Reliability and interpretation of RNA-Seq expression profiles

Paweł P. Łabaj

Chair of Bioinformatics, Department of Biotechnology

Boku University Vienna

www.bioinf.boku.ac.at

chair of bioinformatics

Talk outline

Assessing measurement precision

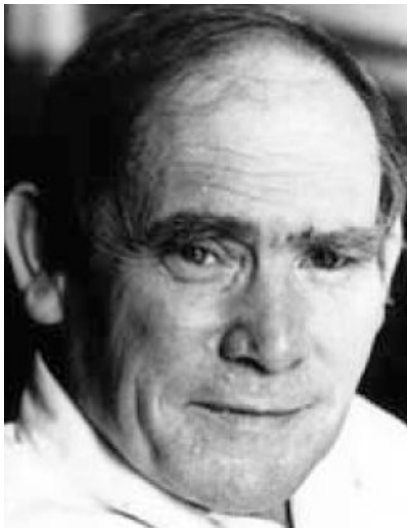
Characterization of established RNA-Seq pipelines

Beyond established approaches

Highly expressed transcripts and read depth

Studying differential signal readout by spike-in mixtures

NGS – 'new' measurement technology

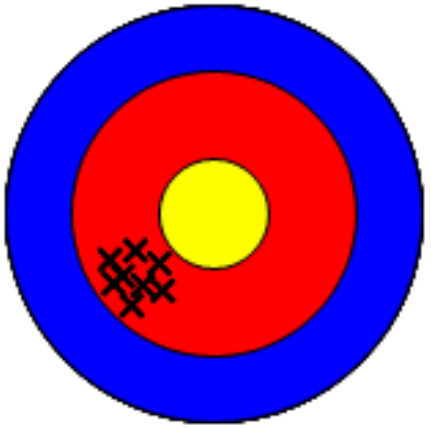
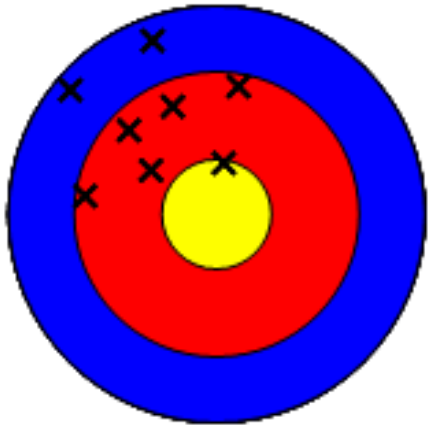


Progress in science depends on *new techniques*, new discoveries, and new ideas …

… probably in that order.

Sydney Brenner, 2002 Nobel Prize Winner

Accuracy *vs* precision

	Accurate	Biased
Precise		
Noisy		

http://withfriendship.com

RNA-Seq precision

*Sequencing of **randomly sampled** fragments!*

Little attention to measurement precision

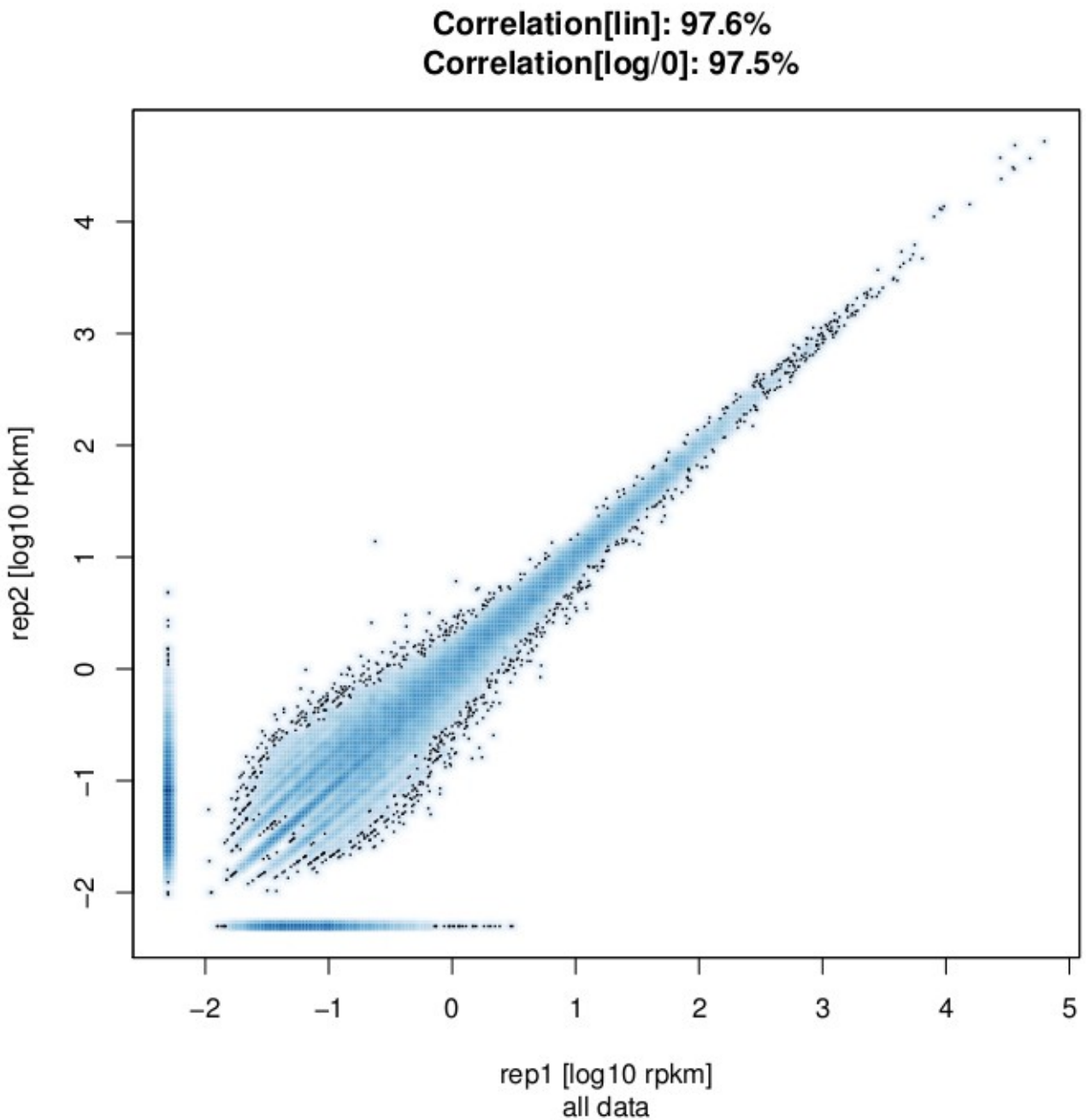
→ initial observations of overall good correlation

(Marioni *et al.* 2008, Wilhelm *et al.* 2008)

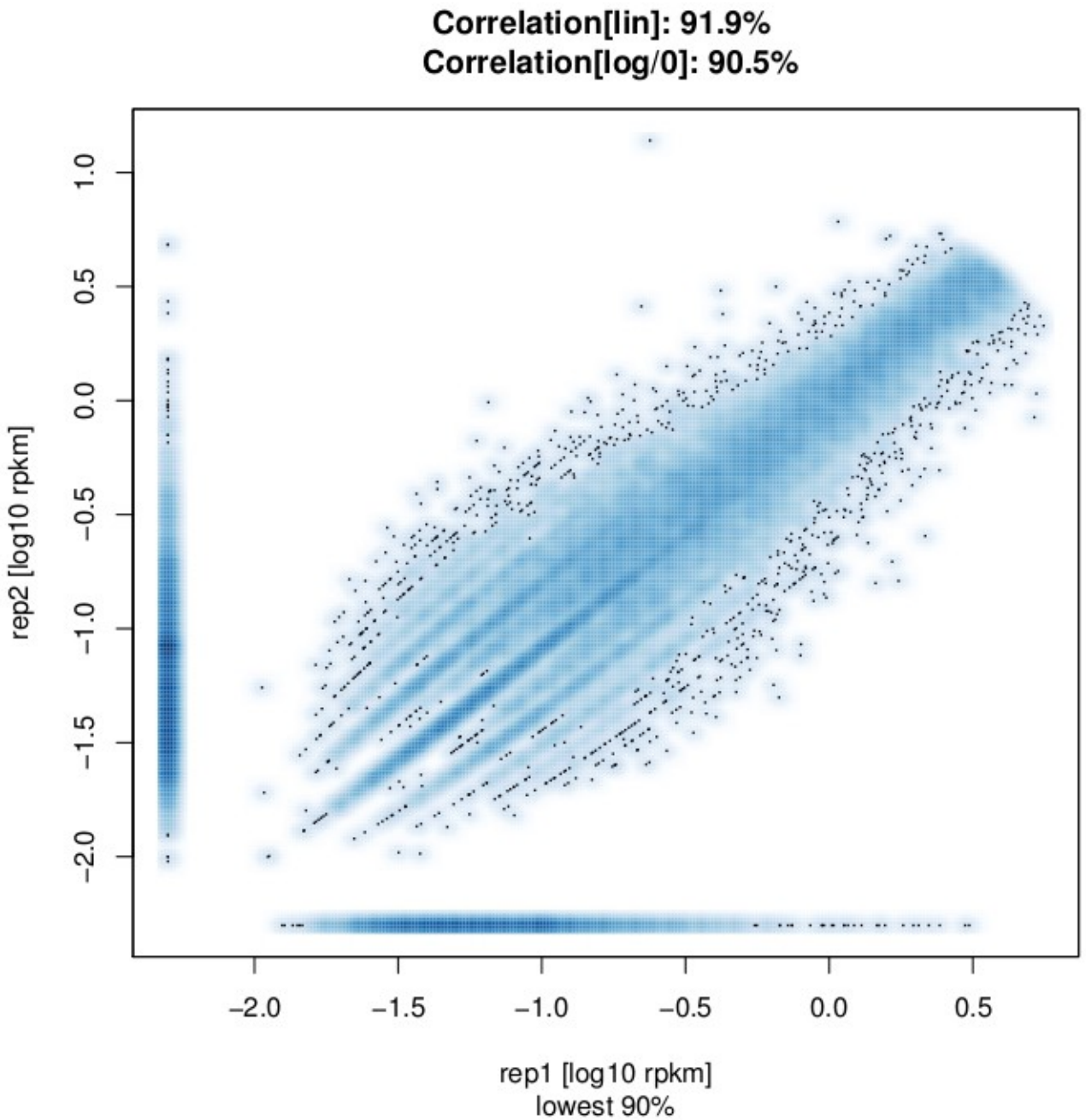
Correlation coefficient can be dominated by extreme values

→ drawback of high dynamic range in RNA-Seq

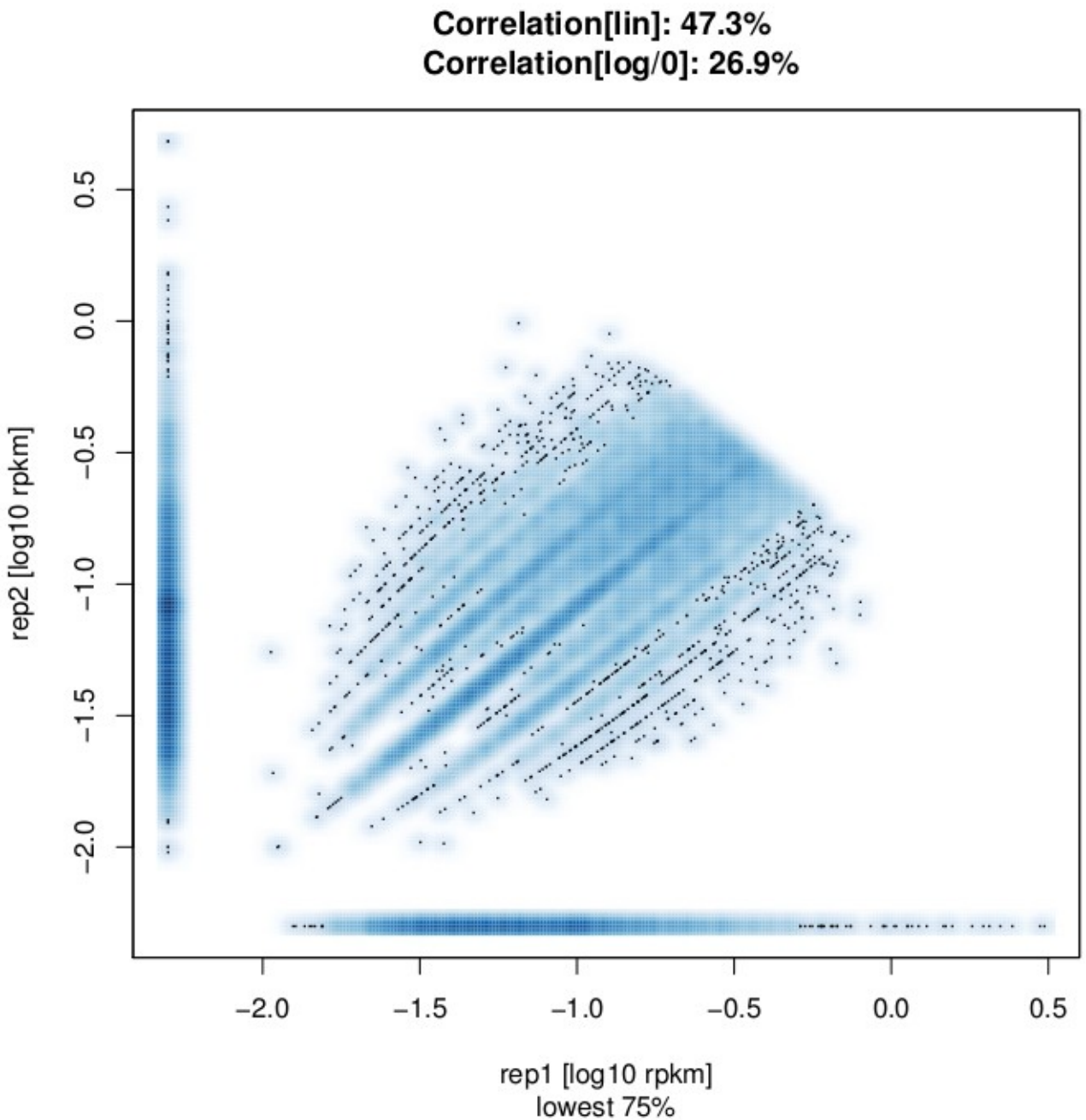
Assessing expression reproducibility



Assessing expression reproducibility

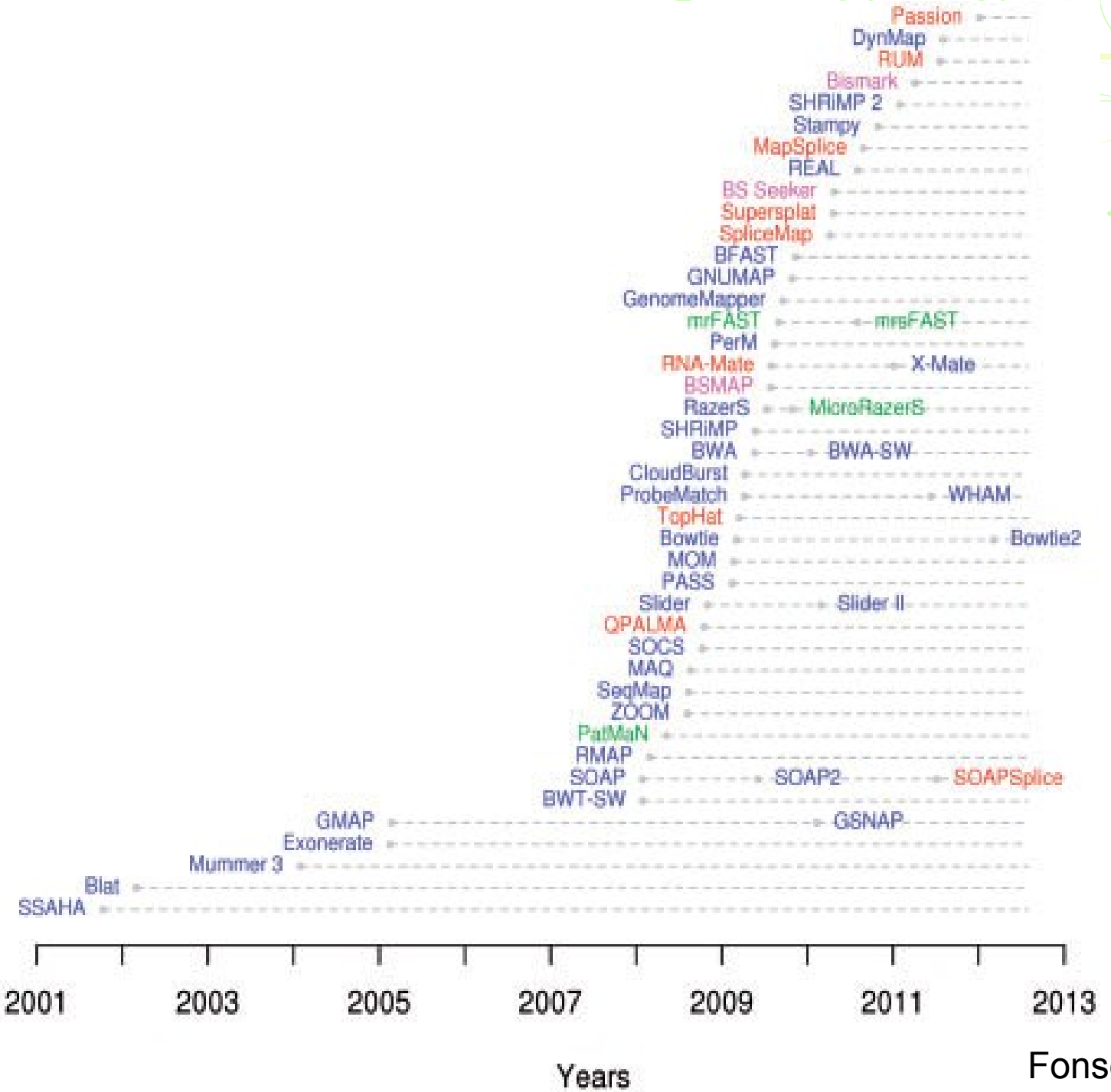


Assessing expression reproducibility



Characterization of established RNA-Seq pipelines

Flood of read mapping tools



Plus:

- **STAR**
- **Subread**

…

Fonseca *et al.* (**2012**) *Bioinformatics*

Simple approach – eg. Bowtie / RPKM

- direct mapping to the transcript sequences
- use of the unique reads for assessing expression levels (RPKM)
  - → exploits only ~ 1 in 5 mapped reads

		Bowtie (transcriptome)		
Replicate	Reads	Mapped reads	Align's	*Unique reads*
1	340	168 (50%)	772	36 (11%)
2	341	167 (49%)	776	35 (10%)
3	311	152 (49%)	699	32 (10%)
Total	993	487 (49%)	2237	103 (10%)

Quantification by *unique reads*

(human cell line sample, 50 bp ABI SOLiD 3+)

More advanced tools

Read – centric: assign probability for each read/fragment to one transcript by maximazing the joined likelihood of read alignments based on the distribution of transcript fragments → estimating the transcript expression

Exon – centric: considers the read abundance on an exonic segment as the cumulative abundance of all transcript isoforms

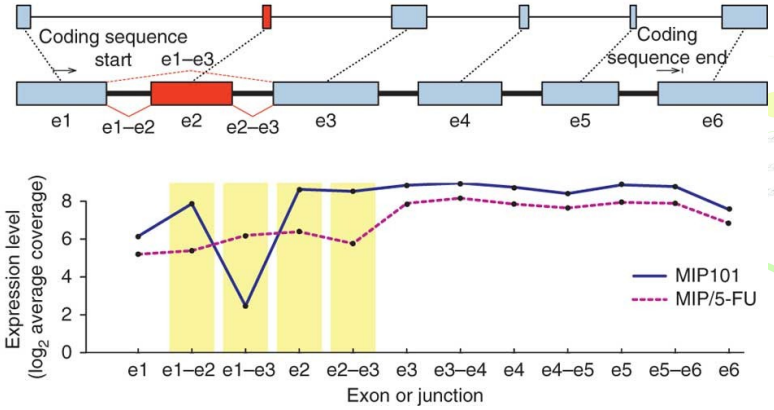
Huang *et al.* RECOMB 2012

More advanced tools

ALEXA – Seq

comprehensive target library from external databases

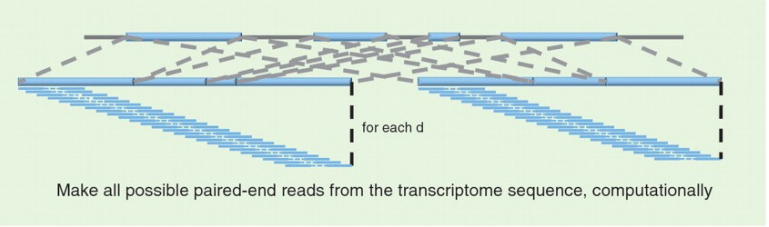
Griffith *et al.* 2010



NEUMA

expected read counts for all possible isoforms

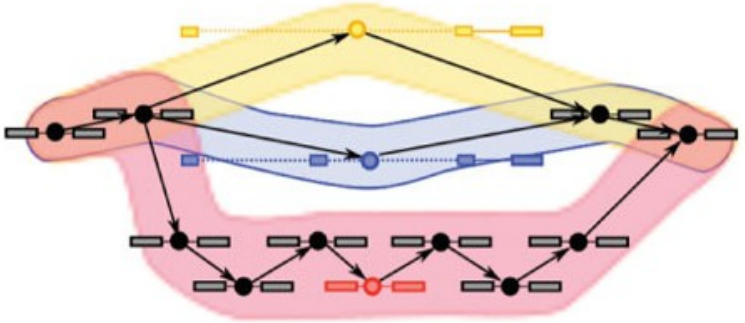
Lee *et al.* 2010



TopHat + Cufflinks

can construct completely *de novo* gene models

Trapnell *et al.* 2009, 2010, 2012



BitSeq

works directly on transcript expression estimates

Glaus *et al.* 2012

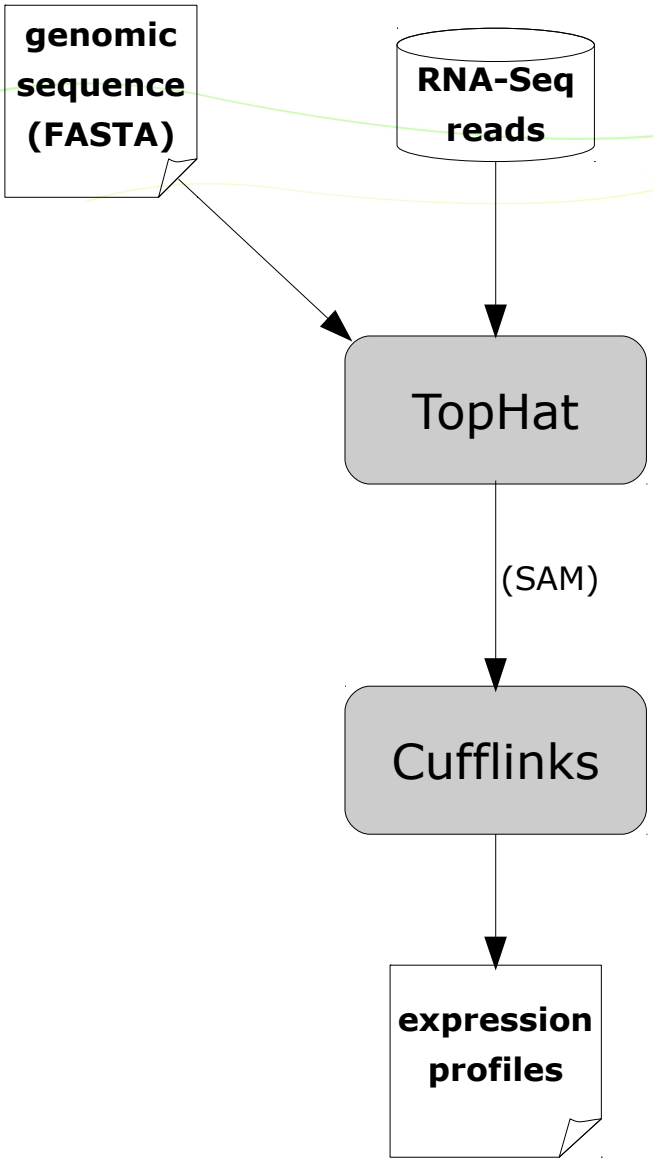
Characterization of the TopHat pipeline

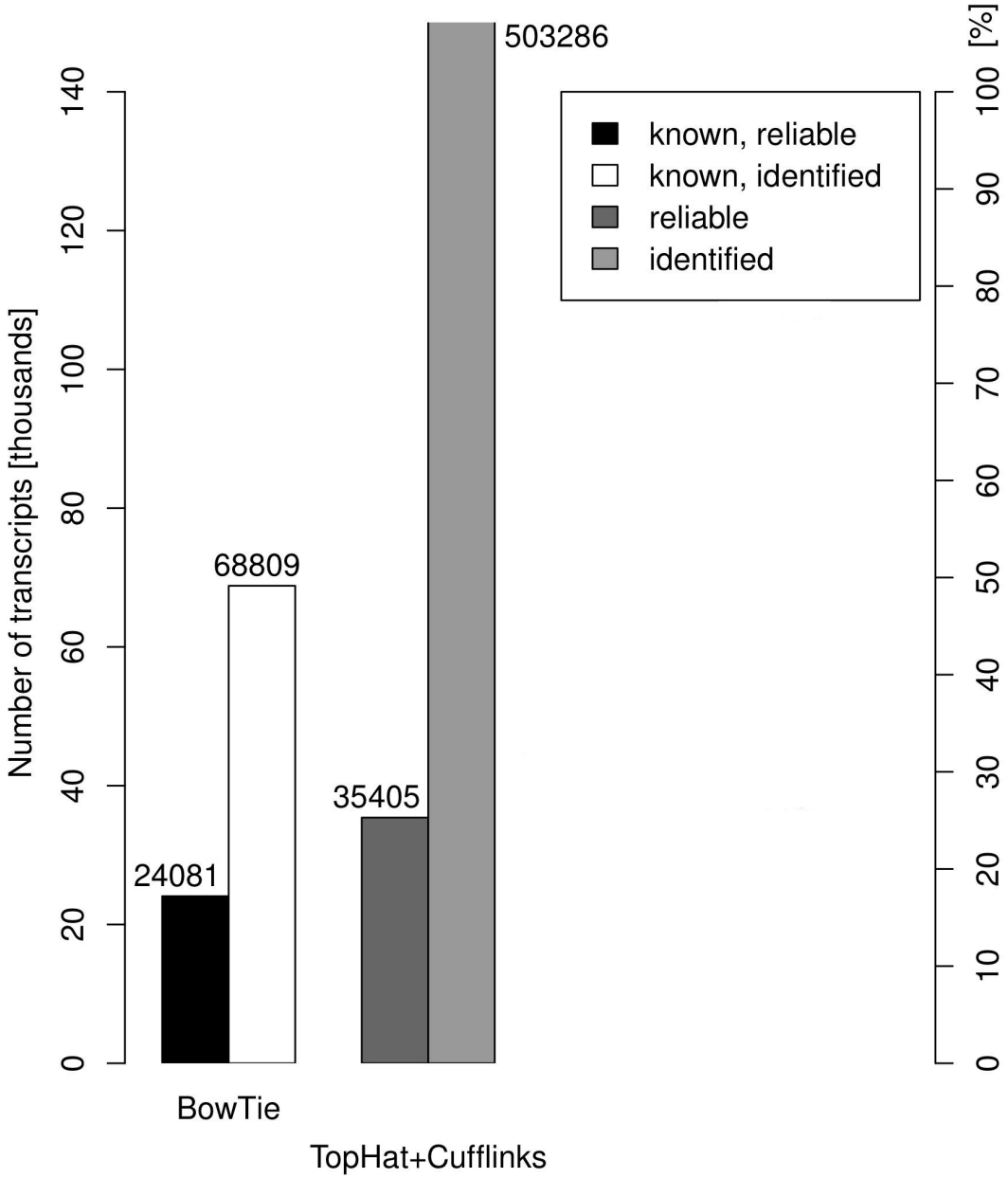
- mapping to the genomic sequence
- *de novo* splice junctions discovery for building gene models
- allows use of all mapped reads for expression estimates

		Bowtie (transcriptome)			TopHat (genome)	
Replicate	Reads	Mapped reads	Align's	*Unique reads*	Mapped reads	*Align's*
1	340	168 (50%)	772	36 (11%)	172 (51%)	241
2	341	167 (49%)	776	35 (10%)	170 (50%)	238
3	311	152 (49%)	699	32 (10%)	155 (50%)	217
Total	993	487 (49%)	2237	103 (10%)	497 (50%)	695

Quantification by *unique reads* *alignments*

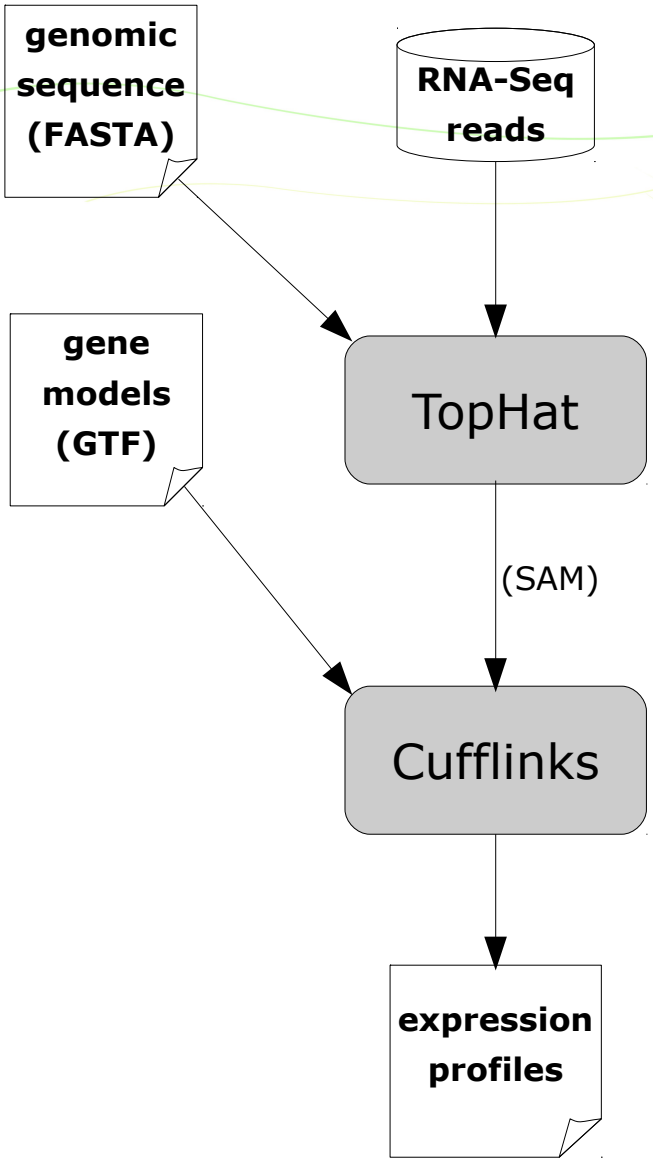
TopHat + Cufflinks +/– models

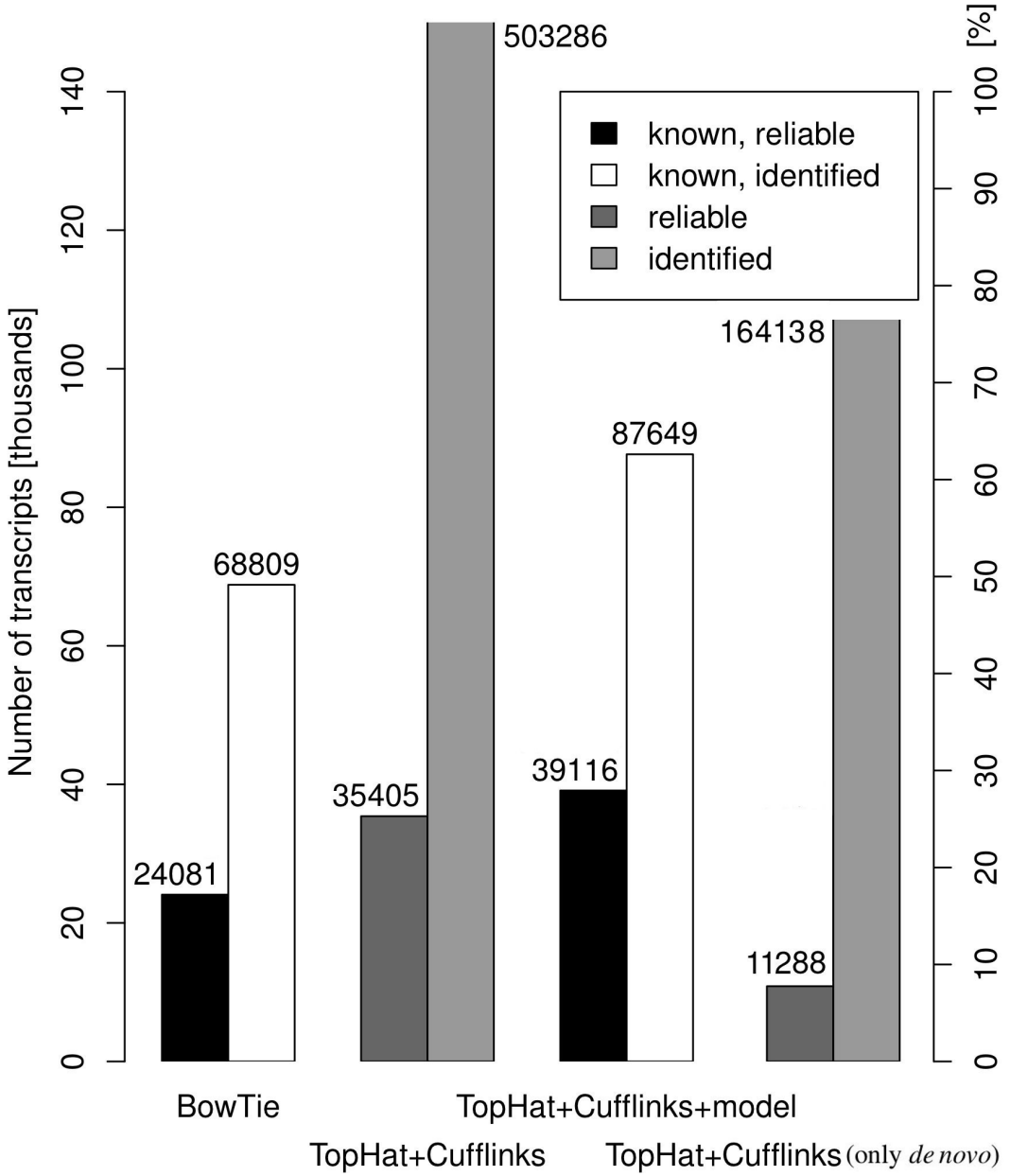




'reliable': < 20% *relative error*

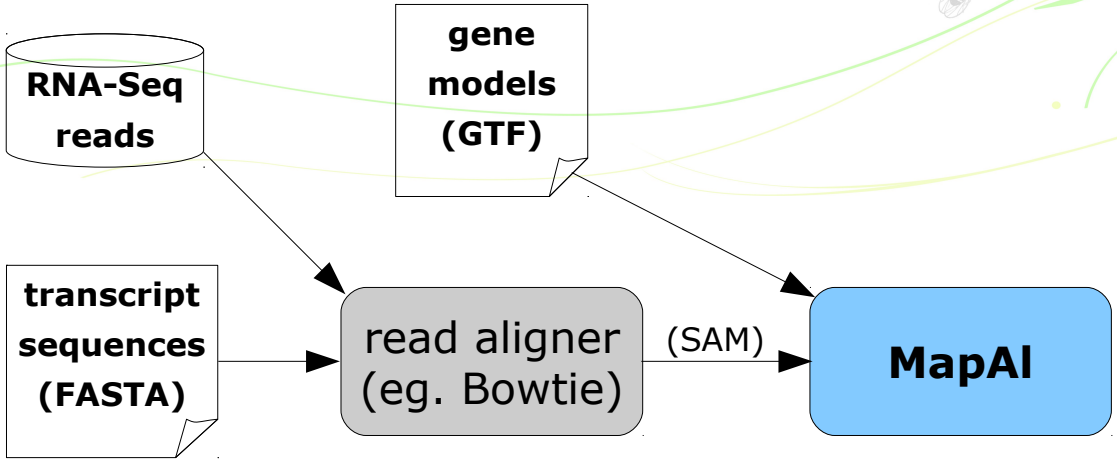
TopHat + Cufflinks +/– models





'reliable': < 20% *relative error*

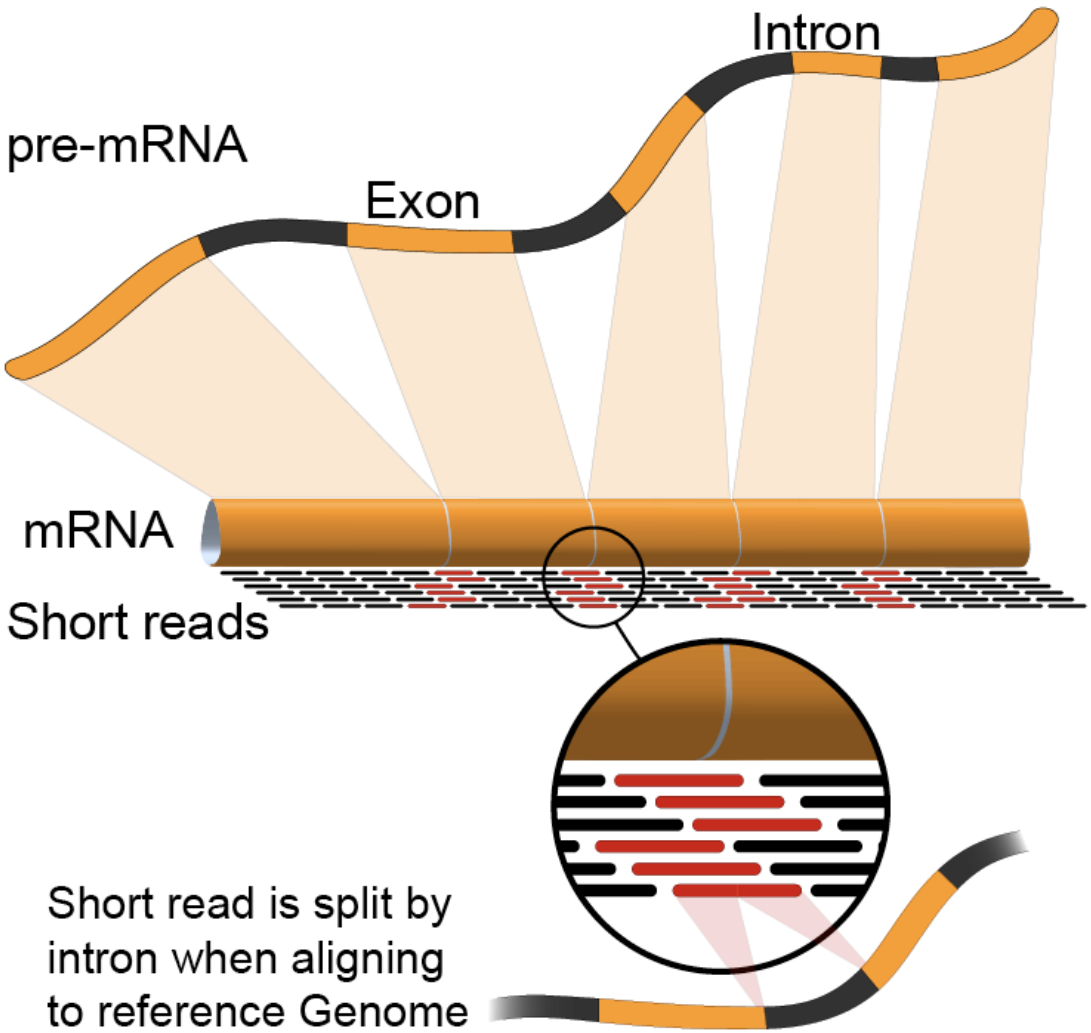
Exploiting gene models at the alignment stage



		TopHat (genome)			Bowtie (combined)		
Replicate	Reads	Mapped reads	*Align's*	Junct	Mapped reads	*Align's*	Junct
1	340	172 (51%)	241	18	168 (50%)	249	45
2	341	170 (50%)	238	17	167 (49%)	247	45
3	311	155 (50%)	217	16	152 (49%)	225	41
Total	993	497 (50%)	695	51	487 (49%)	721	131
Quantification by			*alignments*			*alignments*	

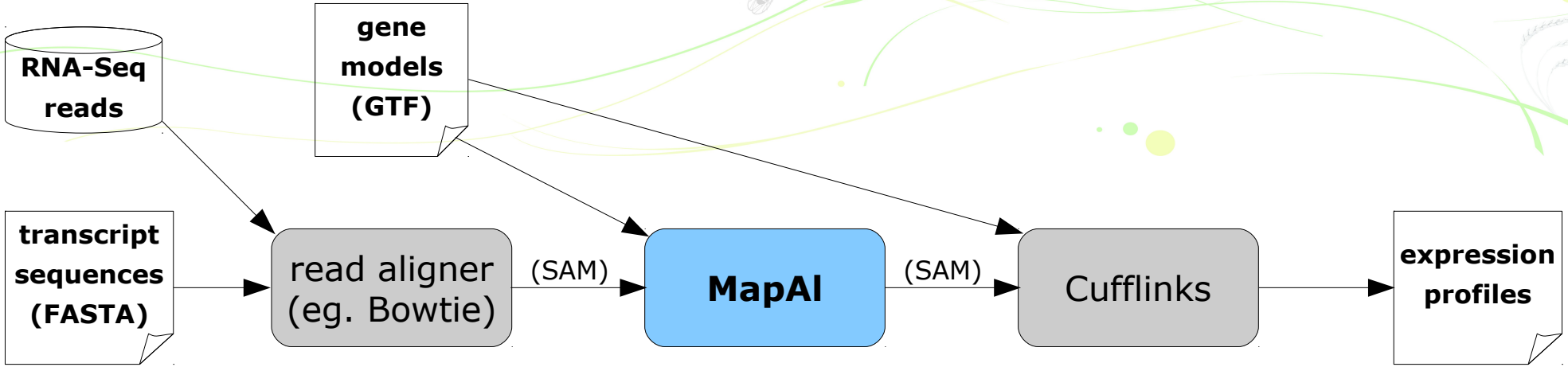
Combined solution is much more sensitive in the identification of known **junctions**.

Exploiting gene models at the alignment stage



- most genes have alternative splice variants
- most reads map to more than one splicing variant
- often splice-junctions identify a specific splicing variant

Exploiting gene models at the alignment stage



		TopHat (genome)			Bowtie (combined)		
Replicate	Reads	Mapped reads	*Align's*	Junct	Mapped reads	*Align's*	Junct
1	340	172 (51%)	241	18	168 (50%)	249	45
2	341	170 (50%)	238	17	167 (49%)	247	45
3	311	155 (50%)	217	16	152 (49%)	225	41
Total	993	497 (50%)	695	51	487 (49%)	721	131

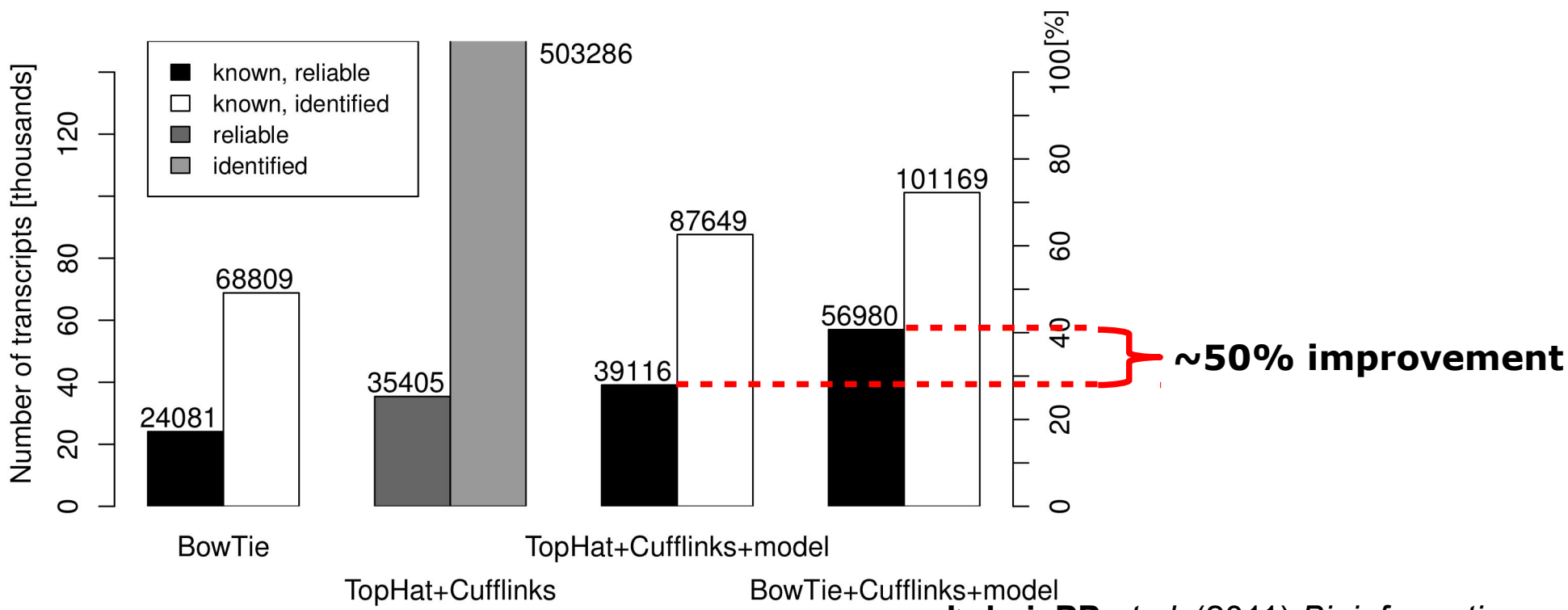
Quantification by *alignments* *alignments*

Combined solution is much more sensitive in the identification of known **junctions**.

These often play a *key role* in identifying the expression of a particular spliceform.

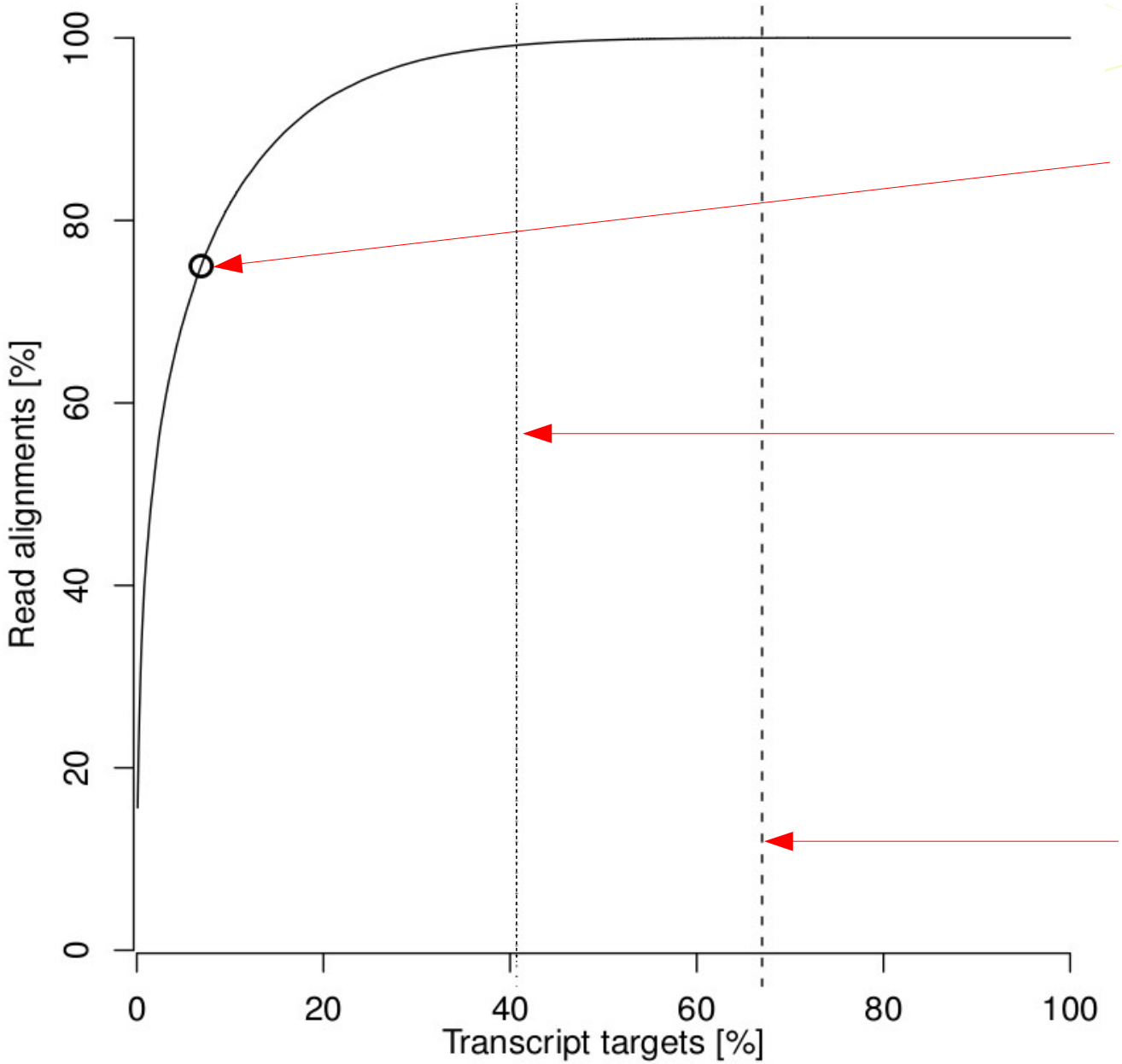
Reproducibility of quantitative expression profiling

Transcripts	Bowtie	TopHat + Cuff-links	TopHat + Cuff-links + model	Bowtie + Cuff-links + model
identified	68,809 (49%)	503,286 (–)	87,649 (63%)	101,169 (**72%**)
reliable	24,081 (17%)	35,405 (–)	39,116 (28%)	56,980 (**41%**)
reliable %	35%	7%	44%	57%



Łabaj, PP *et al.* (2011) *Bioinformatics*
Łabaj, PP *et al.* (2012) *Frontiers in Genetics*

Effects of highly expressed transcripts



> 75% of all read alignments

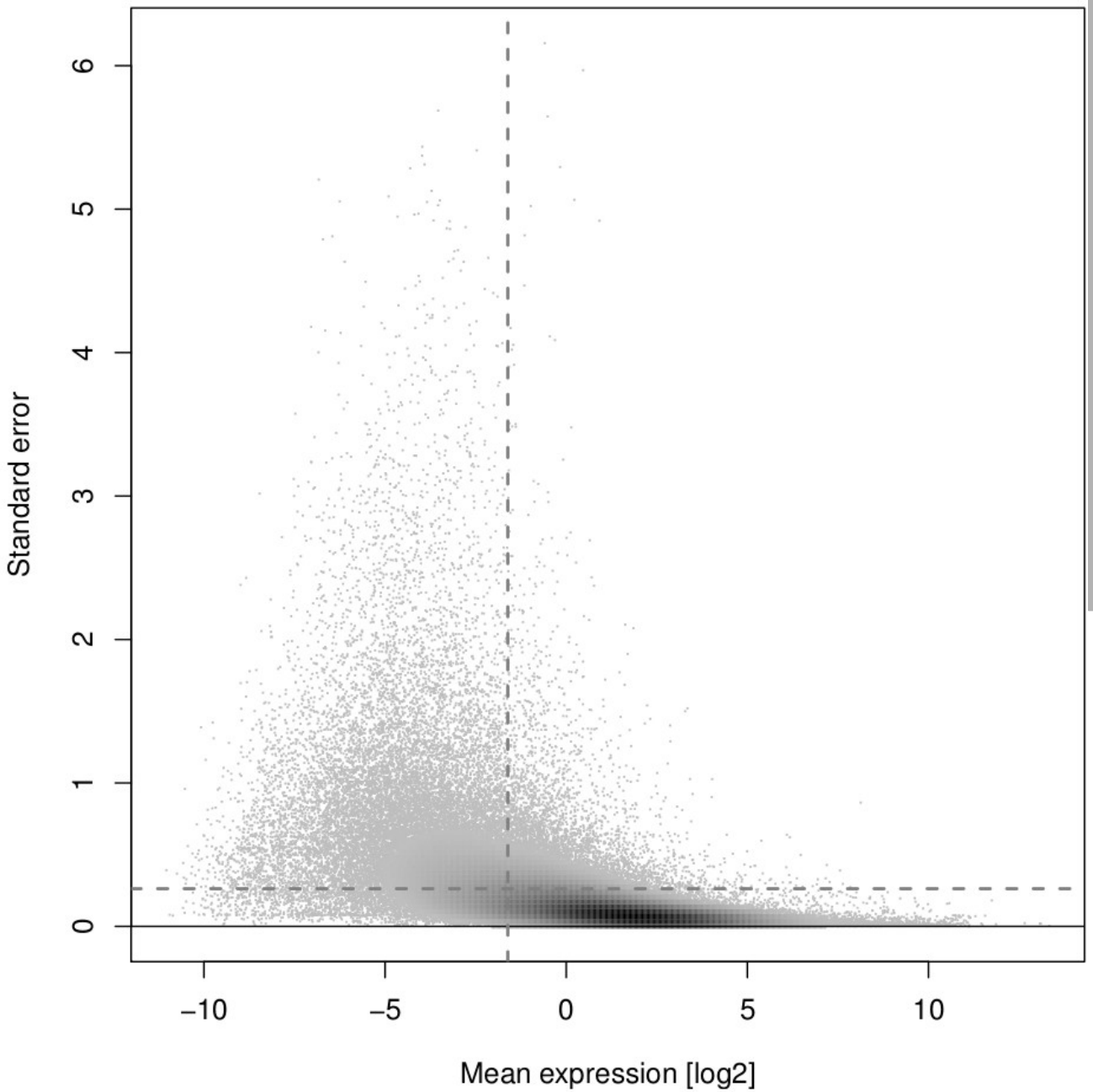
< 7% of known transcripts

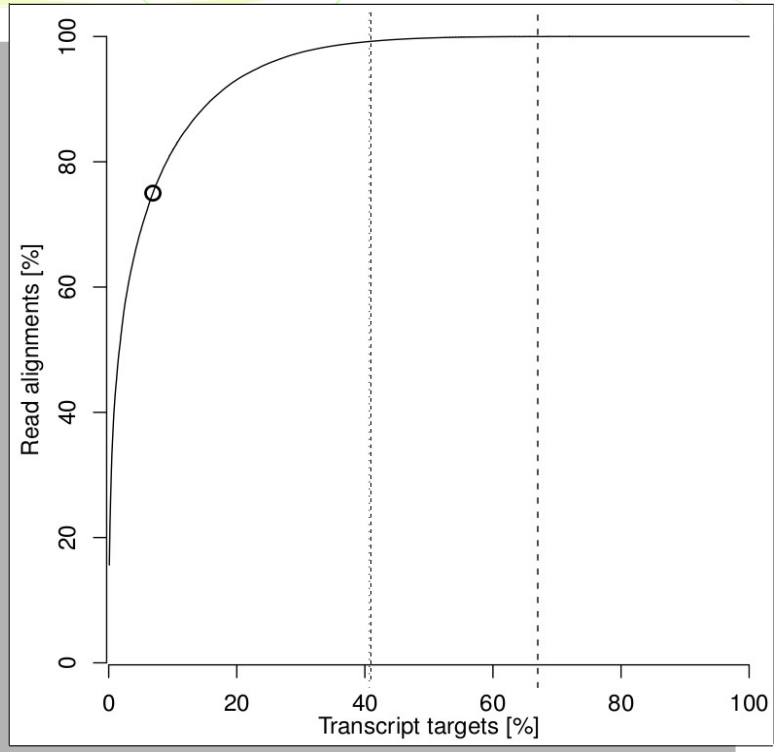
99.5% of all read alignments

41% of known transcripts

detected

Effects of highly expressed transcripts

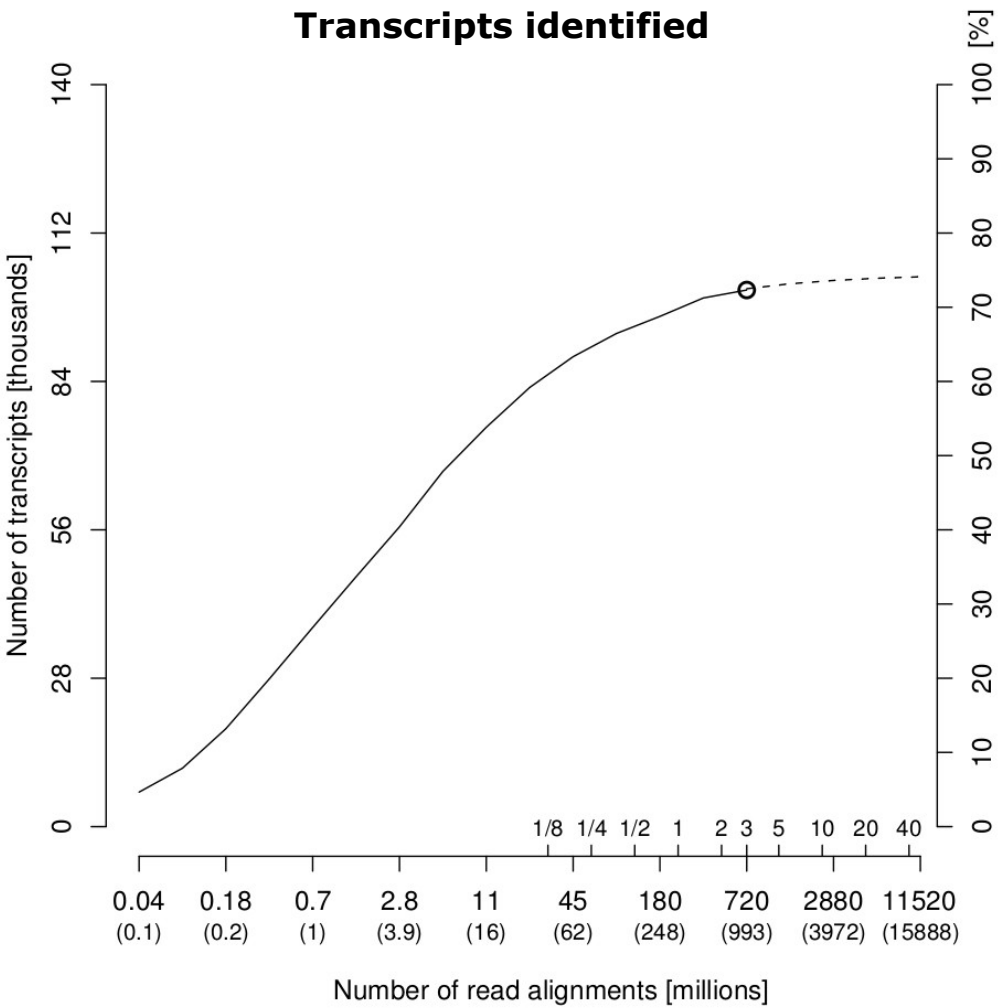


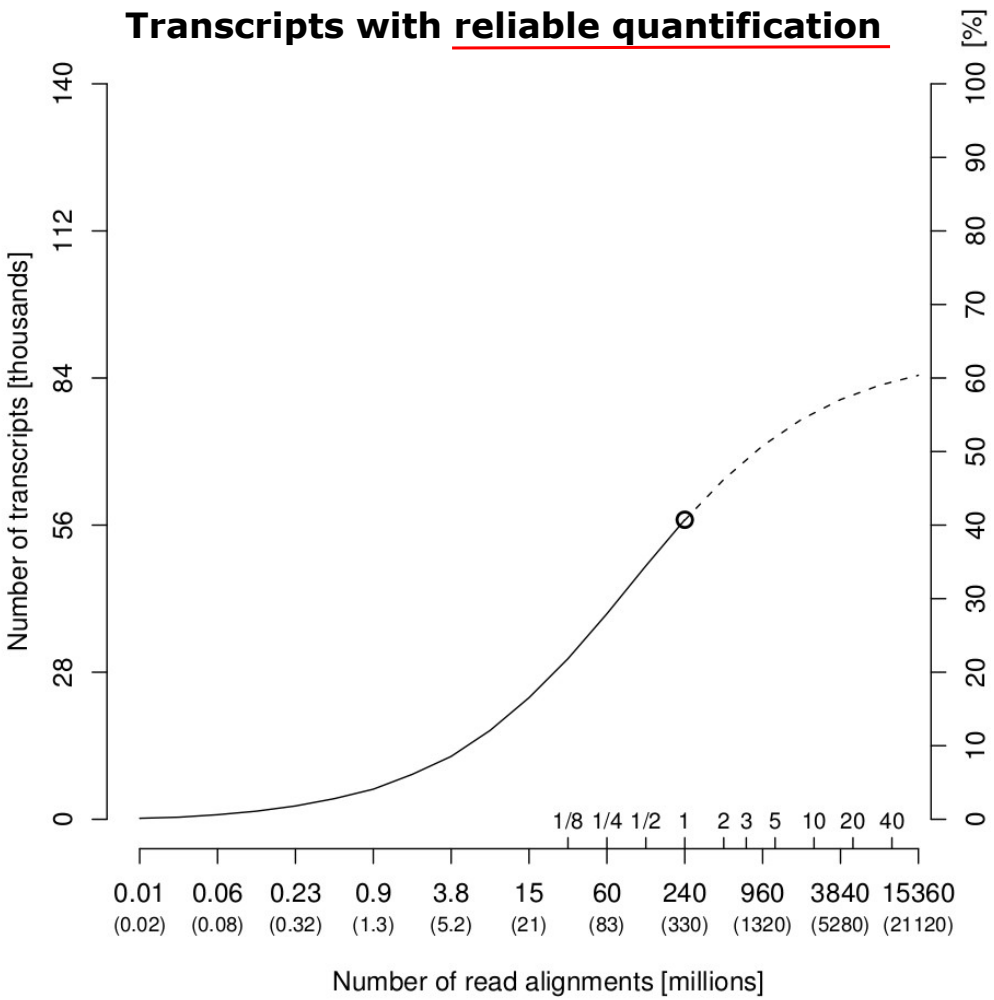


> 75% of all read alignments

< 7% of known transcripts

Impact of read depth





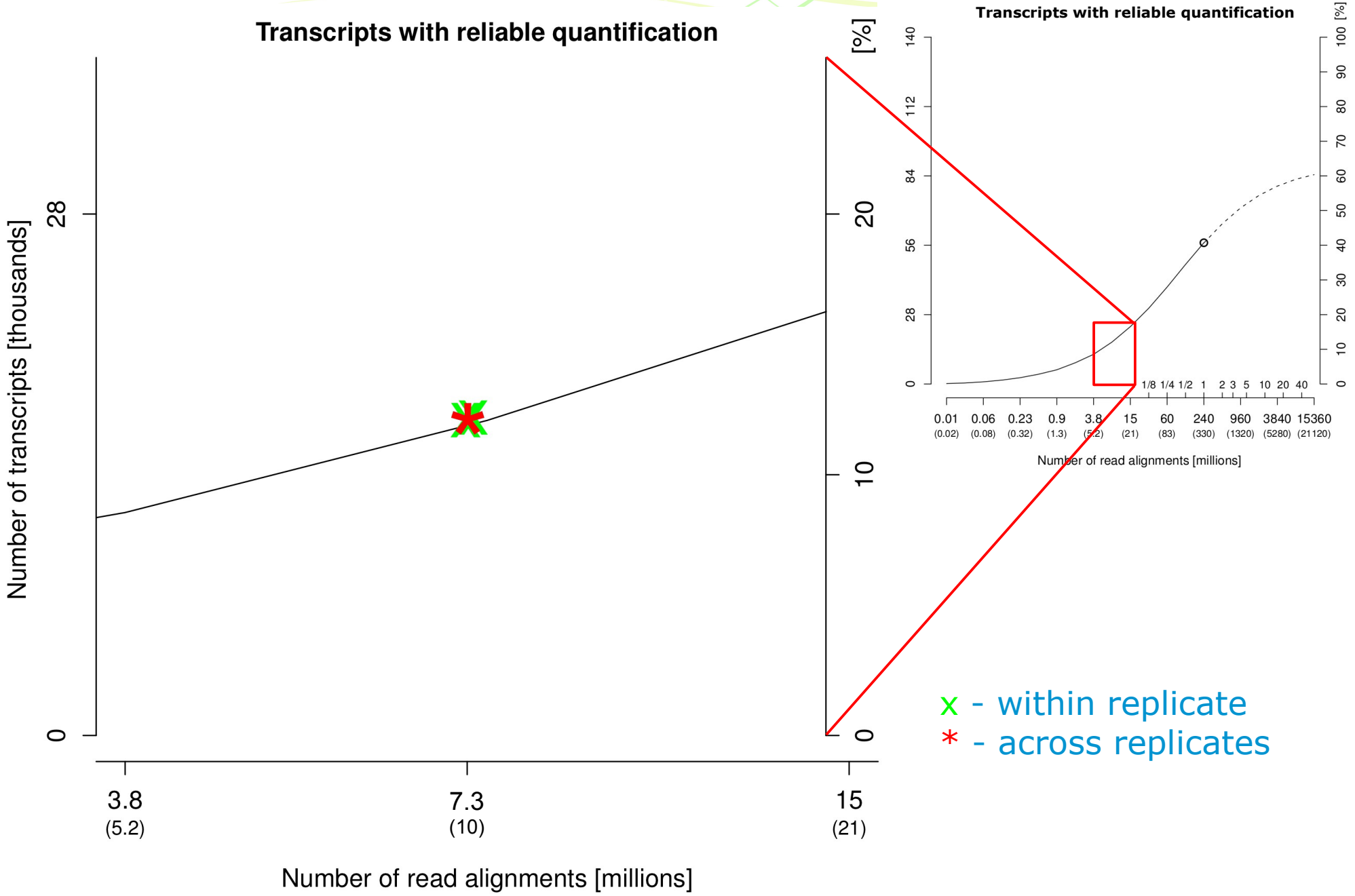
point of diminishing returns

~20% actually not expressed

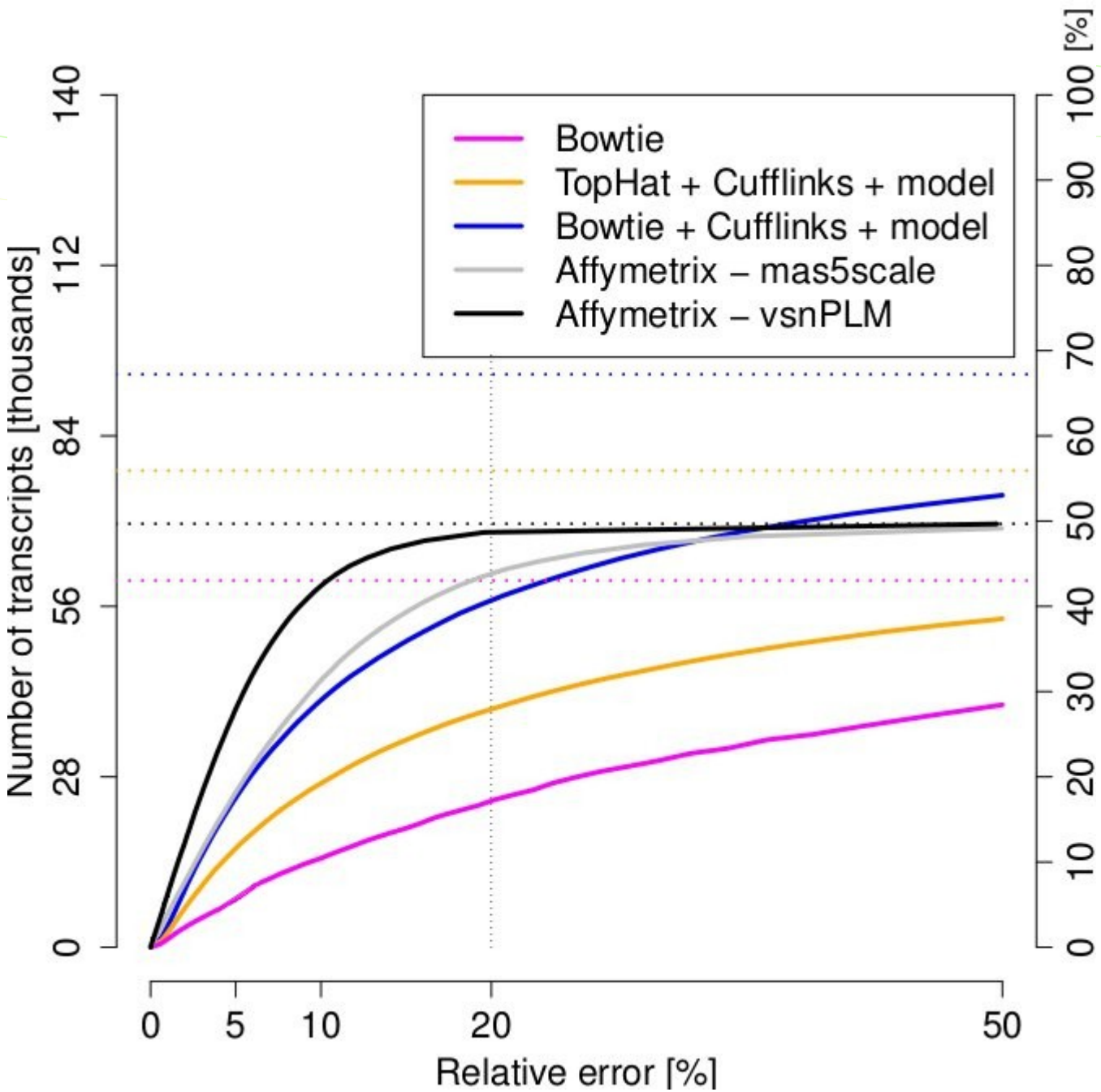
doubling sequencing depth

→ only 5% more reliable targets

Dominance of the sampling effect



RNA-Seq *vs* arrays



Summary and outlook

Exploiting gene models *already at the alignment stage* →

~ 100,000 spliceforms identified (72% of all known)

~ 57,000 measured reliably (41%)

→ an improvement of 50% !

Standard microarrays can reliably measure > 68,000 transcripts

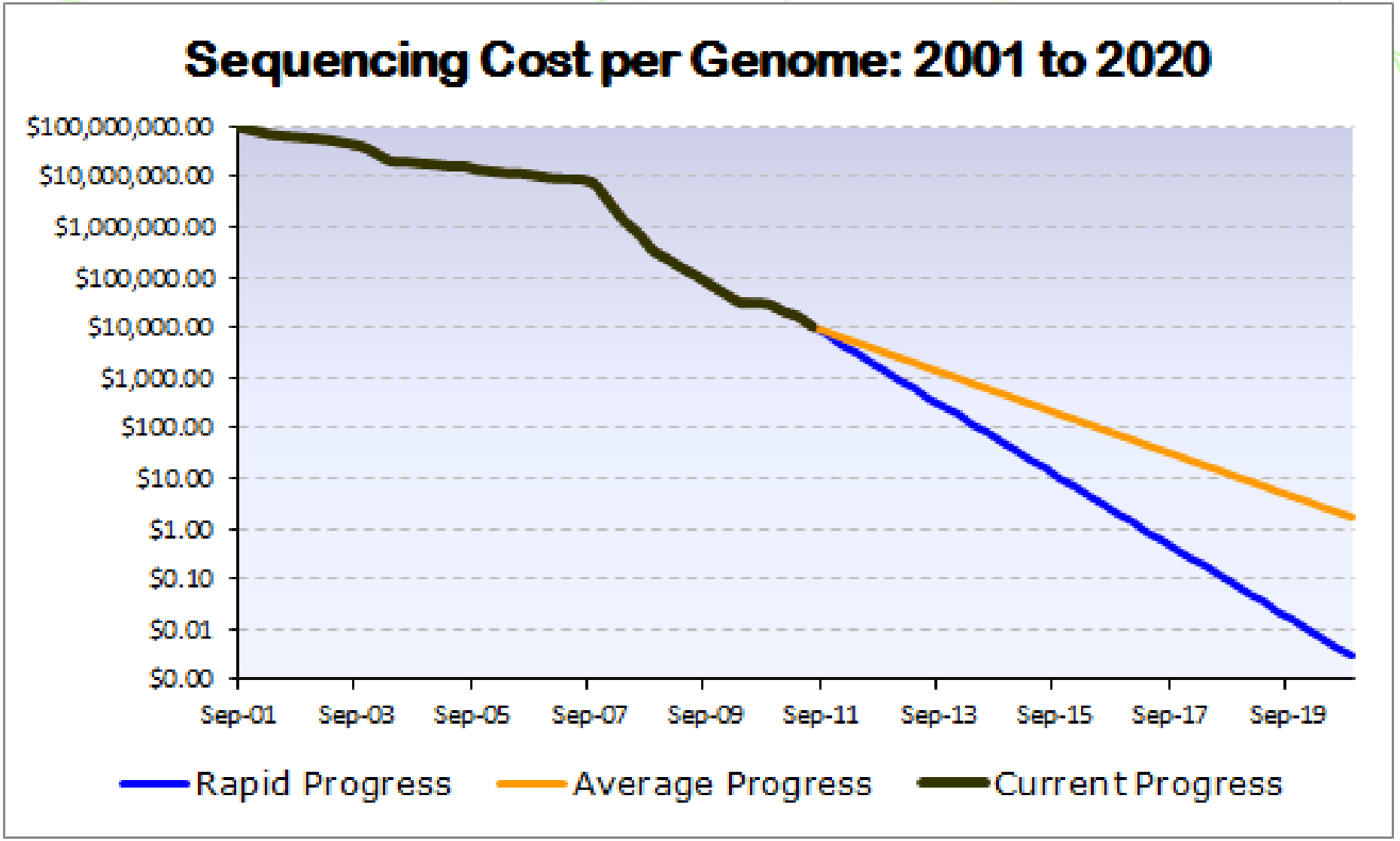
→ 20% more than RNA-Seq …

A doubling of the sequencing depth

- changes little for the number of identified transcripts
- adds 5% to the number of transcripts that can be quantified reliably, with diminishing returns for higher sequencing depths

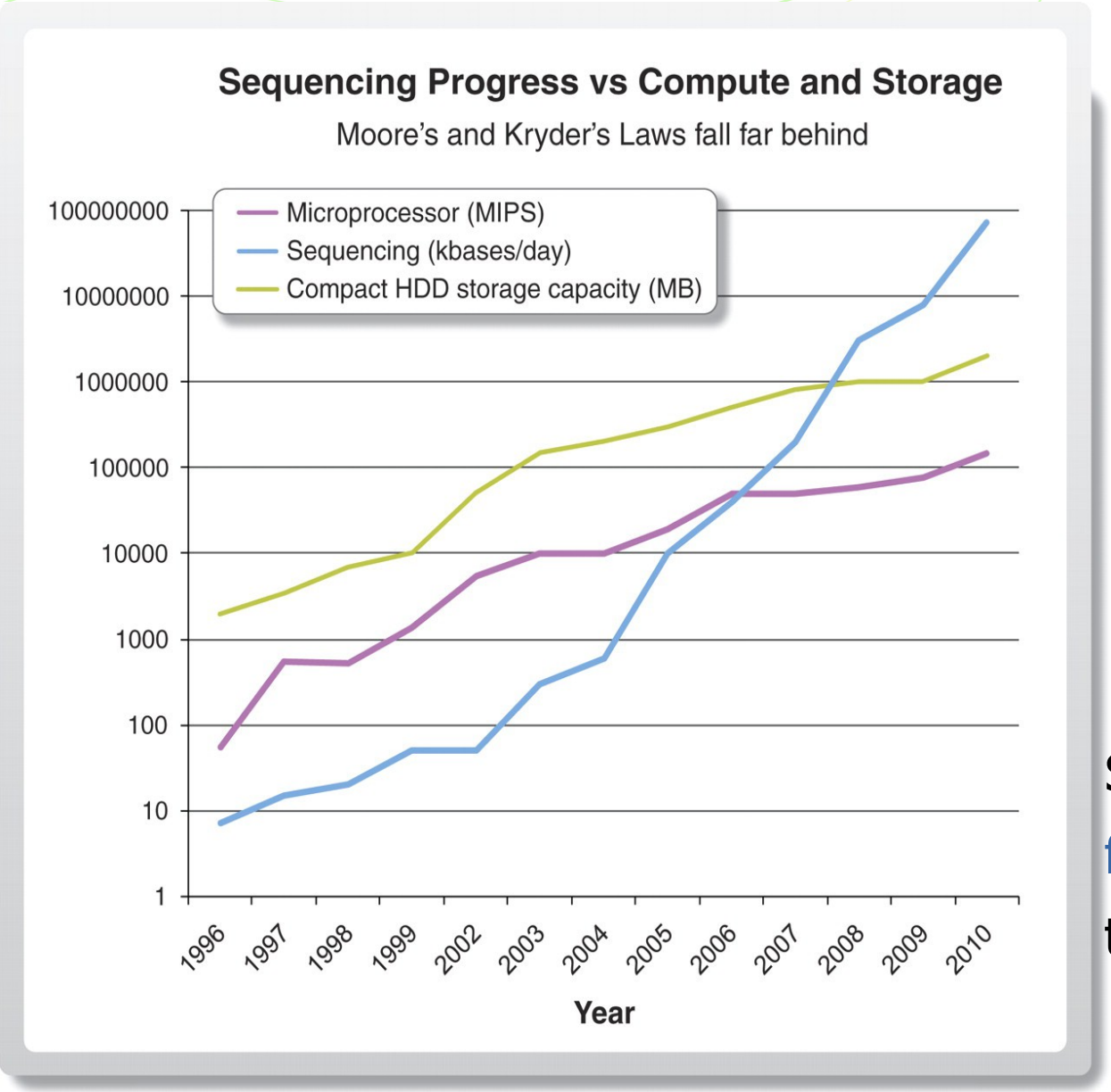
(… 75% of read alignments hit < 7% highly expressed transcripts!)

Falling costs



Sources: National Human Genome Research Institute and DailyFinance.com

Need to work smarter, not harder



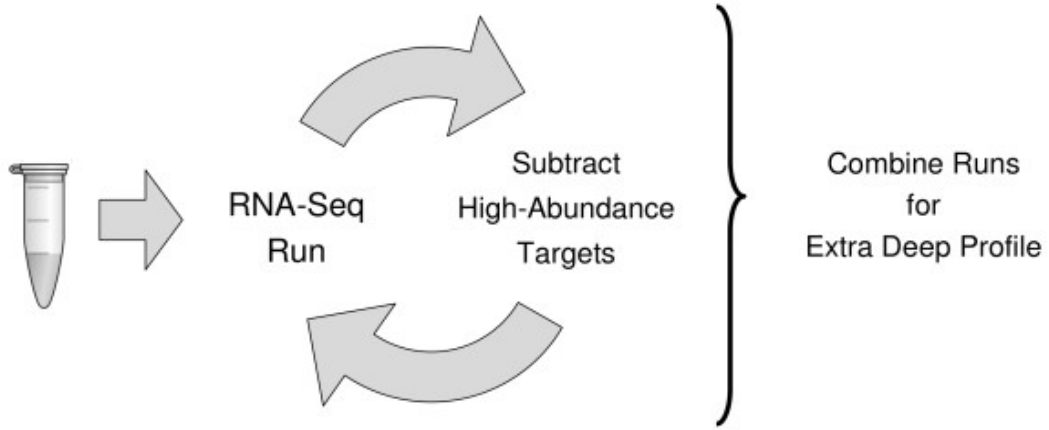
A doubling of processing power every 14 months!

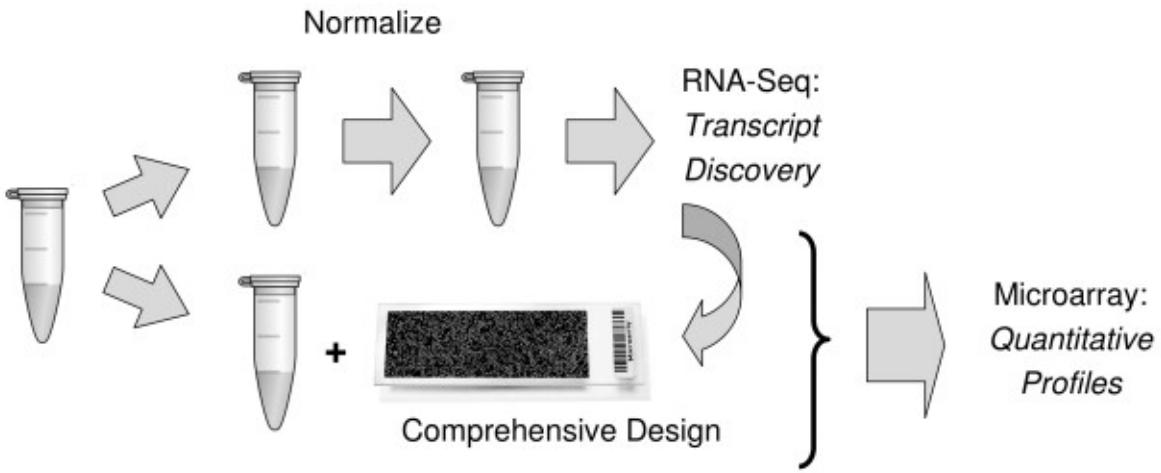
A doubling of storage capacity every 13 months!

A doubling of sequencing output every 5 months!

Sending data oversea by post faster than transferring via network !!!

Summary
- combining complementary strengths





Acknowledgements

Chair of Bioinformatics:

David Kreil

Peter Sykacek

Nancy Stralis-Pavese

...

EMSL, PNL:

Bryan E. Linggi

Lye Meng Markille

H. Steven Wiley

MAQC consortium:

Weida Tong

Leming Shi

Chris E. Mason

...

www.bioinf.boku.ac.at

Thank you for your interest.

Visit us at

www.bioinf.boku.ac.at

THE UNIVERSITY OF WARWICK

chair of bioinformatics

CAMDA2014

13th Annual International Conference on Critical Assessment of Massive Data Analysis

Boston, United States | July 11-12, 2014

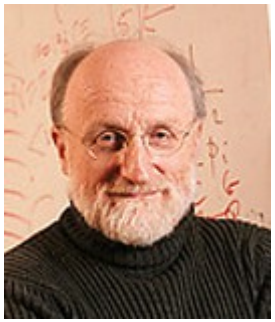
Challenges:

This year, CAMDA's scientific committee set up three challenges to integrate multi-track -omics data:

1. dual dose response profiles for 14 unknown and 2 known compounds from the InnoMed PredTox project of the EU FP7 program,
2. selected cancers from International Cancer Genome Consortium (ICGC), and
3. the prediction of drug compatibility from an extremely large toxicogenomic data set

For additional information see:

http://www.camda.info (coming soon)



Chris Sander

Temple F. Smith

Jun Wang

Abstract Submission Deadline	**20 May 2014**
Poster Submission Deadline	25 May 2014
Acceptance Notification	30 May 2014
Early Registration Closes	1 Jun 2014
CAMDA2014 Conference	11–12 Jul 2014
ISMB 2014 Conference	12–15 Jul 2014
Full Paper Submission	25 Aug 2014

Organizers

Djork-Arné Clevert, JKU Linz, Austria
Joaquin Dopazo, CIPF, Spain
Sepp Hochreiter, JKU Linz, Austria
Lan Hu, Dana-Farber Cancer Institute, U.S.A.
David Kreil, Boku University, Austria
Simon Lin, Marshfield Clinic, U.S.A.

ISMB 2014 boston, usa