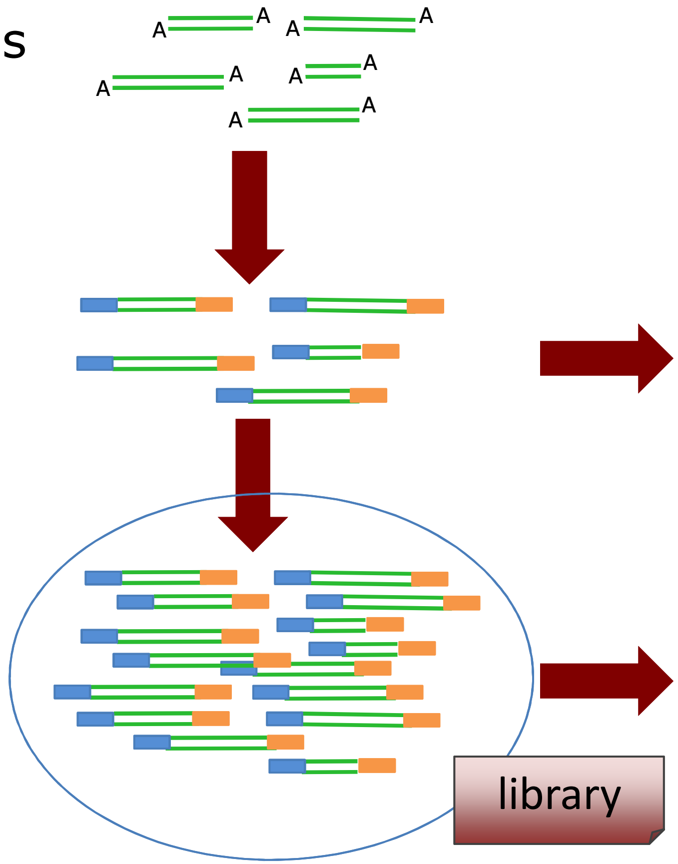
*Solexa Pair-End

RNA-seq protocol (II)

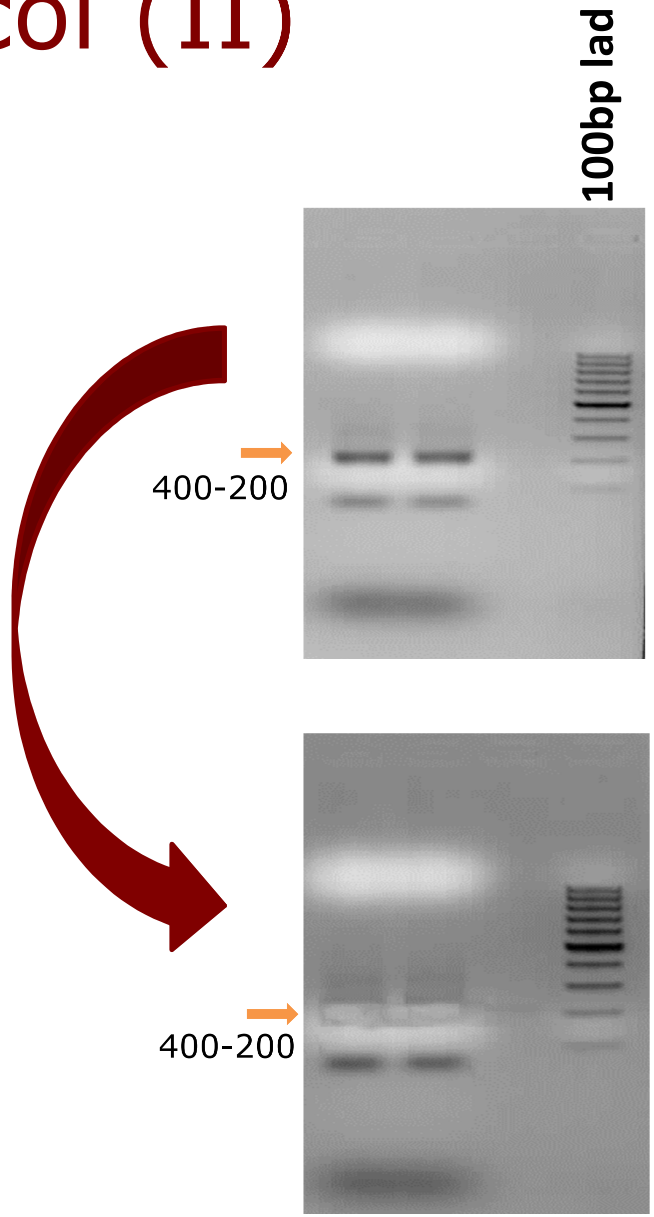
adenylation 3’ ends

ligate adapters

amplification



SEQUENCING!



How does sequencing depth affects to the estimation of differential expression in RNAseq data?

RNA-seq vs. sequencing depth

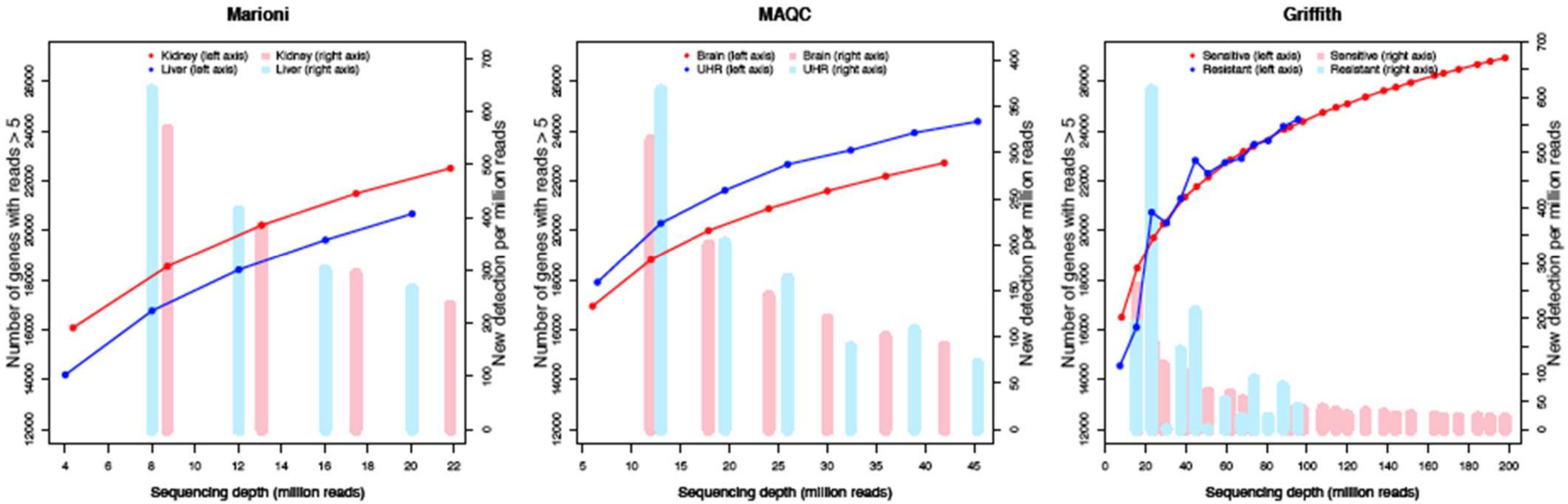
MARIONI:
- Solexa
- 5 lanes
- Kidney vs liver
- 22 million reads

MAQC:
- Solexa
- 7 lanes
- Brain vs UHR
- 45 million reads

Griffith:
- Solexa
- 22 lanes
- 2 cancer lines
- 200 million reads

Saturation in RNA-seq

Saturation Curves and New Detection Rates (NDR)

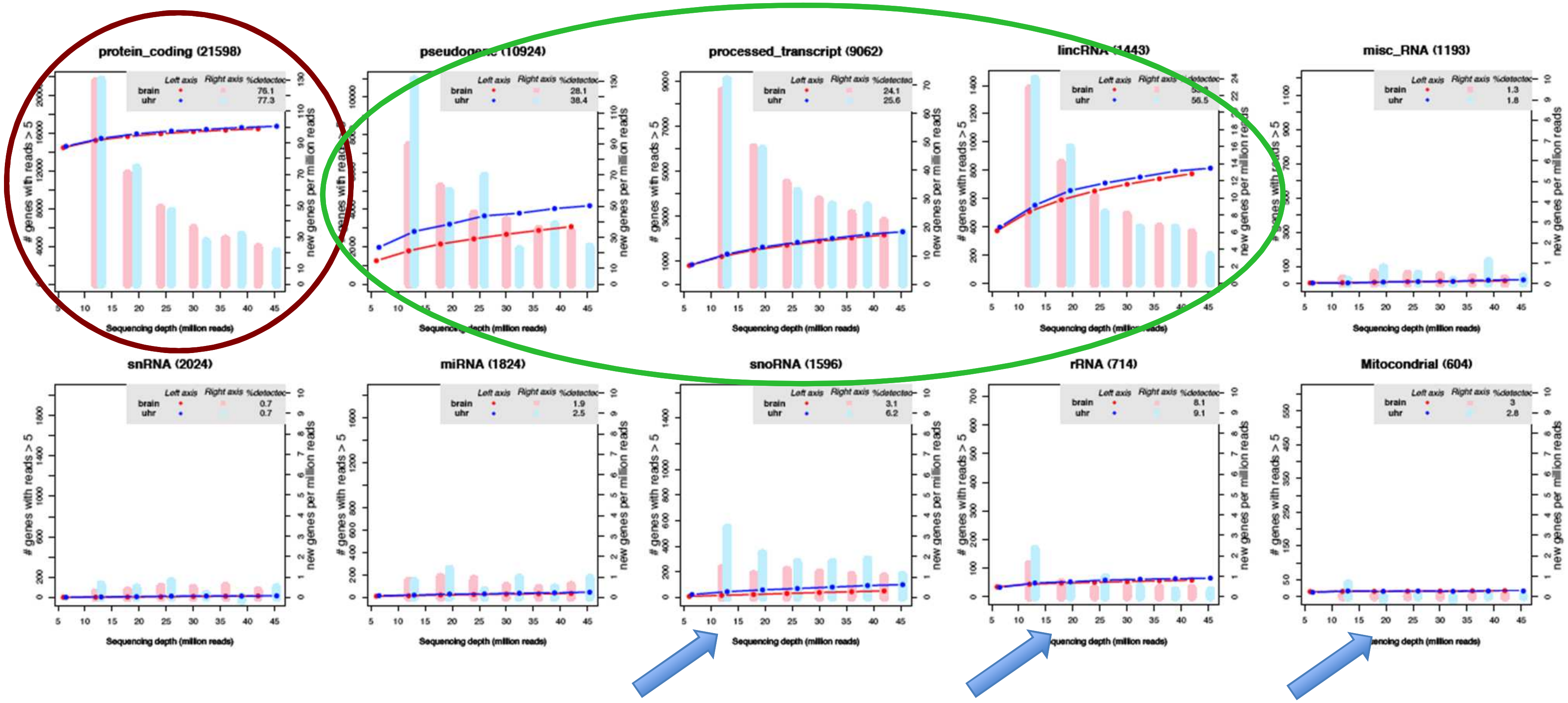


Saturation does not seem to be reached even in large datasets !!!

Saturation per transcript biotype

MAQC (45 M)

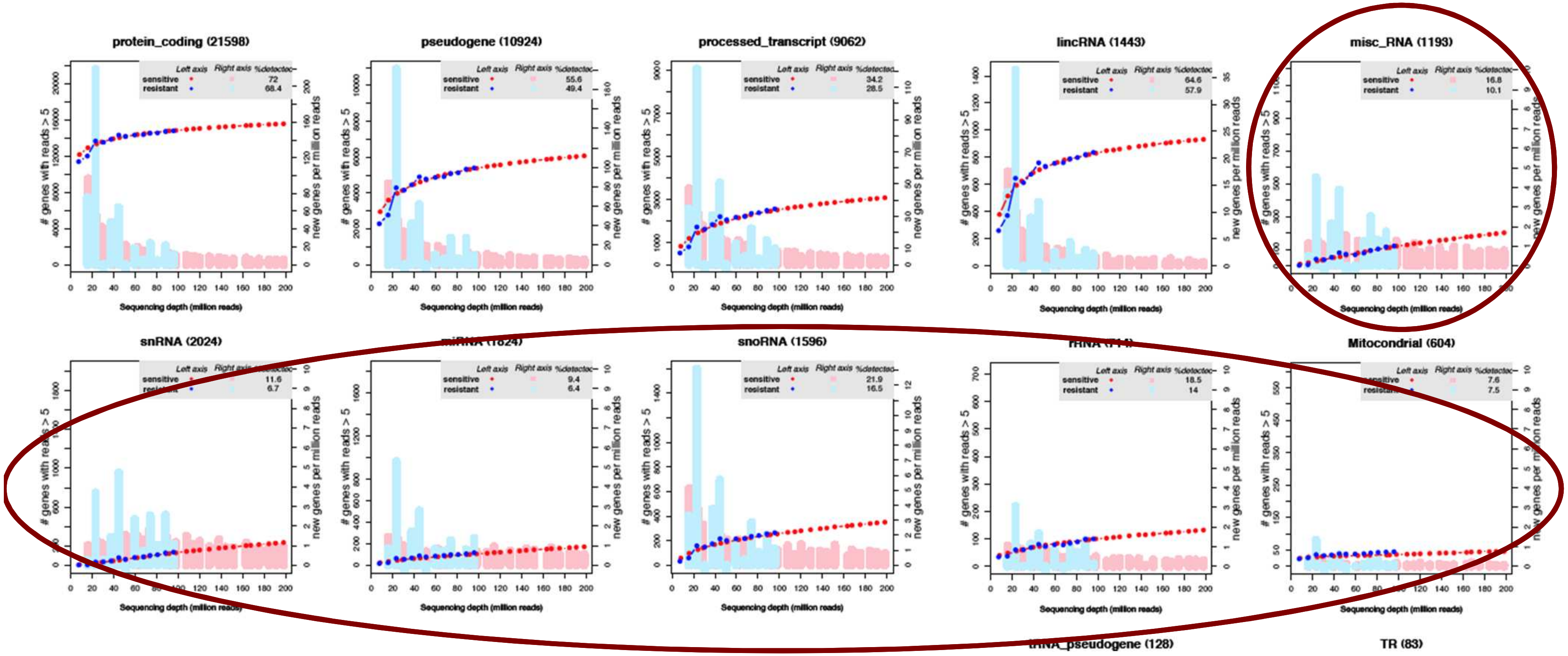
Important expression of Pseudogenes, processed_transcripts and lincRNAs



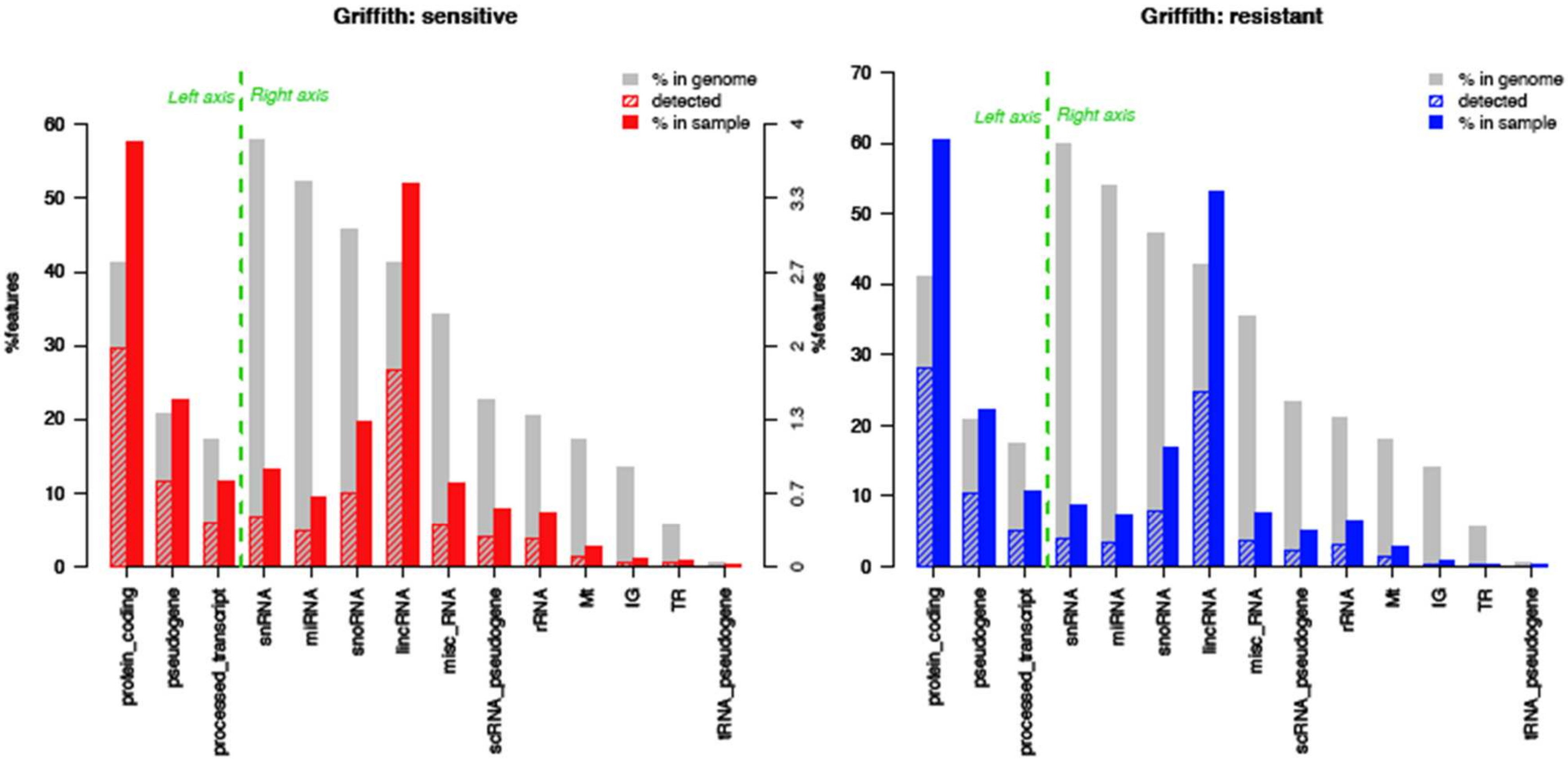
Saturation per transcript biotype

Griffith (200 M)

Off-target RNA species increase at high sequencing depths

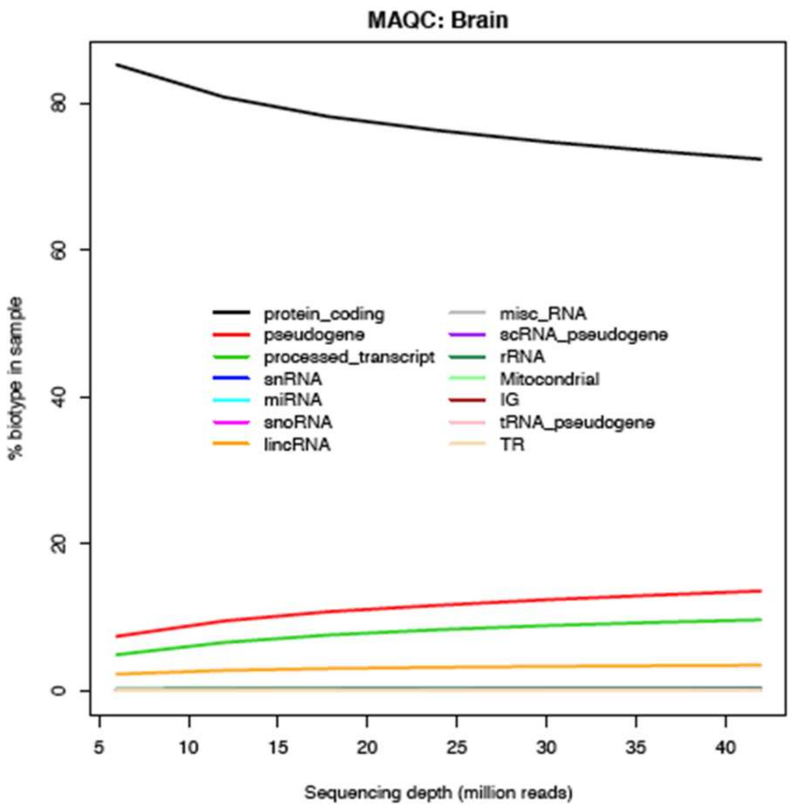


RNA-seq detection per biotype



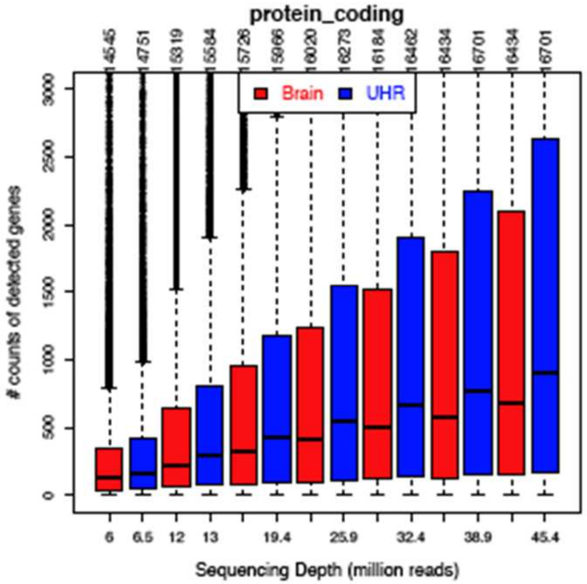
Sequencing depth affects dataset transcript distribution

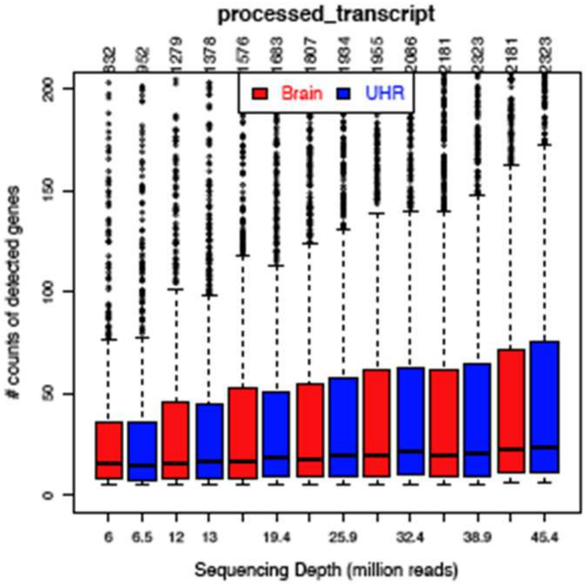
For differential expression comparing samples should have similar library sizes.

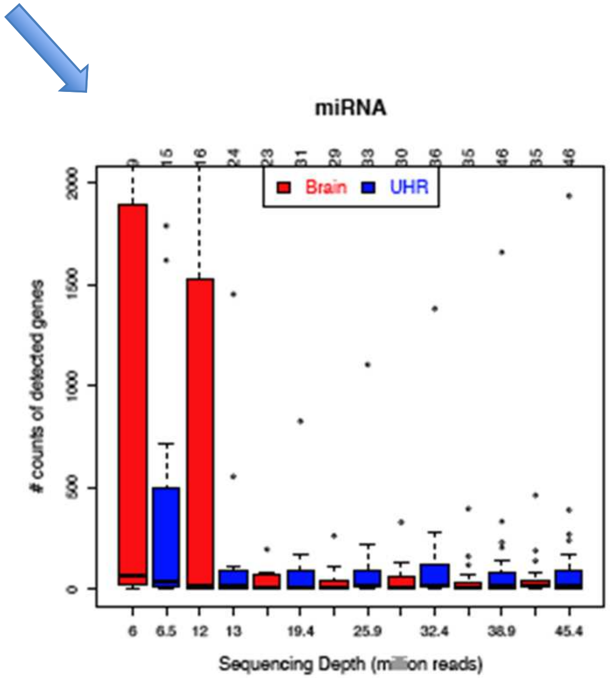


Expression levels increase with sequencing depth at different rates

Few, abundant RNA species sneak into sequencing output!



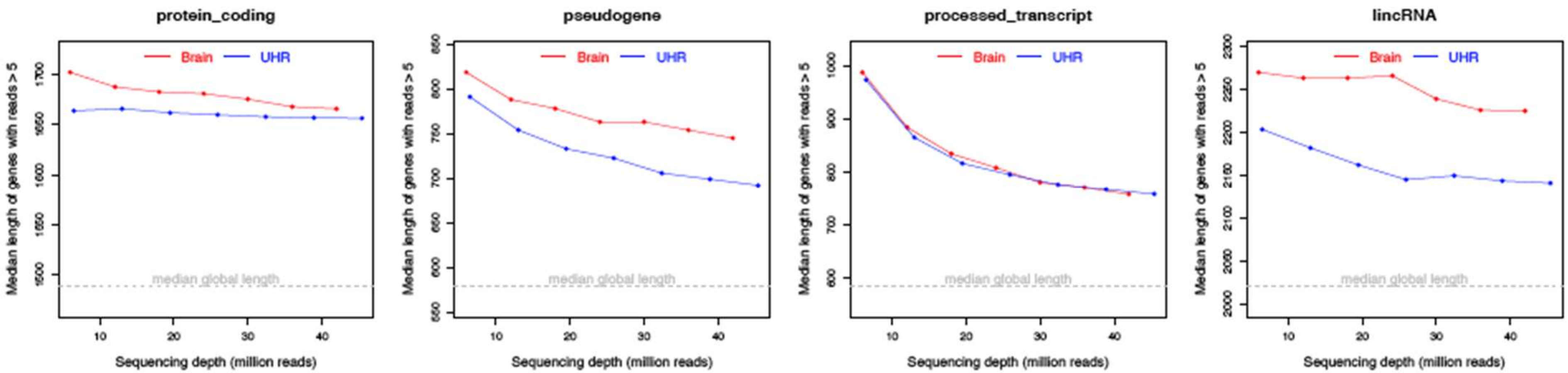




Sequencing depth influences the length of detected transcripts

* As more it is sequenced, small genes are easier detected.

* Still, RNAseq is biased towards longer genes



RNAseq & differential expression

RNAseq & differential expression

* Robinson and Smith (2007, 2008, 2010): **edgeR**

  Exact test based on ***negative binomial*** *distribution.*

* Marioni *et al.* (2008):

  Likelihood ratio test based on ***Poisson model****.*

* Anders and Huber (2010): **DESeq**

  Exact test based on ***negative binomial*** *distribution.*

* *Srivastava and Chen (2010): Gpseq*

  Likelihood ratio test for two-parameter generalized Poisson model.

* Wang *et al.* (2010): DEGseq (MATR & MARS)

  MA-plots based methods, assuming ***normal*** *distribution for M | A.*

* Hardcastle and Kelly (2010): **baySeq**

  Empirical ***Bayesian*** *method to compute posterior probabilities of models, based on Poisson or Negative Binomial data distribution.*

RNAseq & differential expression

* Robinson and Smith (2007, 2008, 2010): **edgeR**

  *Exact test based on **negative binomial** distribution.*

* Marioni *et al.* (2008):

  *Likelihood ratio test based on **Poisson model**.*

* Anders a

  Exac

* *Srivastav*

  Likel *oisson model.*

* Wang *et a*

  *MA-plots based methods, assuming **normal** distribution for M | A.*

* Hardcastle and Kelly (2010): **baySeq**

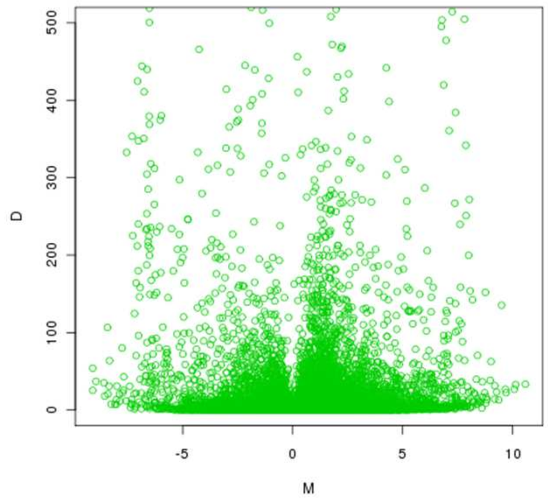
  *Empirical **Bayesian** method to compute posterior probabilities of models, based on Poisson or Negative Binomial data distribution.*

* Parametric assumptions

* Need of replicates

* Unstable with low expressed genes

NOISeq

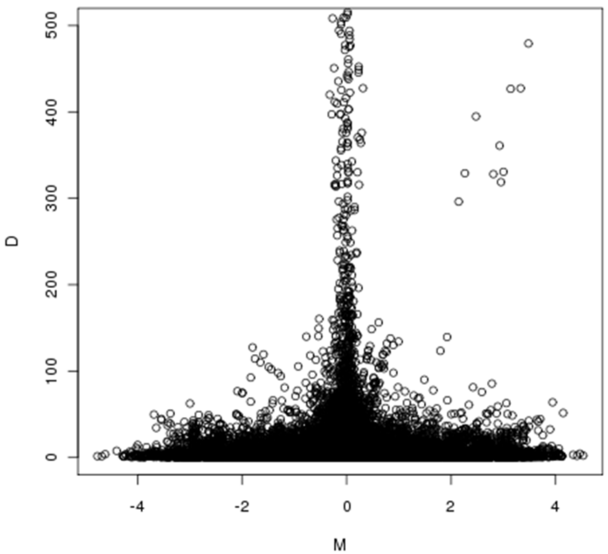


* **No parametric** assumptions. **No** need of **replicates**.

* Statistics for each gene, exon, transcript, tag, etc.:

  M = *log$_2$(expression in condition 1 / expression in condition 2)*

  D = |*expression in condition 1 – expression in condition 2*|



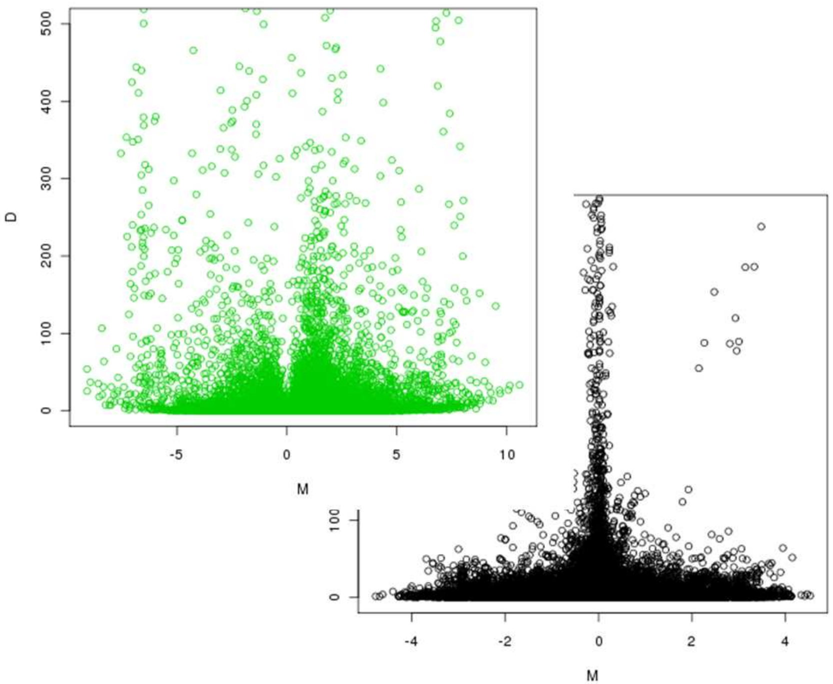
- **Noise distribution**: M-D null distribution estimation.

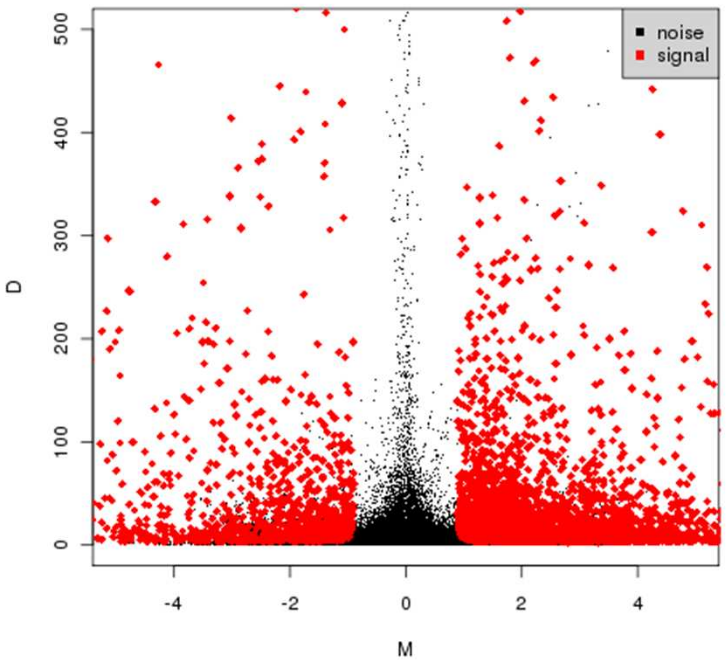
  NOISeq-real: uses available replicates

  NOISeq-sim: simulates replicates from a multinomial distribution with probabilities derived from the counts in the samples

NOISeq

Probability for a gene of being differentially expressed (*deg*):
Computed by comparing M-D values of that gene against noise distribution



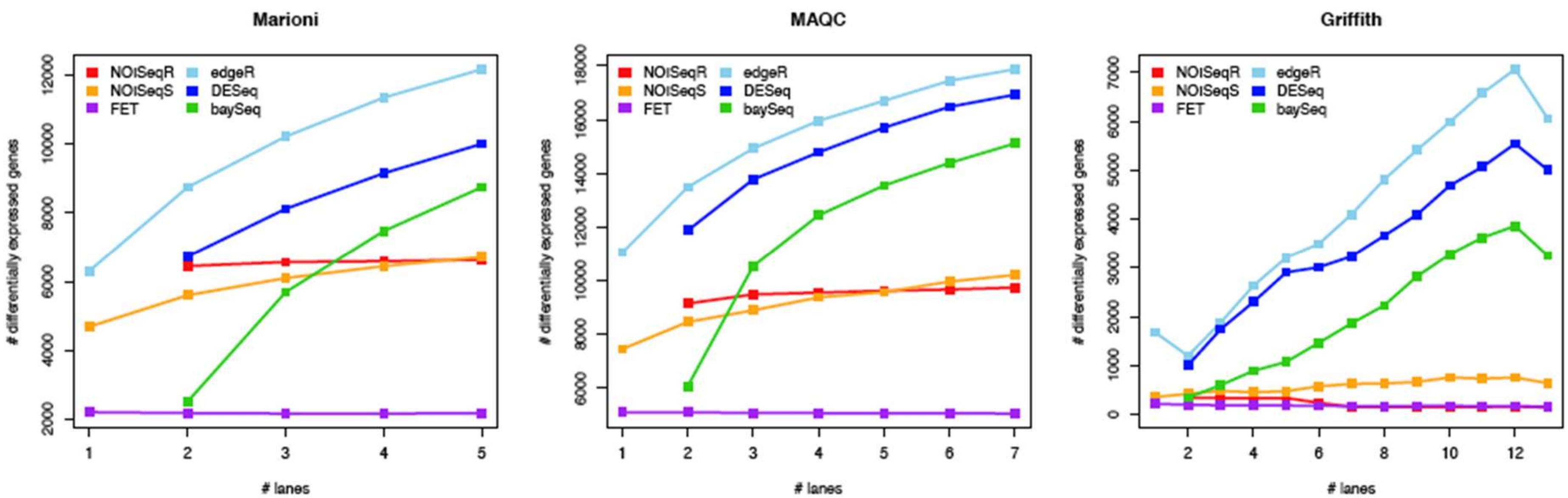


A gene is declared as ***deg*** if this **probability** is higher than **0.8**

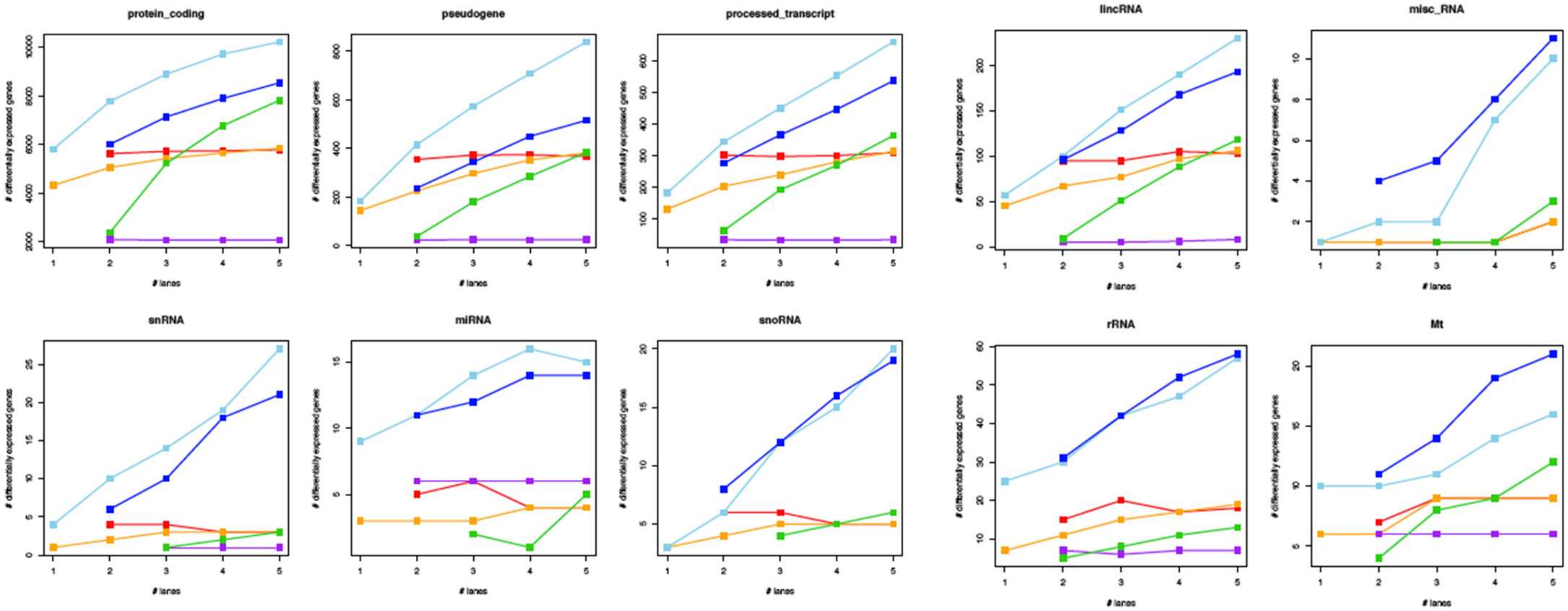
Differential expression vs Sequencing depth

* edgeR, DESeq and baySeq, d.e.g. depend on sequencing depth

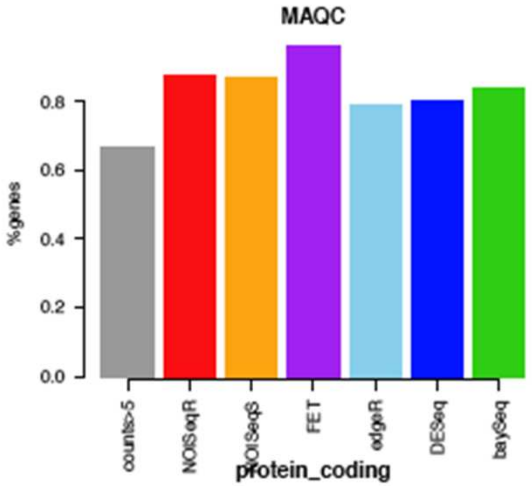
* FET and **NOISeq** are constant

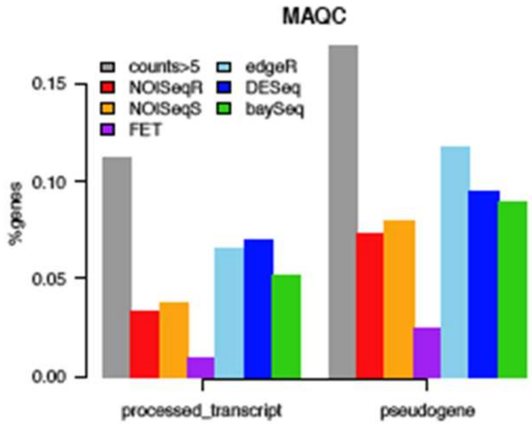


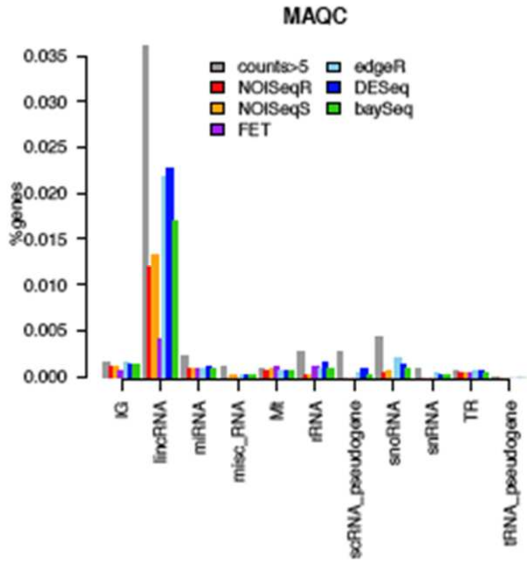
Incremental d.e.g. by biotype



Differential expression by biotype

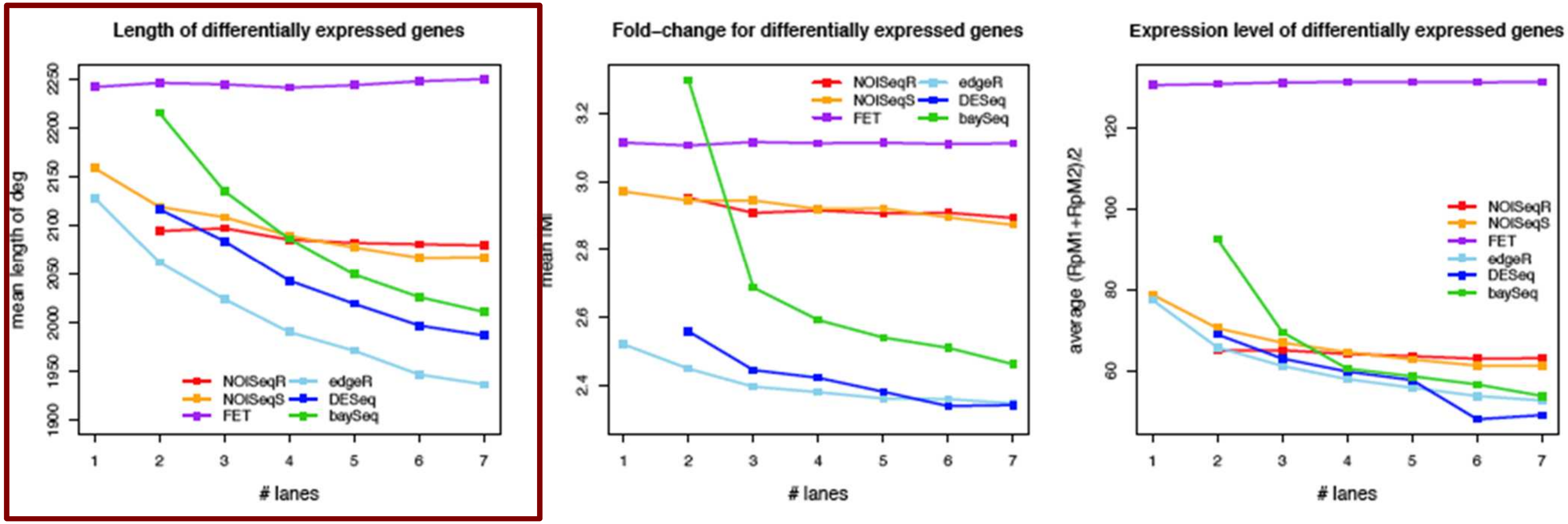






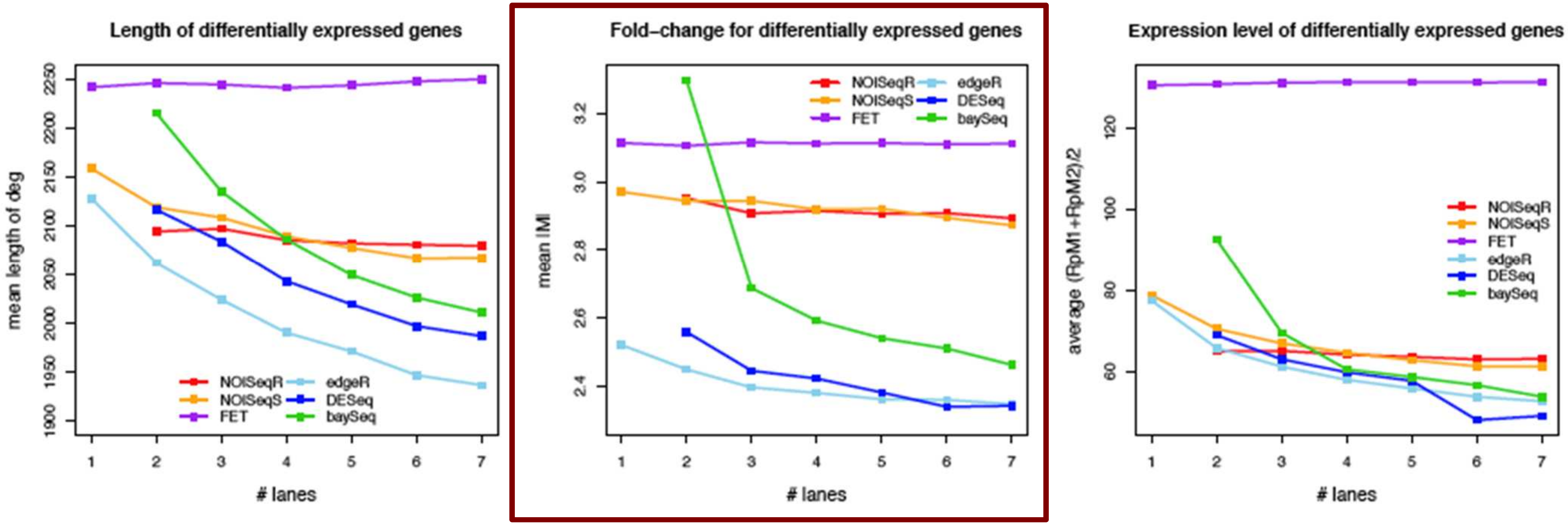
Sequencing depth & characteristics of selected genes

NOISeq is robust to the **length** of detected genes, the **fold-change** of differential expression and the mean **expression level**



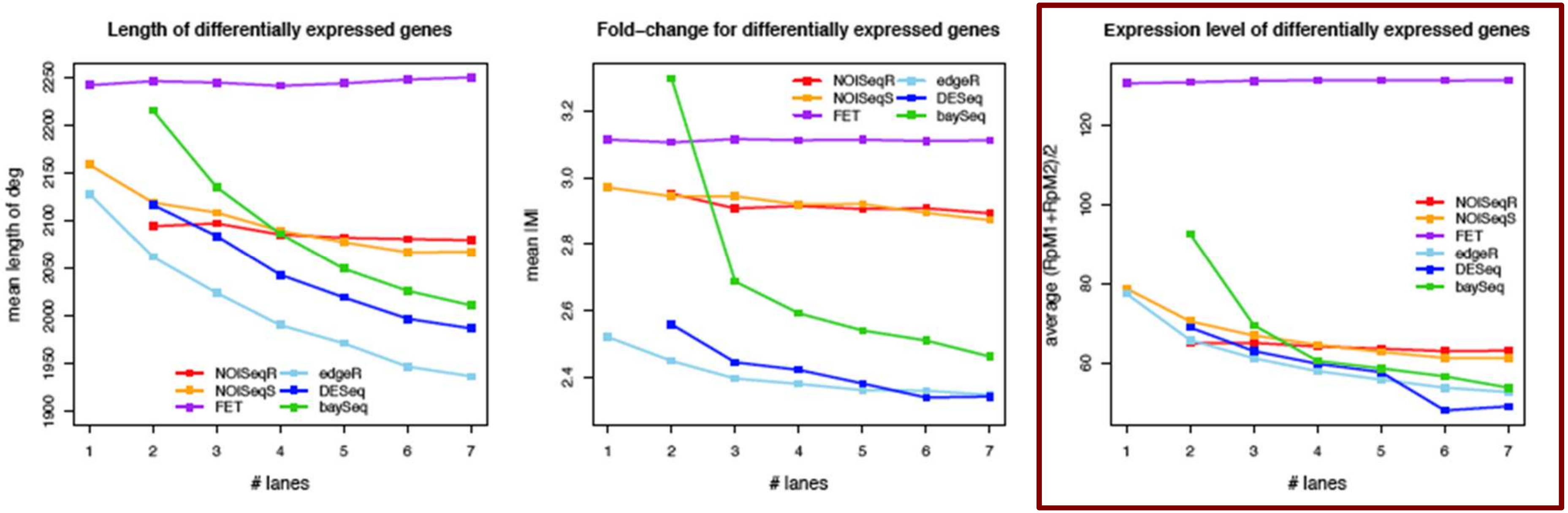
Sequencing depth & characteristics of selected genes

NOISeq is robust to the **length** of detected genes, the **fold-change** of differential expression and the mean **expression level**



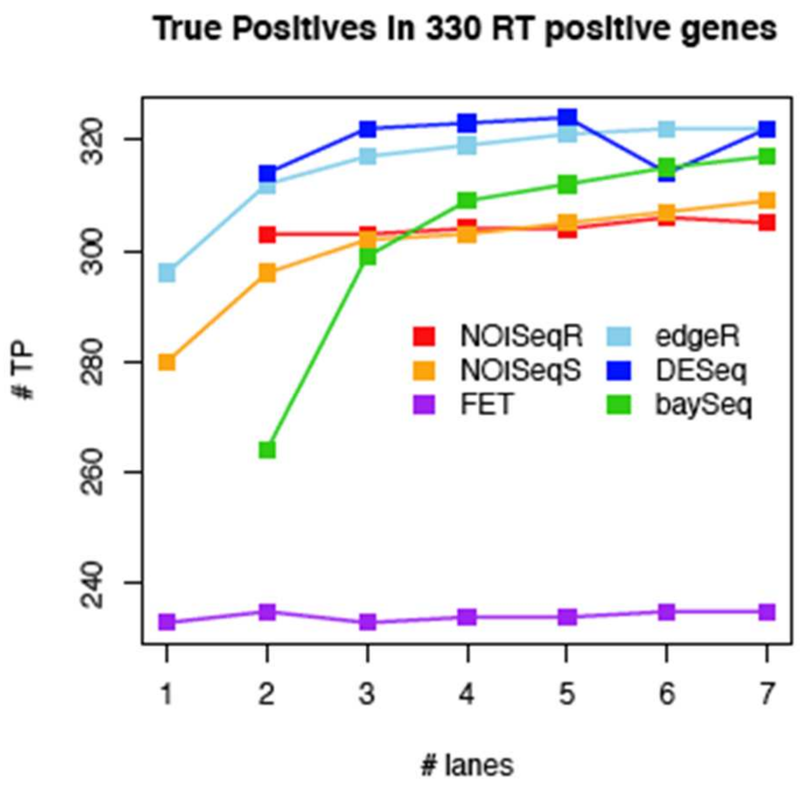
Sequencing depth & characteristics of selected genes

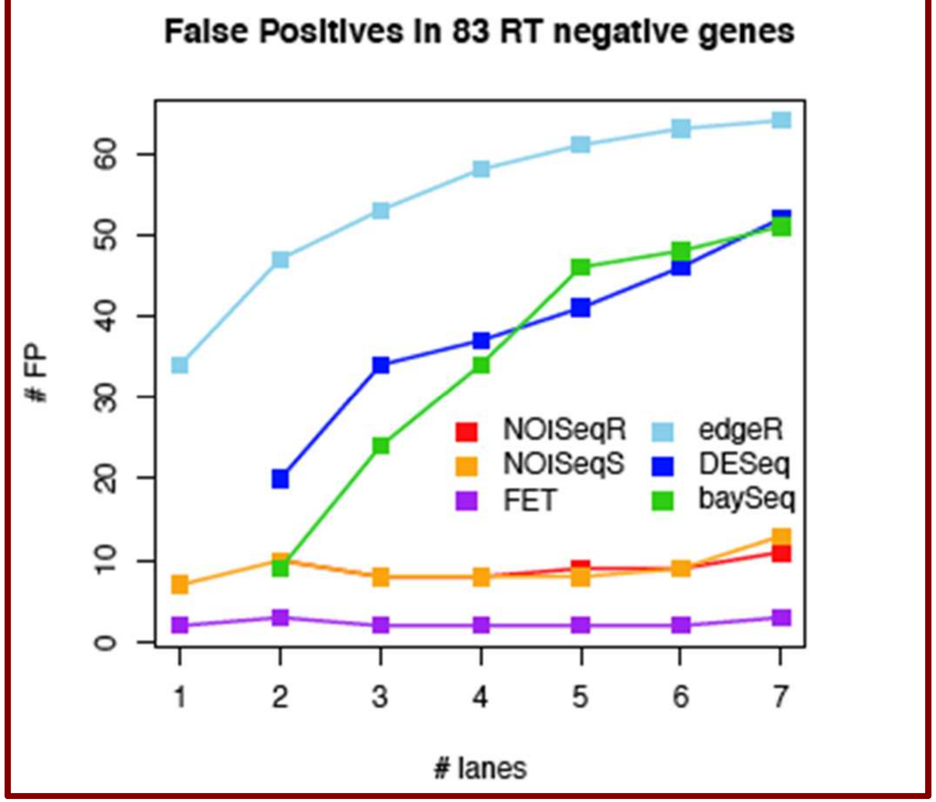
NOISeq is robust to the **length** of detected genes, the **fold-change** of differential expression and the mean **expression level**



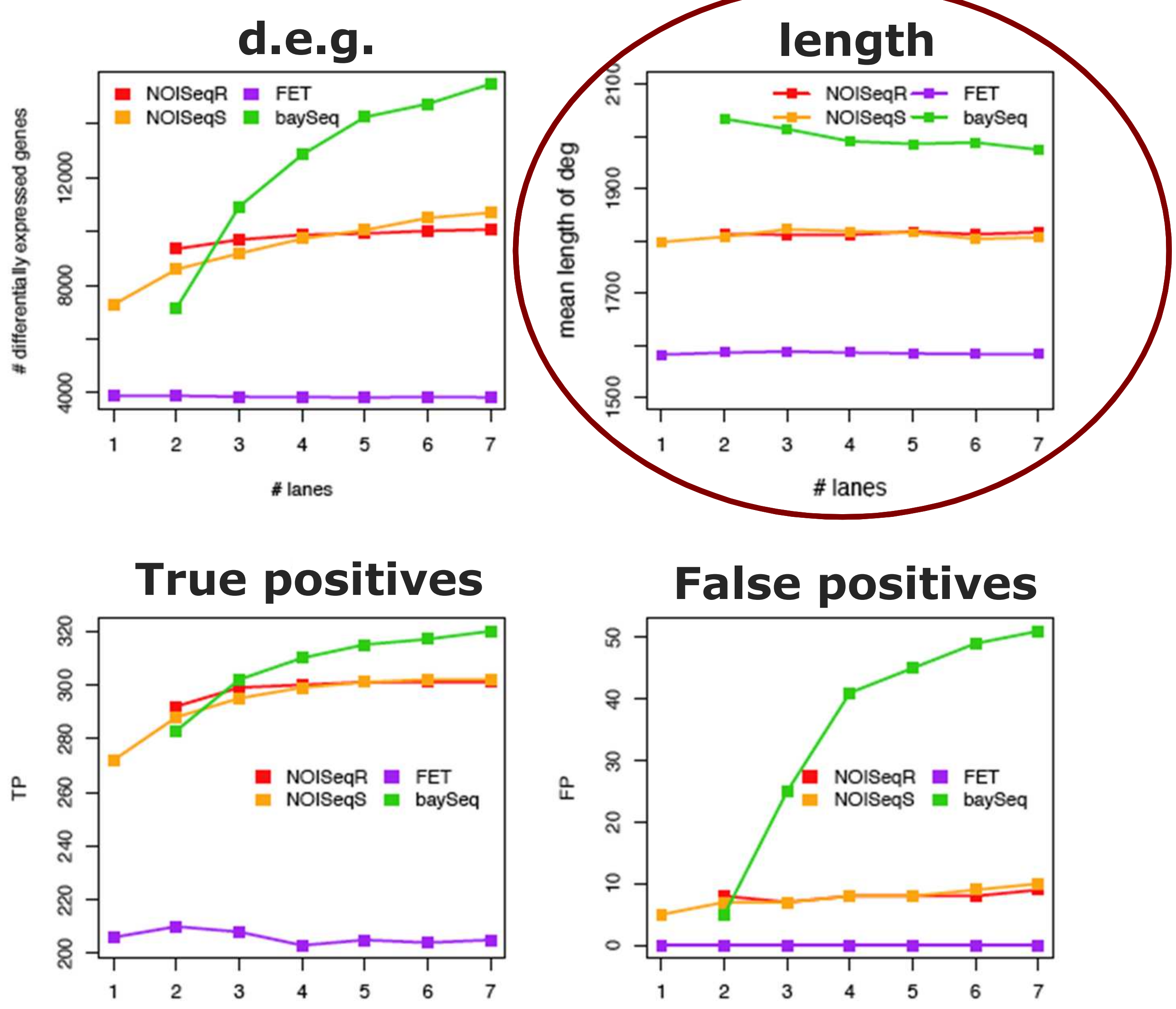
False discoveries at high sequencing depth

Parametric methods tend to identify **more false positives** (up to 70%) as more it is sequenced. **NOISeq** controls FDR





Normalization by length (RPKM) maintain sequencing depth biases



More understanding of RNAseq data

- Sequencing depth affects the **composition** of the RNASeq dataset
- **Short** transcripts are in disadvantage
- Most parametric RNAseq d.e. methods tend to **overdetection** as library size increases.

More understanding of RNAseq data

- Sequencing depth affects the **composition** of the RNASeq dataset
- **Short** transcripts are in disadvantage
- Most parametric RNAseq d.e. methods tend to **overdetection** as library size increases.

- **NOISeq** takes a **non-parametric** approach that better adapts to the noise with large reads numbers.
- NOISeq is robust to sequencing depth biases.

More understanding of RNAseq data

- Sequencing depth affects the **composition** of the RNASeq dataset
- **Short** transcrips are in disadvantage
- Most parametric RNAseq d.e. methods tend to **overdetection** as large library increases.

- **NOISeq** takes a **non-parametric** approach that better adapts to the noise with large reads numbers.
- NOISeq is robust to sequencing depth biases.

- Identification of **low expression genes** with RNASeq is possible but differential expression assessment remains difficult.

Acknowledgements

Sonia Tarazona
Fernando Garcia
Aaron Weimann
Stefan Götz
Samuel Martín
David Jovaní

PRINCIPE FELIPE
CENTRO DE INVESTIGACION

Genomics of Gene Expression