Reducing technical variability and bias in RNA-seq data

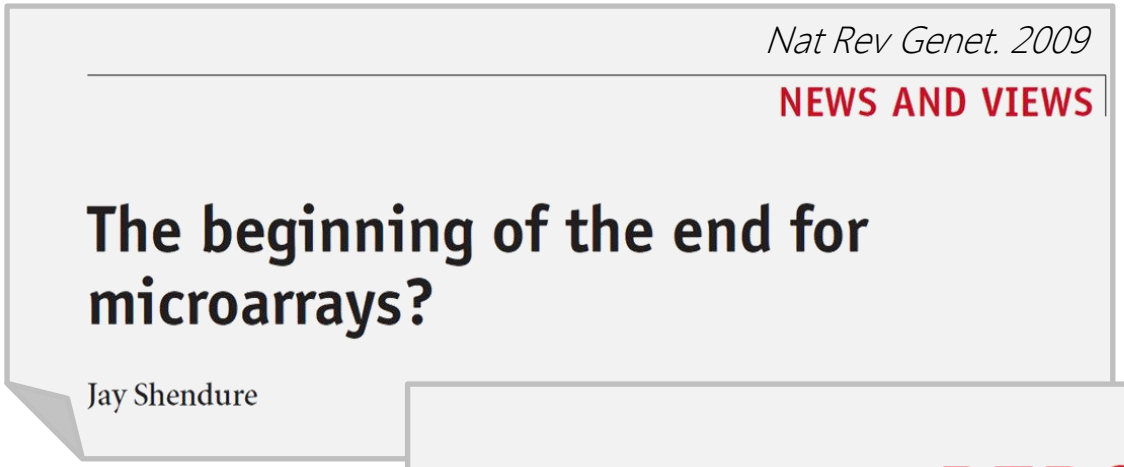
Francesca Finotello

NETTAB 2012

November 14-16 2012, Como, Italy

RNA-seq methodology

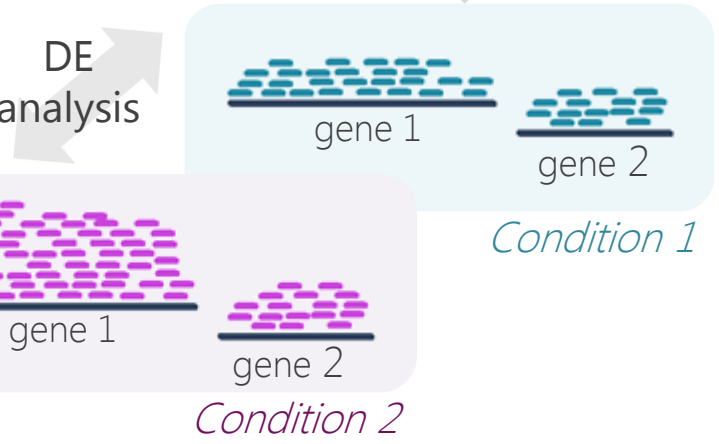
RNA-Seq is a recent methodology (Nagalakshmi, Science 2008) for **transcriptome profiling** that is based on Next-Generation Sequencing



widely adopted in quantitative transcriptomics and seen as a valuable alternative to microarrays

RNA-seq data

	Condition 1	*Condition 2*
gene 1	27	80
gene 2	15	56
...	...	...
gene N	50	20

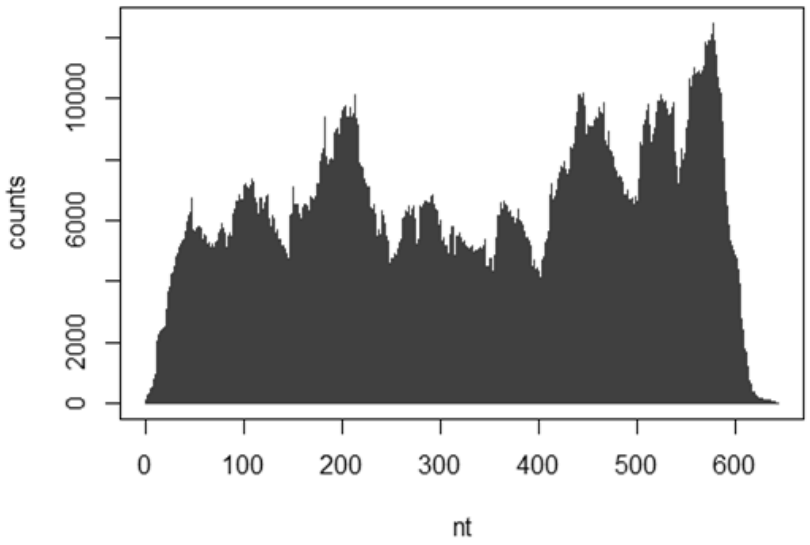


Counts

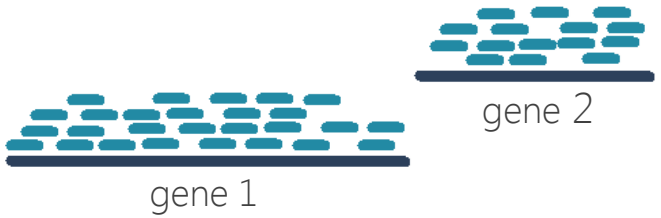
number of reads aligned on a gene
digital measure of gene expression

RNA-seq biases

> RNA-seq [...] can capture transcriptome dynamics across different tissues or conditions **without sophisticated normalization** of data sets.
>
> *- Wang, Nat Methods. 2008*



- Read coverage is not uniform along genes/transcripts
- Different samples can be sequenced at different sequencing depths
- Longer genes are more likely to have higher counts



- Most of reads arise from a restricted subset of highly expressed genes

Outline

- Definition of an alternative approach for computing counts
- Assessement of bias with standard and novel approach
- Evaluation of effects on quantification and differential expression analysis
- Conclusions and future developments

Outline

- Definition of an alternative approach for computing counts
- Assessment of bias with standard and novel approach
- Evaluation of effects on quantification and differential expression analysis
- Conclusions and future developments

New approach maxcounts

- Consider the reads aligned to an exon
- For each exon i, in sample j

  N_{jip} are the number of reads covering exon base p
- ***maxcounts*** are computed as the maximum of per-base counts:

$$M_{ji} = \max\left(N_{jip}\right)$$

Methods

Reads mapped on reference genomes with TopHat, not allowing multiple alignments (`-g 1` option)

Counts *(totcounts)* and per-base counts computed with bedtools (Quinlan, 2010)

maxcounts computed with custom scripts (C++ and Perl)

Differences in sequencing depths corrected via TMM (Robinson, 2010)

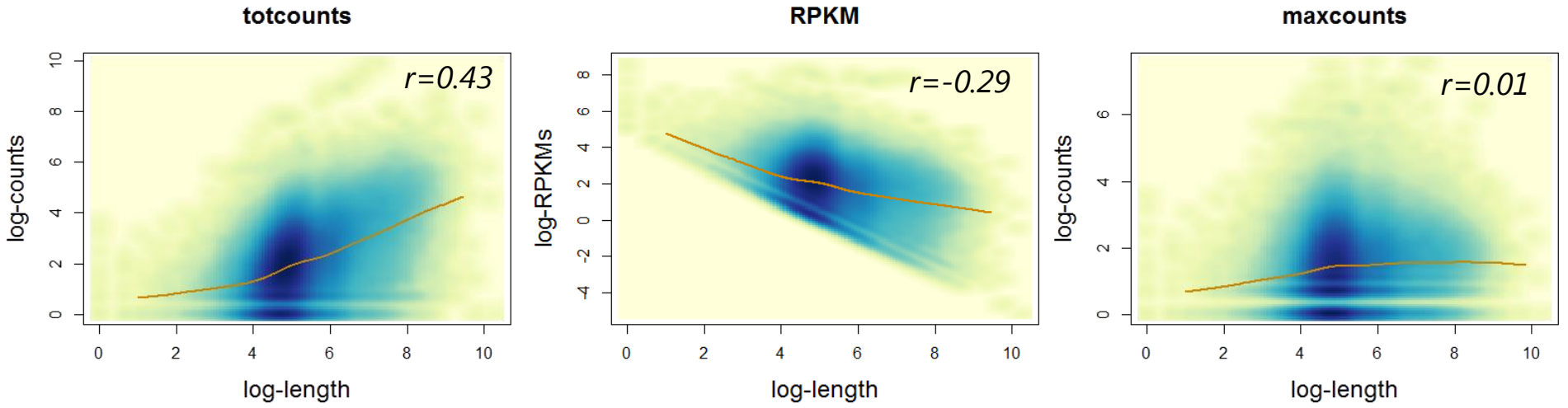
Outline

- Definition of an alternative approach for computing counts
- Assessment of bias with standard and novel approach
- Evaluation of effects on quantification and differential expression analysis
- Conclusions and future developments

Biases exon length

Data set: Griffith, 2010

Smoothed scatter plot of counts vs. exon length (log-log)
Cubic-spline fit of mean log-counts, bins of 100 exons each



		Exp. 1	Exp. 2
e1	[100 bp]	100	80
e2	[95 bp]	120	115
…	…	…	…
e100	[2000 bp]	2120	2000
	∑ counts	15 000	10 000

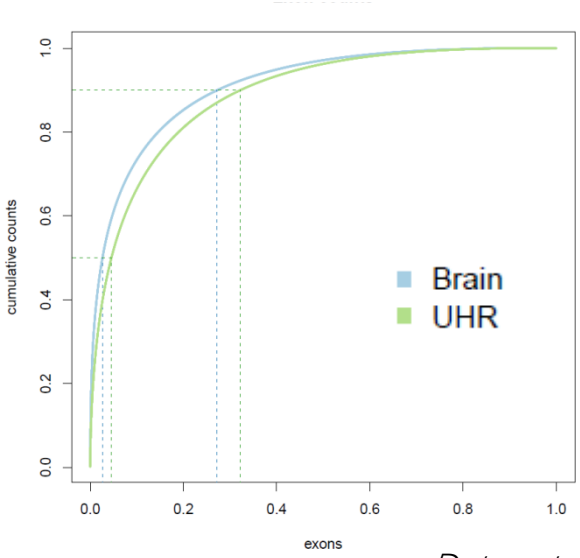
RPKM

Reads Per Kilobase of exon model per Million mapped reads

$$RPKM_{ij} = \frac{N_{ij}}{N_{.j}/_{10^6} \cdot L_i/_{10^3}}$$

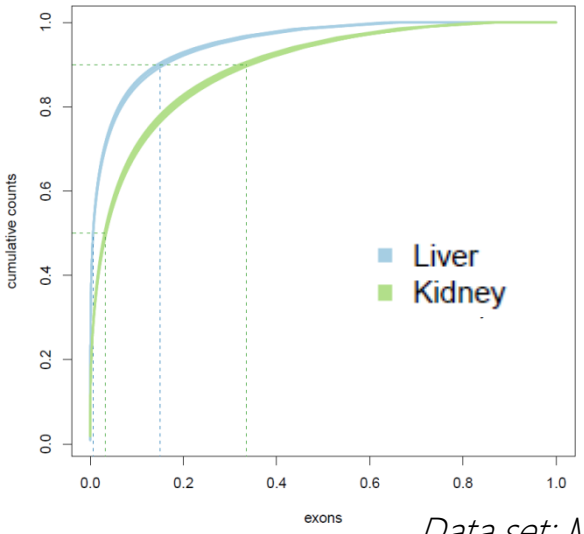
- Length bias also at exon level
- RPKMs overcorrect
- *maxcounts* strongly reduce length bias

Counts distribution across exons



- 3-5% exons contain 50% of counts
- 27-32% exons contain 90% of counts

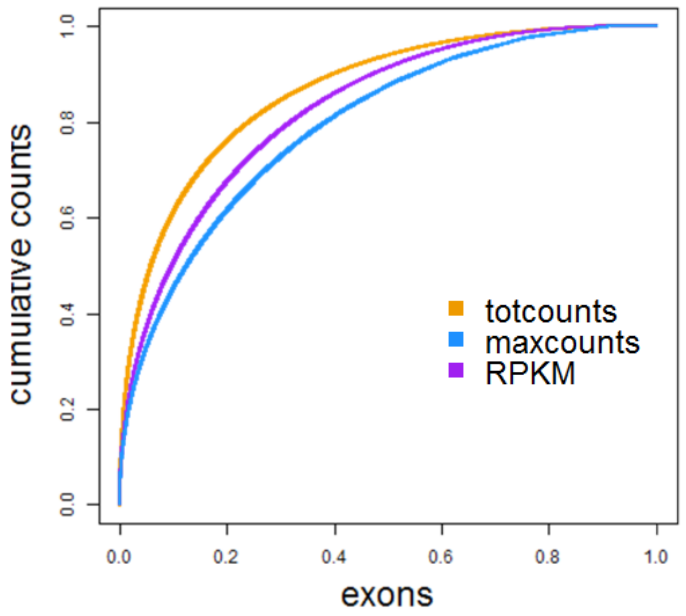
Data set: Bullard, 2010



- 1-3% exons contain 50% counts
- 15-34% exons contain 90% counts

Data set: Marioni, 2008

Data set: Griffith, 2010

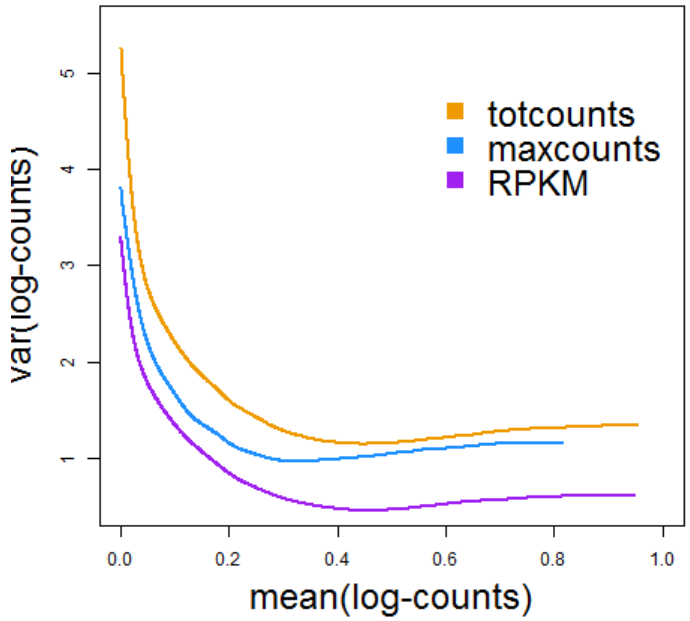


- *maxcounts* have a less steep curve than *totcounts* and RPKMs
- i.e. counts are more evenly distributed across exons

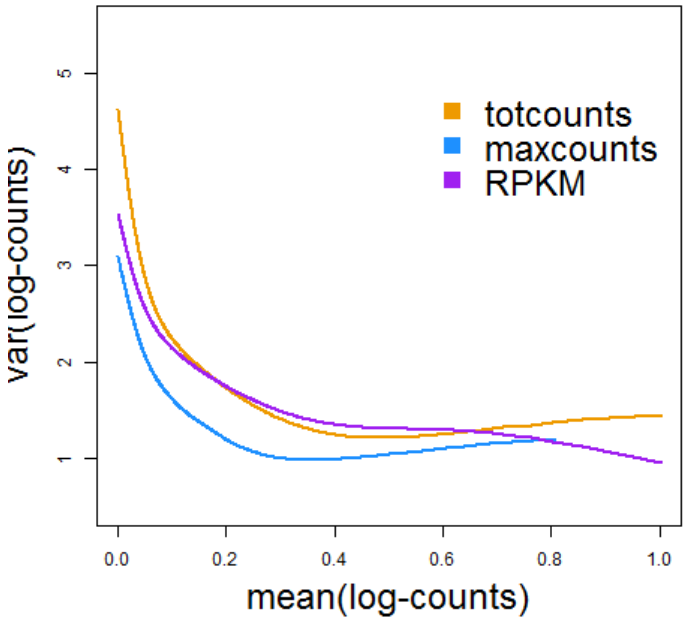
Variance technical replicates

Variance vs. mean of log-counts/RPKMs across technical replicates

Data set: Bullard, 2010



Data set: Griffith, 2010



- *maxcounts'* variance is always lower than *totcounts'* variance
- RPKMs' variance depends on data set
- Assessment on other data sets

Outline

- Definition of an alternative approach for computing counts
- Assessment of bias with standard and novel approach
- Evaluation of effects on quantification and differential expression analysis
- Conclusions and future developments

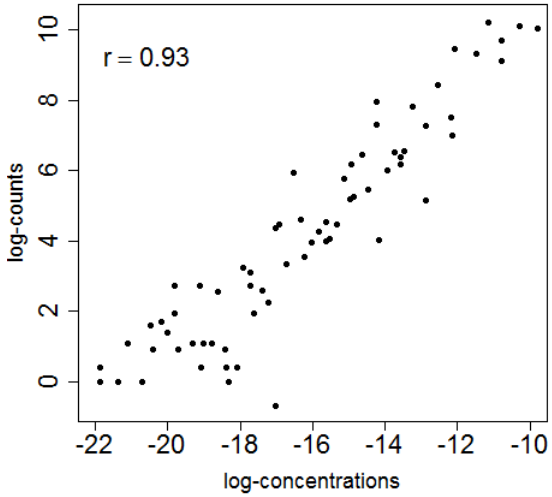
Quantification spike-in RNAs

Data set: Jiang, 2011

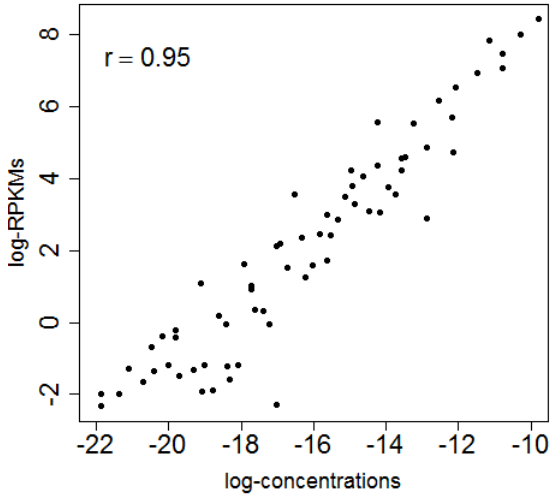
Spike-in RNAs (ERCC Consortium)

- Single-isoforms
- Known sequence and concentration

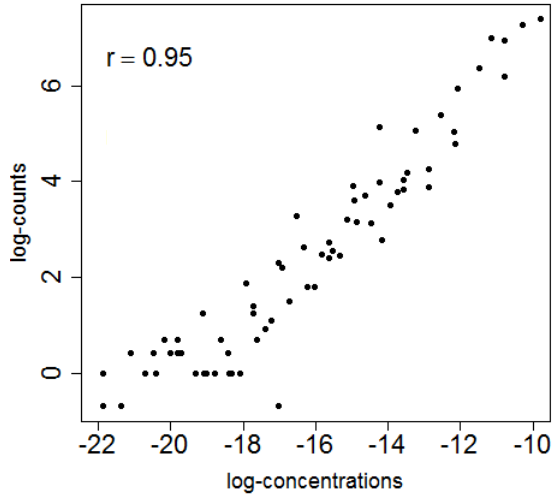
totcounts



RPKMs



maxcounts

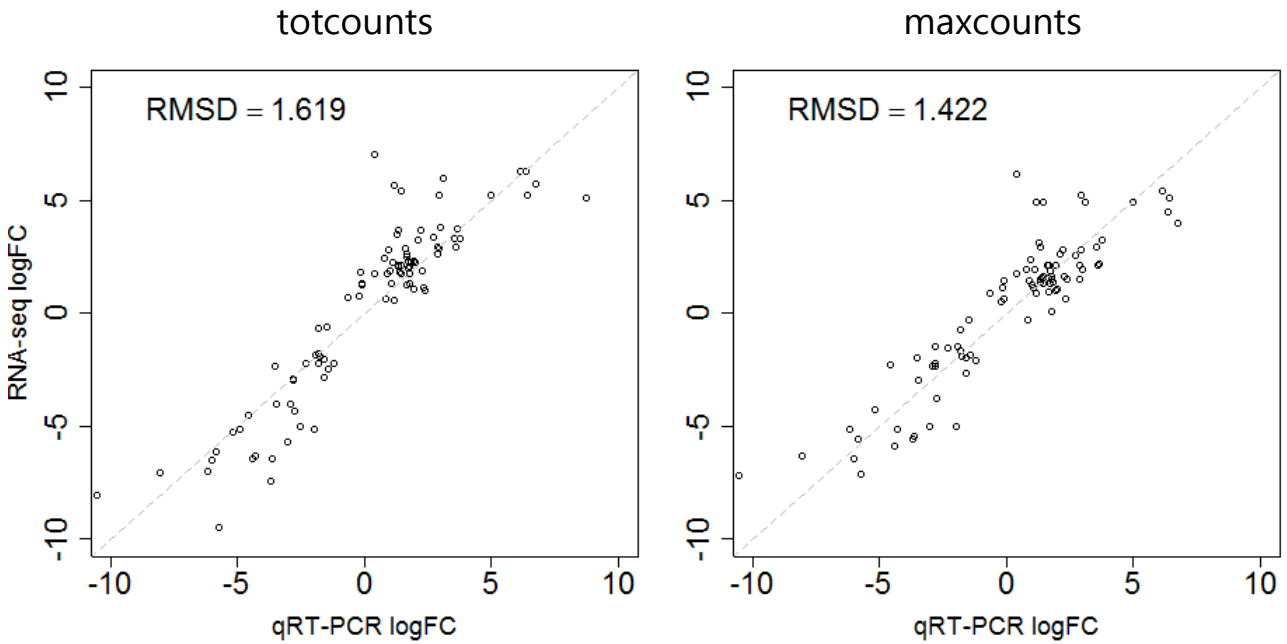


- All measures have high concordance with concentrations
- Transcripts length 270-2000 nt (performance on shorter transcripts?)

DE analysis log-fold-changes

Data set: Griffith, 2010

DE analysis with edgeR (Robinson, 2010) → log-fold-changes (logFC)
Negative Binomial distribution of data required (no RPKMs)



RMSD
Root-mean-square deviation → difference between logFC predicted from *maxcounts* or *totcounts* and from qRT-PCR (gold-standard)

$$RMSD(\hat{\theta}) = \sqrt{E(\hat{\theta} - \theta)^2}$$

maxcounts have a lower RMSD → higher concordance with qRT-PCR

Outline

- Definition of an alternative approach for computing counts
- Assessment of bias with standard and novel approach
- Evaluation of effects on quantification and differential expression analysis
- Conclusions and future developments

Conclusions & future developments

	length bias	count distrib.	tech. variance	spike-in quant.	DE analysis
totcounts (std approach)	-	-	-	+	+
RPKM	+	+	+	++	
maxcounts	++	++	+	++	++

Work in progress and future developments

- Benchmark on more data sets (biological replicates, spike-in RNAs)
- Use other DE methods downstream
- Aggregate exon *maxcounts* to have a measure at gene/transcript level
- Define a robust pre-processing pipeline to avoid artifacts
- Develop an alternative strategy for computing *maxcounts* and implement all versions in a bedtools module

UNIVERSITÀ DEGLI STUDI DI PADOVA

FONDAZIONE EDMUND MACH

ISTITUTO AGRARIO
DI SAN MICHELE ALL'ADIGE

Aknowledgements

Enrico Lavezzo

Luisa Barzon

Stefano Toppo

Paolo Fontana

Paolo Mazzon

Barbara Di Camillo