MEDS5420 RNA-Intro

Michael Guertin

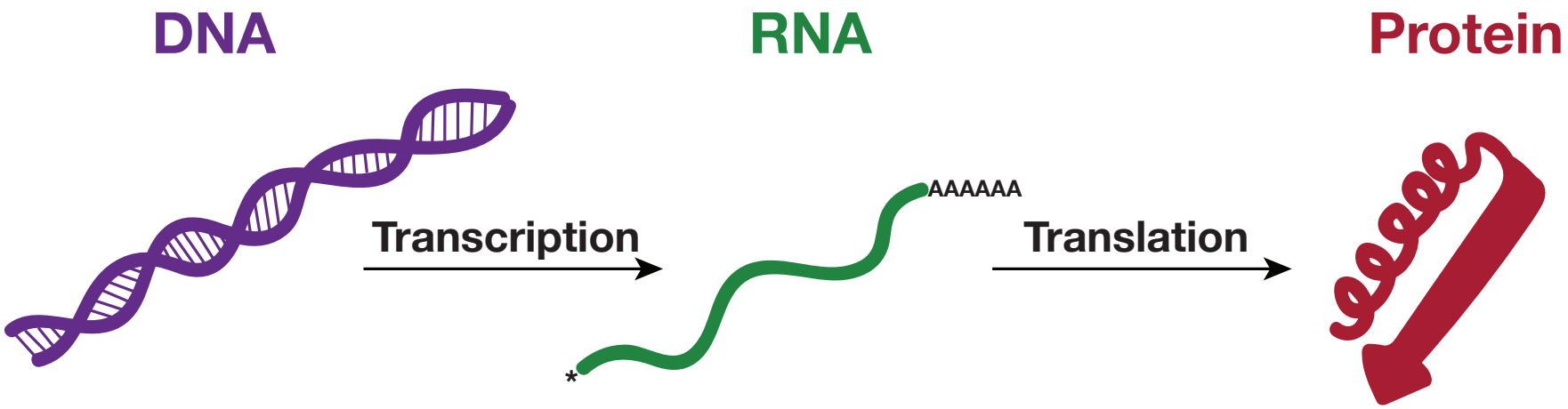
(Many slides stolen from Leighton Core)

March 29, 2023

Transcriptome

Each cell within an organism has an identical genome (more or less); gene expression dictates cellular phenotypes.

Gene Regulation:
From transcription to protein degradation



RNA species

- rRNA: 80%
- tRNA: 15%
- mRNA: ~3%
- miRNA: <1%
- eRNA: <1%
- lncRNA: <1%
- *RNA: <1%

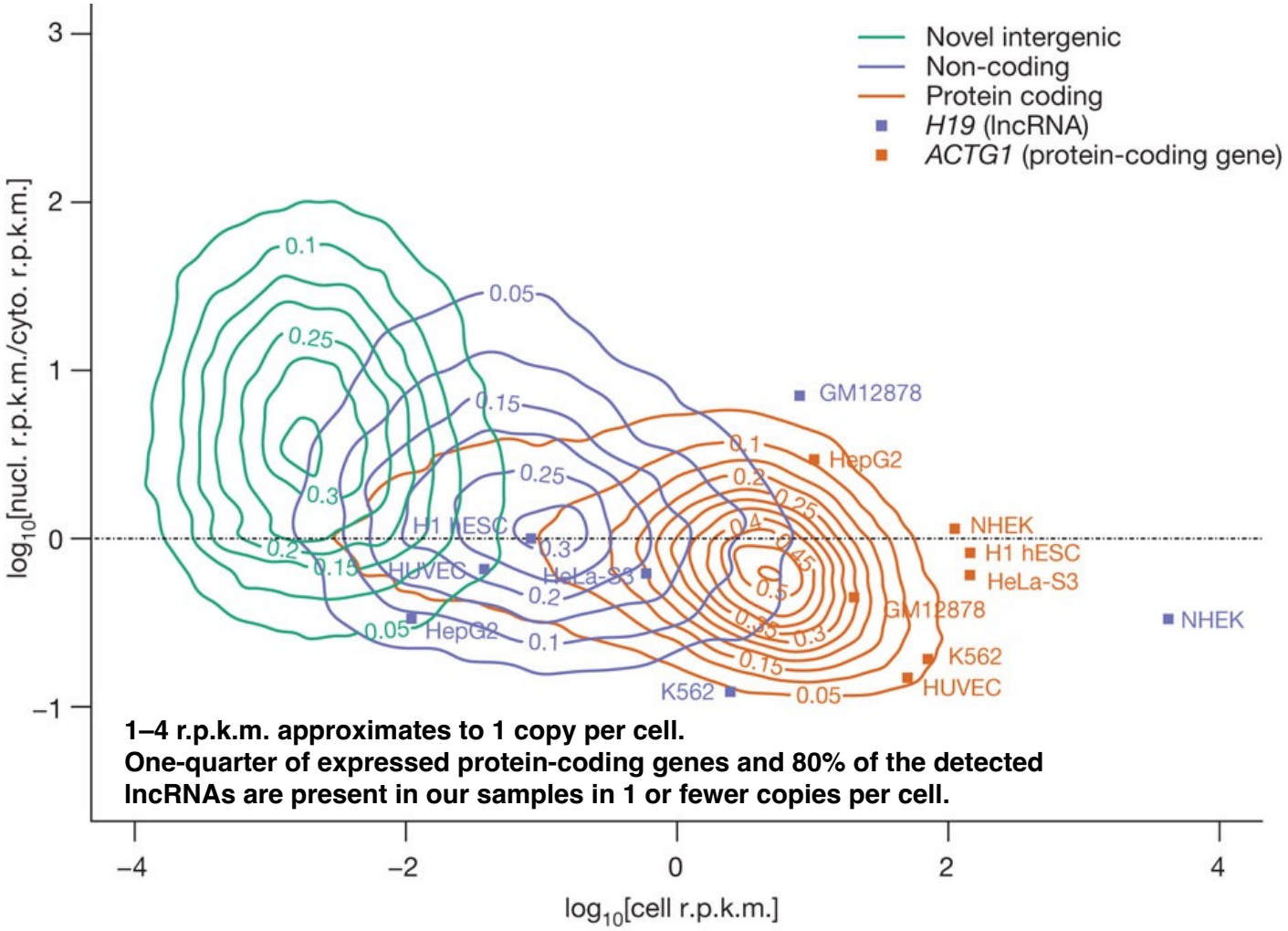
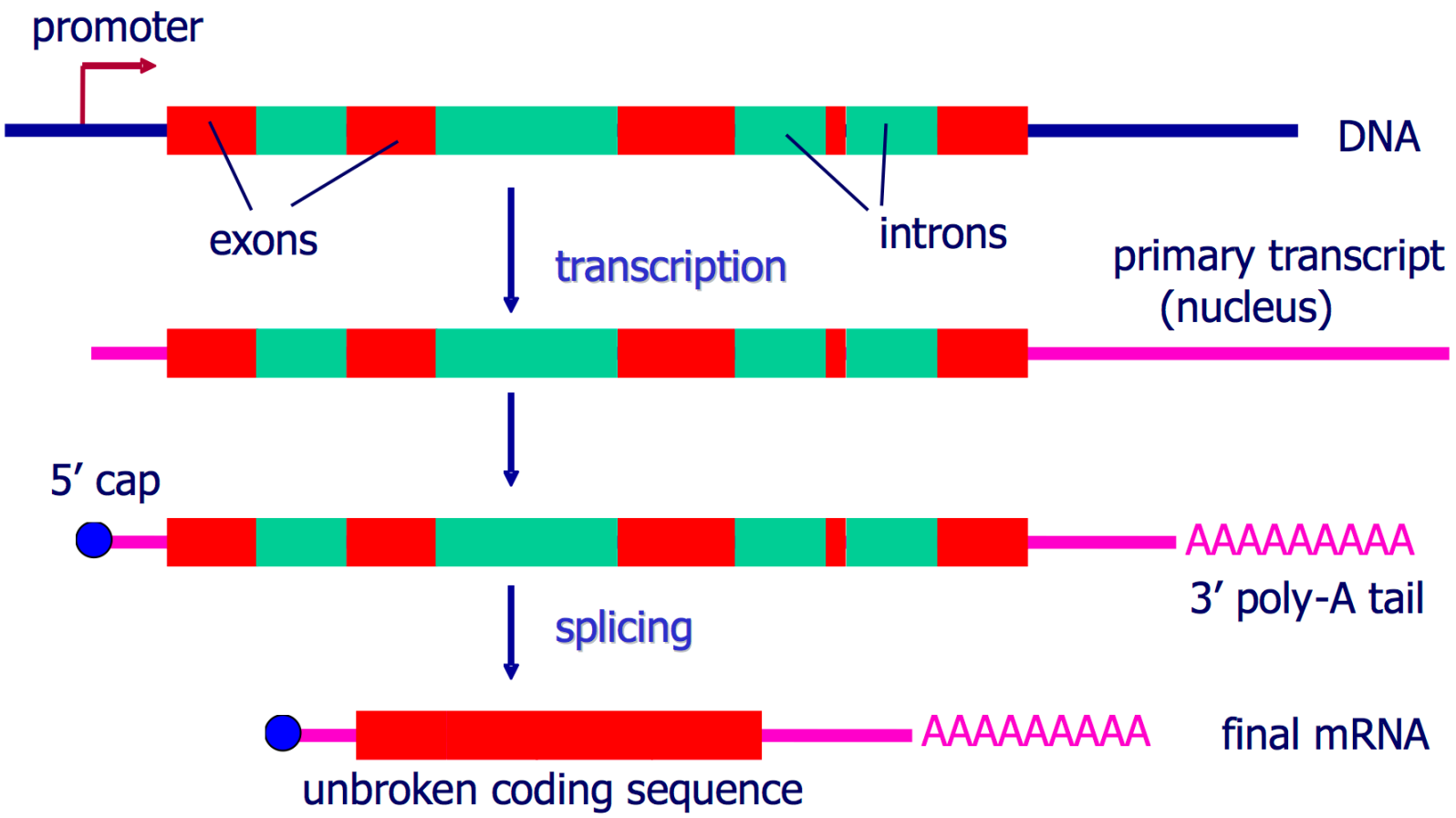


Fig 3. Djebale et al (2012) Nature 489:101

mRNA is generated from longer pre-mRNA



mRNA composition dictates the identity of a cell

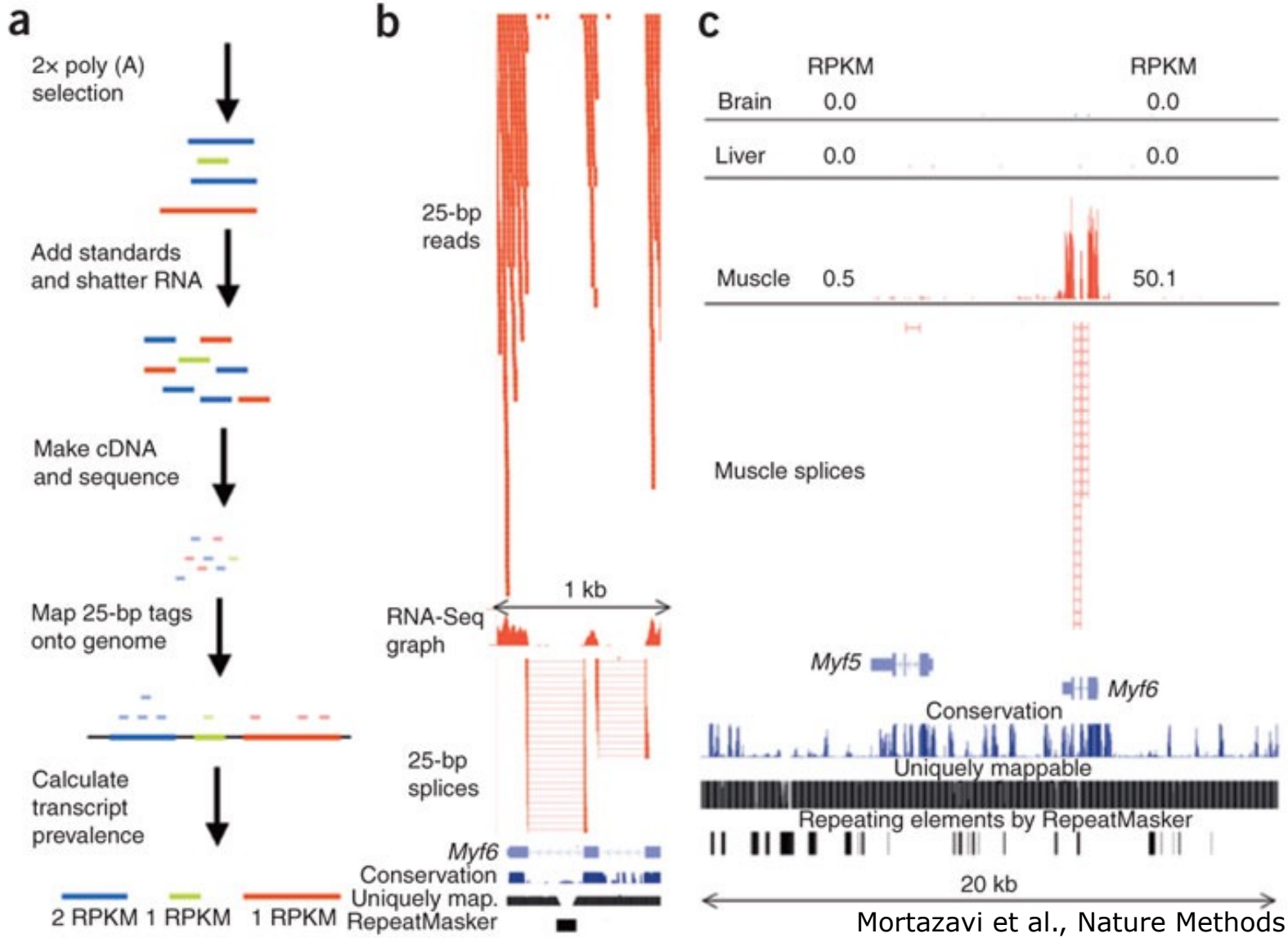
- Although mRNA is ~3% of total RNA, it is the most biologically significant because it specifies the proteome and biochemical capacity of the cell.
- How do we measure the transcriptome?
- What drives lineage-specific transcriptomes during development and throughout a cell's life?

RNA-seq

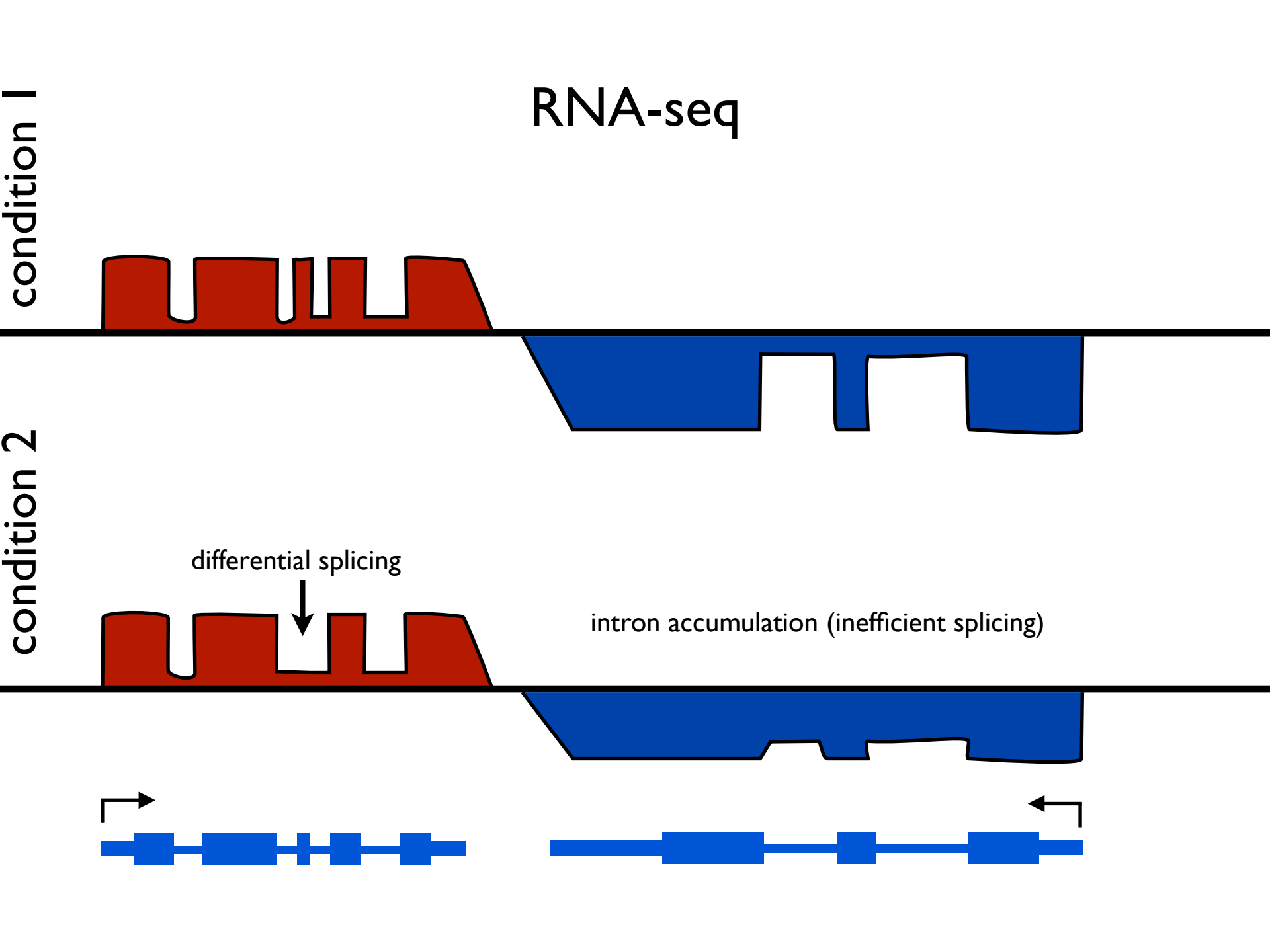
- What RNA is found in the cell?
  - Species-general; good for studies of non-model organisms
  - Good for looking steady-state RNA levels and splicing variants
  - Very few biases: rRNA depletion; poly-A selection
  - Easily interpreted

Nagalakshmi, et al. Science (2008)

RNA-seq



Mortazavi et al., Nature Methods 2008


RNA-seq
condition 1
condition 2
differential splicing
intron accumulation (inefficient splicing)

RNA-seq

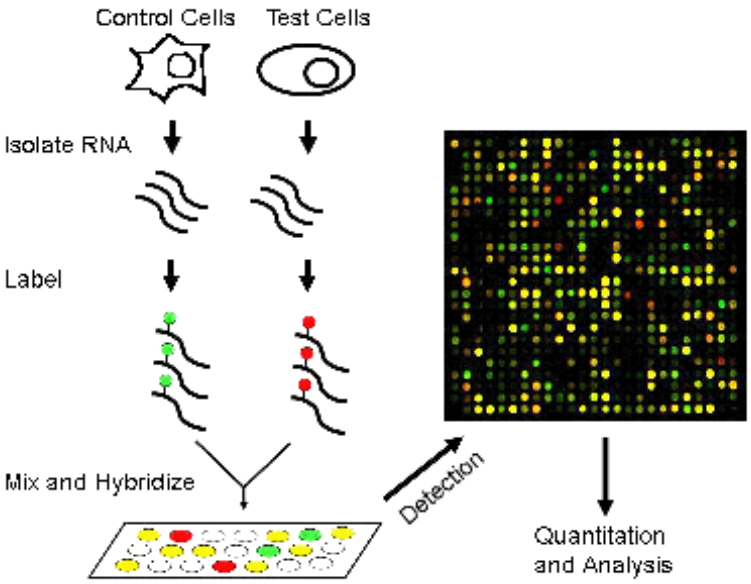
- There are many variants of RNA-seq.
- I consider strand-specific, rRNA depleted, random hexameter priming RNA-seq as the gold standard for addressing many questions.
- I am happy to discuss what RNA-seq protocol is right for you—it can depend on your biological question

RNA-seq

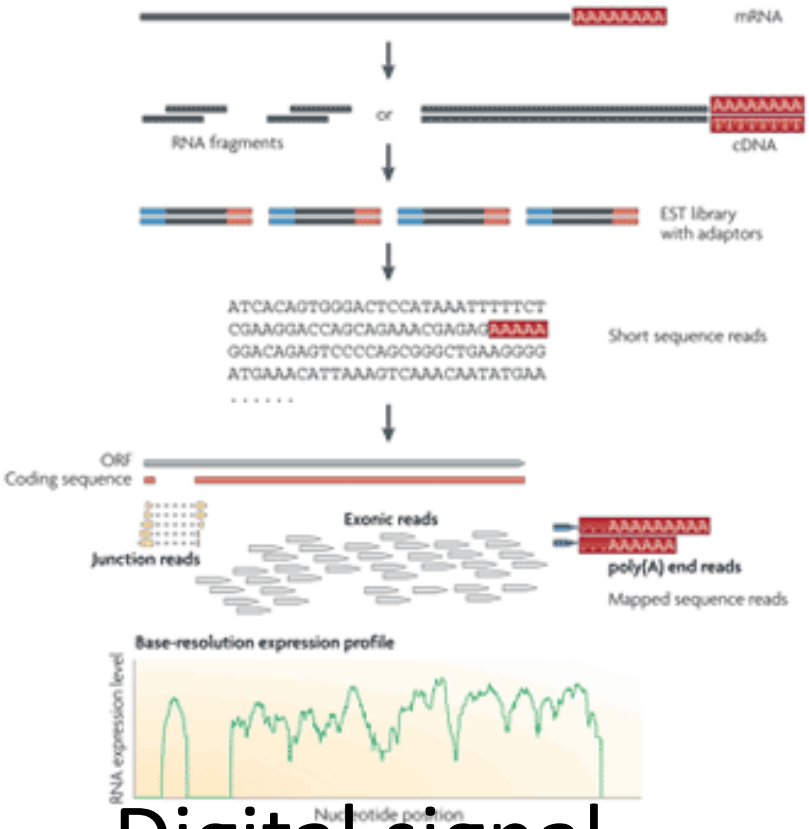
- Why?
- Considerations
  - Methods
  - Replicates
  - Mapping pipelines
  - Normalization
  - Differential Expression
    - Considerations
    - Software

Genomic measurements of RNA abundance

Microarray hybridization



RNA sequencing



Analog signal

Signal is a ratio of conditions

Relative abundance

Digital signal

Independent sample quantification

Closer to Absolute abundance (with spike-ins)

Why RNA-seq? More benefits and opportunities

- All transcripts are sequenced, not just ones for which probes are designed (e.g. microarrays)
- Can discover new exons, transcribed regions, genes or non-coding RNAs
- No cross-hybridization
- Digital readout (counting) instead of analog signal (ratios of hyb. signal)
- Can compare expression between genes
- Limited only by sequencing depth – detect low abundance transcripts
- Genuine whole transcriptome sequencing:
  - the ability to look at alternative splicing
  - allele-specific expression
  - RNA editing

Experimental and sequencing considerations

- Before library prep:
  - RNA population
  - Spike-in controls?
  - RNA quality
  - Type of kit or library prep method
  - Number of replicates

- After library prep:
  - Sequencing depth
  - Processing pipelines
  - Normalization methods
  - Differential gene expression analysis

https://genome.ucsc.edu/ENCODE/experiment_guidelines.html

Considerations: RNA population

- Poly-A+ RNA
  - Good for detecting mRNA
- Total RNA
  - Good for detecting non-coding RNA
  - Must remove rRNA (>80% of RNA in cell)
- Targeted RNA capture:
  - Disease-associated panels of genes
  - Detecting isoforms
  - Detecting low-abundant RNAs

Considerations: Spike-in controls

Resource

Synthetic spike-in standards for RNA-seq experiments

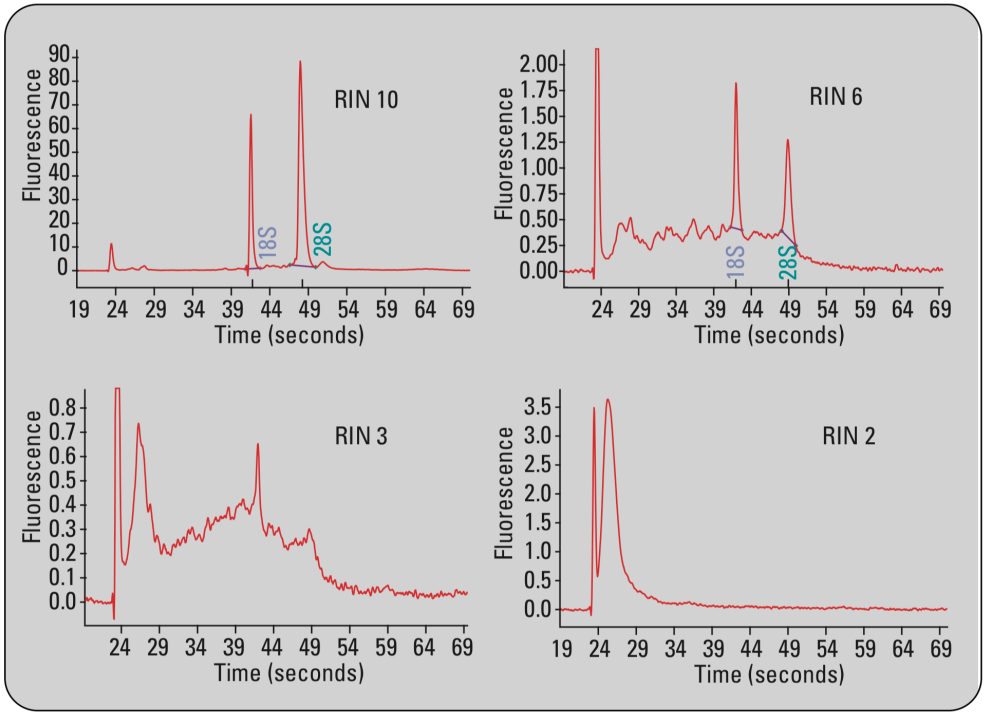
Lichun Jiang,[1,5] Felix Schlesinger,[2,3,5,7] Carrie A. Davis,[2] Yu Zhang,[1,6] Renhua Li,[1] Marc Salit,[4] Thomas R. Gingeras,[2] and Brian Oliver[1]

- Multi-group effort: External RNA Control Consortium (ERCC)
  - headed by National Institute of Standards and Technology (NIST)
- ERCC spike-ins are 96 synthetic RNAs with varying length, GC content, and 20 order of magnitude in concentration.
- Allow measurement of sensitivity, accuracy, and biases of RNA-seq
- Allow absolute quantification of RNAs and normalization between samples.
- Can make yourself, obtain clones from ERCC, or purchase from vendors.

http://genome.cshlp.org/content/21/9/1543.full.pdf+html

Considerations: RNA QC before library prep

- Total RNA Quality
  - BioAnalyzer RIN (RNA Integrity Number)
- Absence of genomic DNA
  - qPCR assay



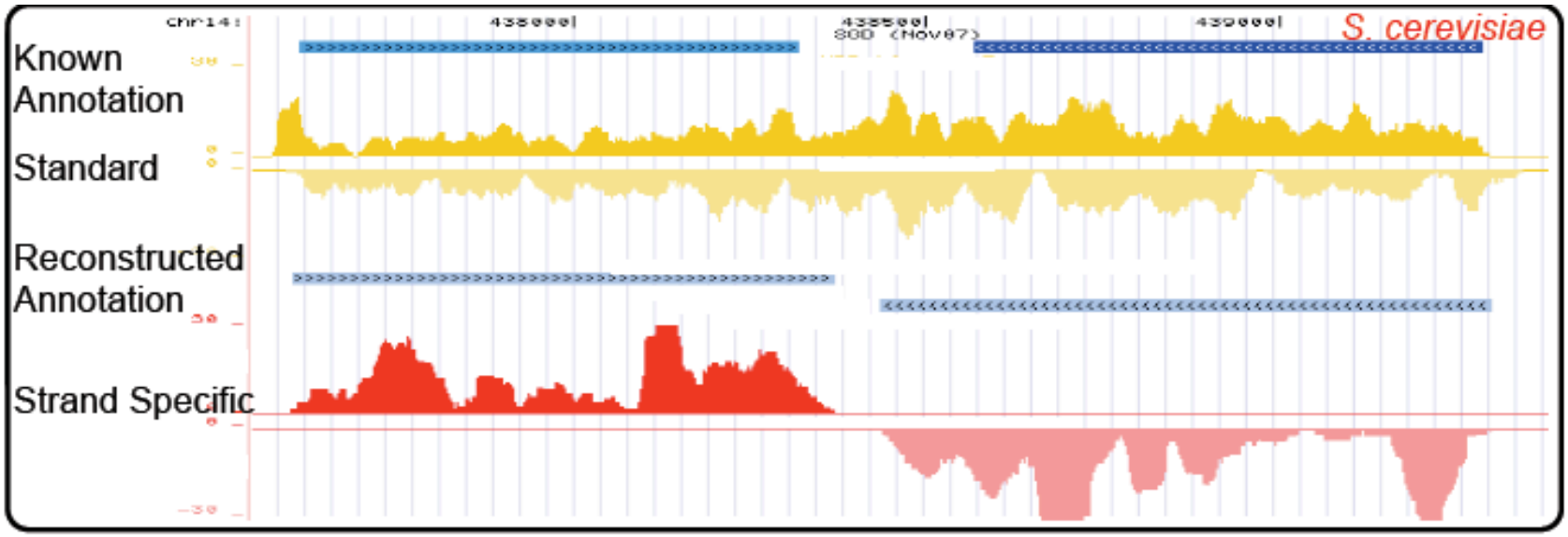
http://www.agilent.com/cs/library/applications/5989-1165EN.pdf

Considerations: RNA QC before library prep

- Total RNA Quality
  - BioAnalyzer RIN (RNA Integrity Number)
- Absence of genomic DNA
  - qPCR assay
- mRNA Purity
  - BioAnalyzer % rRNA < 5%
- mRNA Quantity
  - Minimum of 10 nanograms

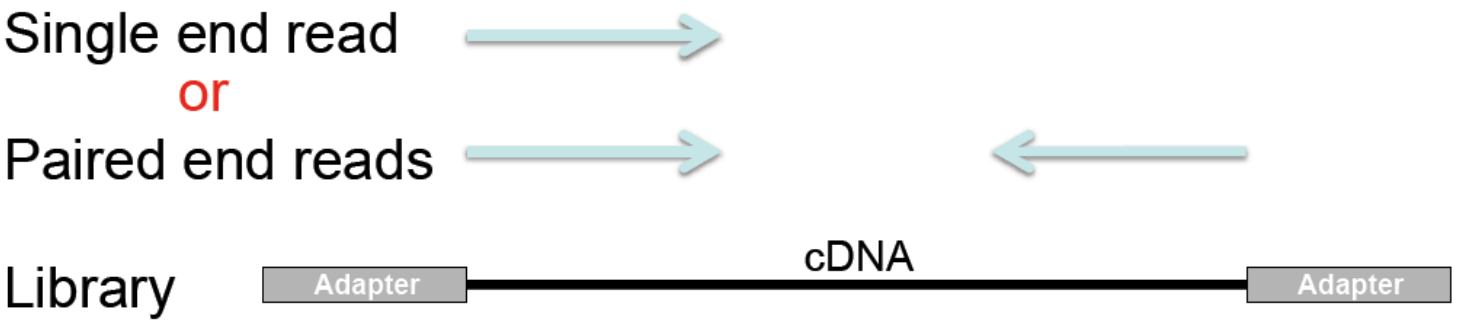
http://www.agilent.com/cs/library/applications/5989-1165EN.pdf

Considerations for library prep: strand specificity



- Identify strand of origin for non-coding RNA
- Identify antisense RNA
- Define ends of adjacent or overlapping transcripts transcribed in opposite directions

Considerations for library prep: Single or paired-end



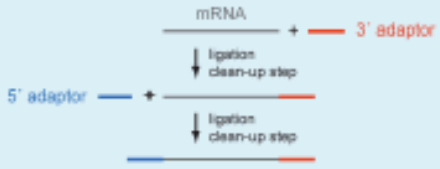
	Single End	Paired End
Cost	Lower	Higher
Sequencing Run Time	Shorter	Longer
Data per library	Less	More
Informativeness	Generally Less	Generally More

Strand specific methods and kits

RNA ligation

3' and 5' adaptors ligated sequentially to RNA with cleanup

Lister et al. (2008) *Cell*



"Illumina" RNA ligation

3' pre-adenylated adaptors and 5' adaptors ligated sequentially to RNA without cleanup

S. Luo & G. Schroth (pers. comm.)

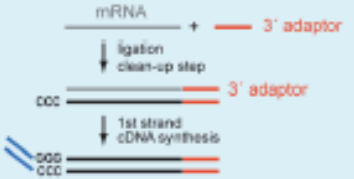
SMART (Switching Mechanism At 5´ end of RNA Template)

Non-template 'C's on 5´ end of cDNA; template switching, PCR

Cloonan et al. (2008) *Nat. Methods*

SMART – RNA ligation

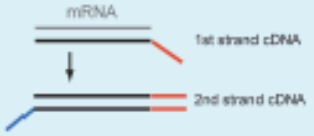
Adaptor ligated on 3' end of RNA; non-template 'C's on 5' end of cDNA; template switching, PCR



NNSR (Not Not So Random priming)

RT – 1st & 2nd strand cDNA synthesis with adaptors on ends of random primers

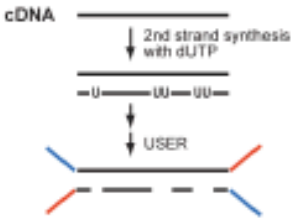
Armour et al. (2009) *Nat. Methods*



dUTP 2nd strand

2nd strand synthesis with dUTP, remove 'U's after adaptor ligation and size selection

Parkhomchuk et al. (2009) *NAR*



Bisulfite

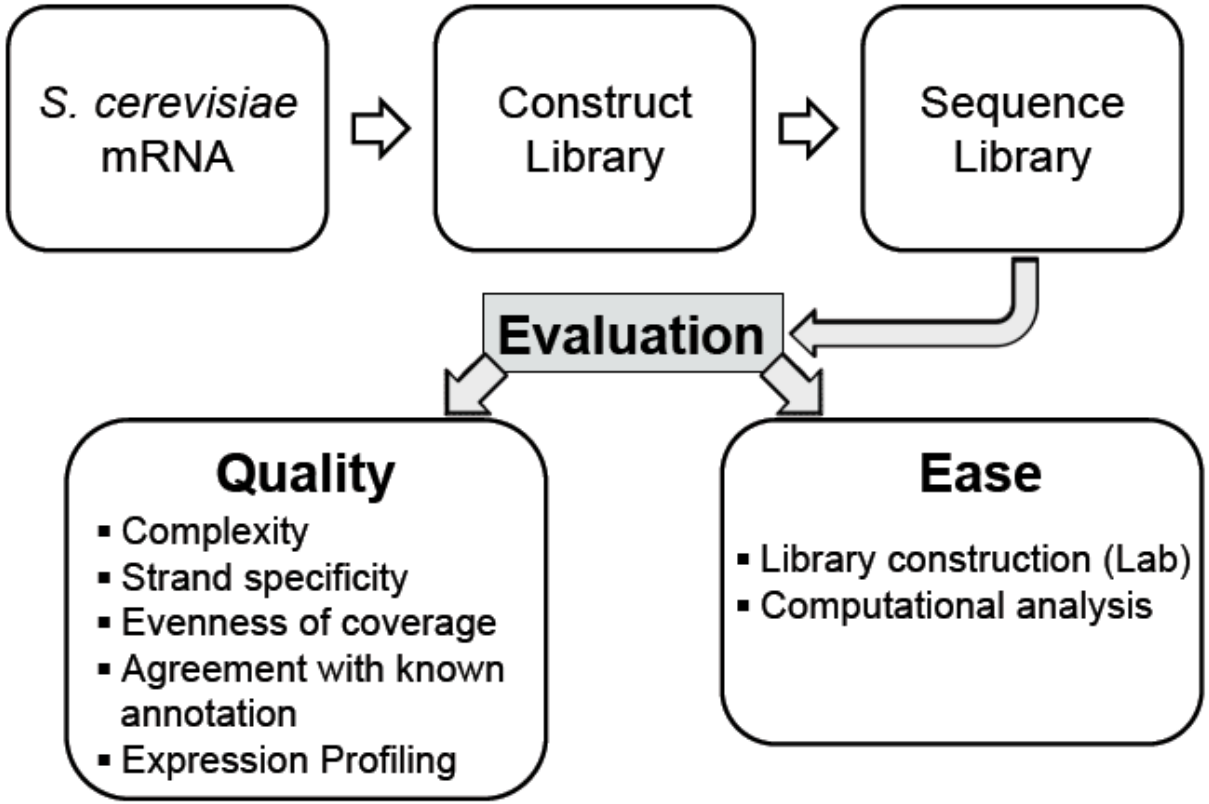
Convert 'C's to 'U's in RNA

He et al. (2008) *Science*
Schaefer et al. (2009) *NAR*

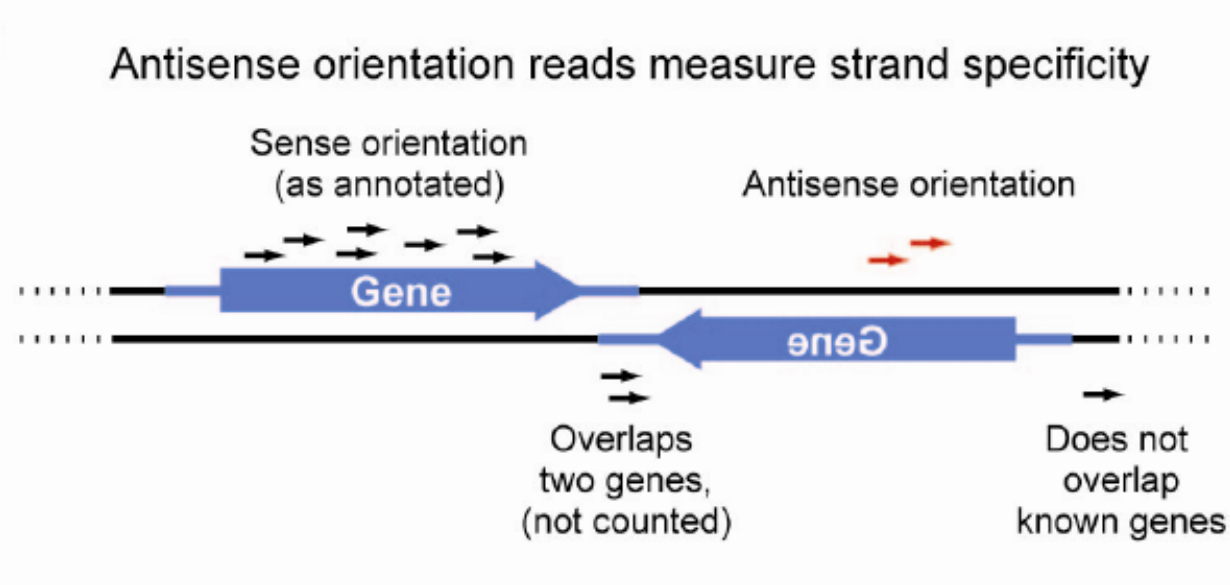
Evaluating RNA-seq library preparation methods

Comprehensive comparative analysis of strand-specific RNA sequencing methods

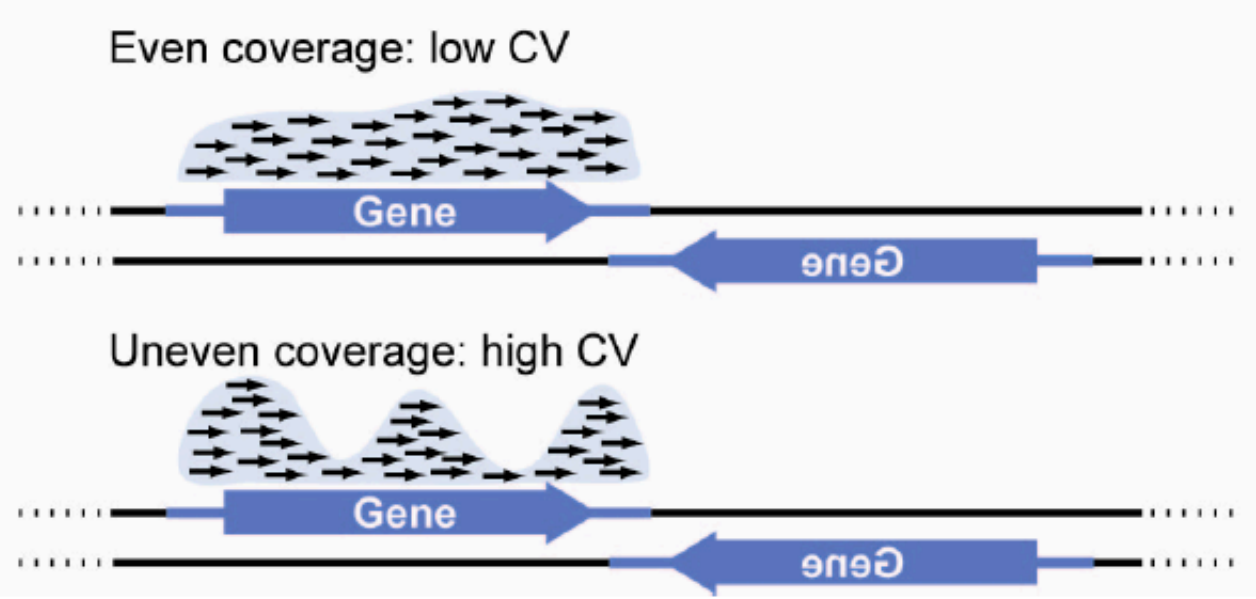
Joshua Z Levin[1,6], Moran Yassour[1–3,6], Xian Adiconis[1], Chad Nusbaum[1], Dawn Anne Thompson[1], Nir Friedman[3,4], Andreas Gnirke[1] & Aviv Regev[1,2,5]



Evaluating RNA-seq methods: Strand specificity

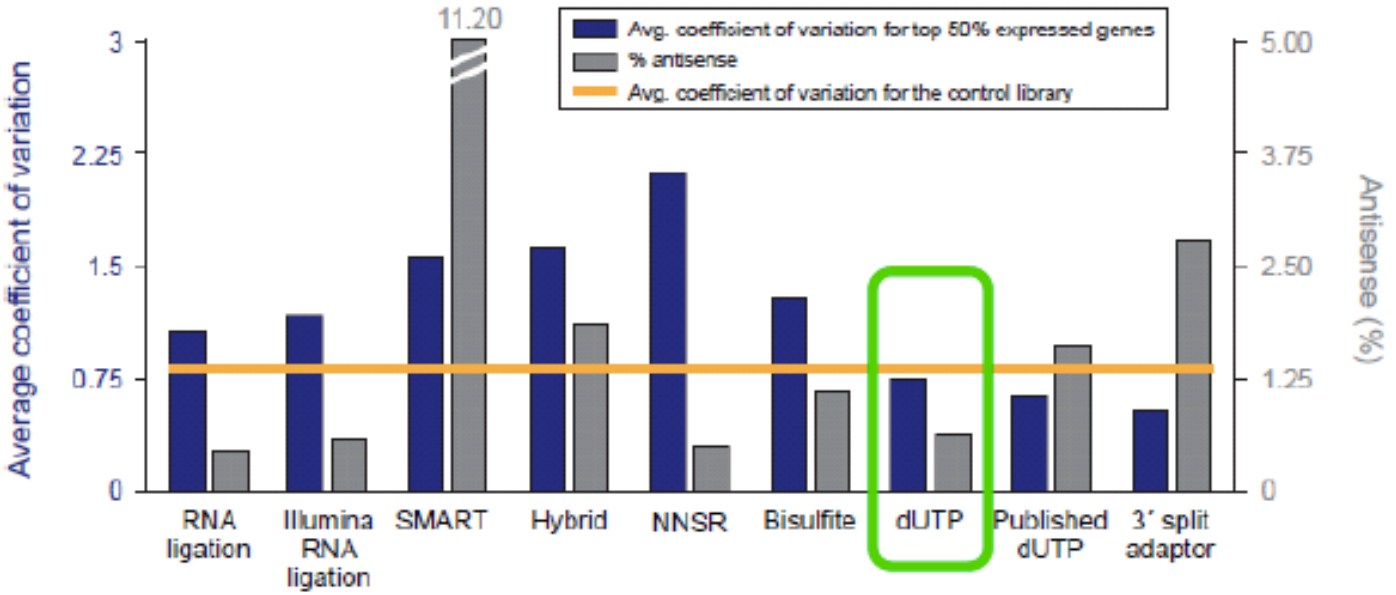


Evaluating RNA-seq methods: Evenness of Coverage



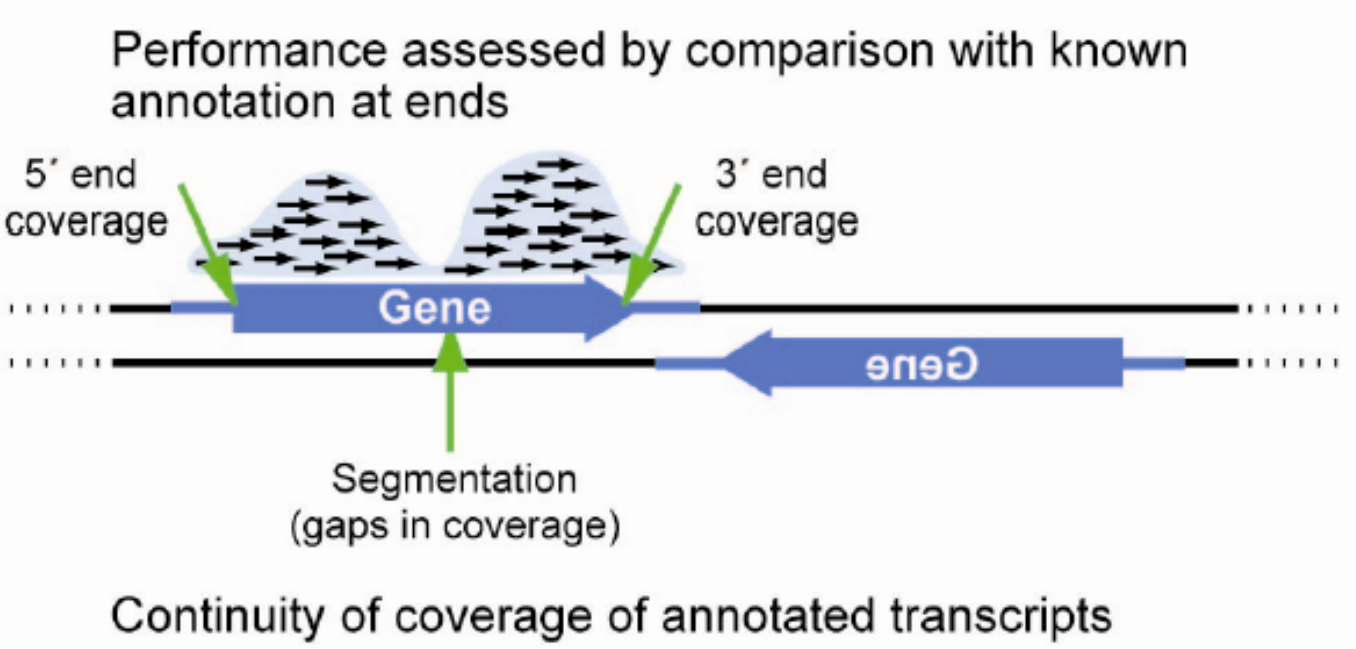
- Coefficient of Variation (CV) = standard deviation / mean
  & is a measure of evenness

Strand specificity and Evenness of Coverage



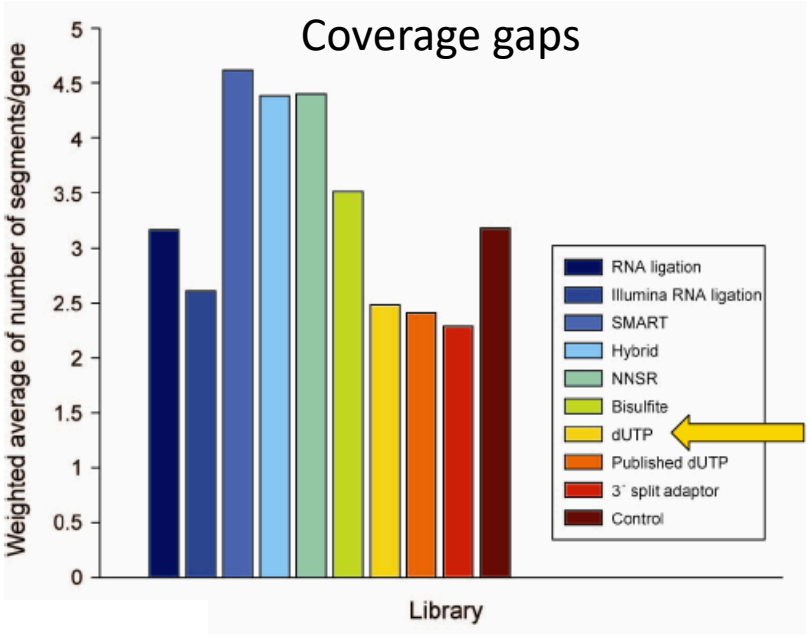
- **For both measures, lower is better**
- **dUTP library performs best**

Evaluating RNA-seq methods: coverage gaps

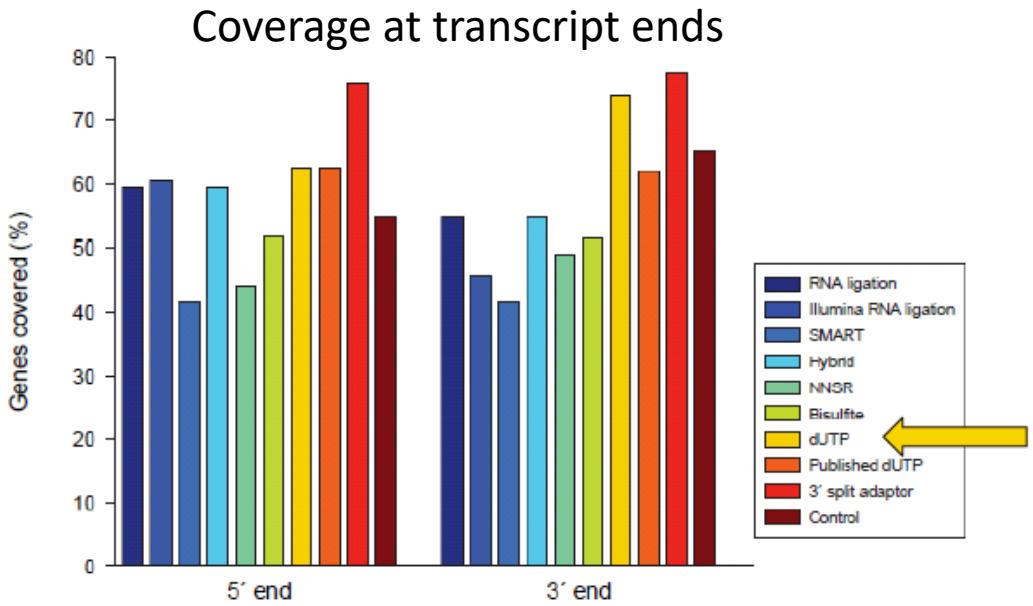


Evaluating RNA-seq methods: coverage gaps

Lower is better



Higher is better



Molecular Biology for RNA-seq: this approach will probably work for you

1) RNA isolation

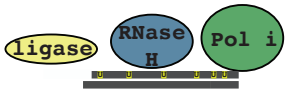
2) Negative selection of rRNA

3) Fragment RNA

4) Anneal random hexamer DNA primers

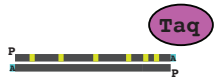
5) Reverse Transcription

6) Second strand synthesis



7) Blunt end repair and phosphorylate 5′

8) A-tail



9) Ligate sequencing adapters

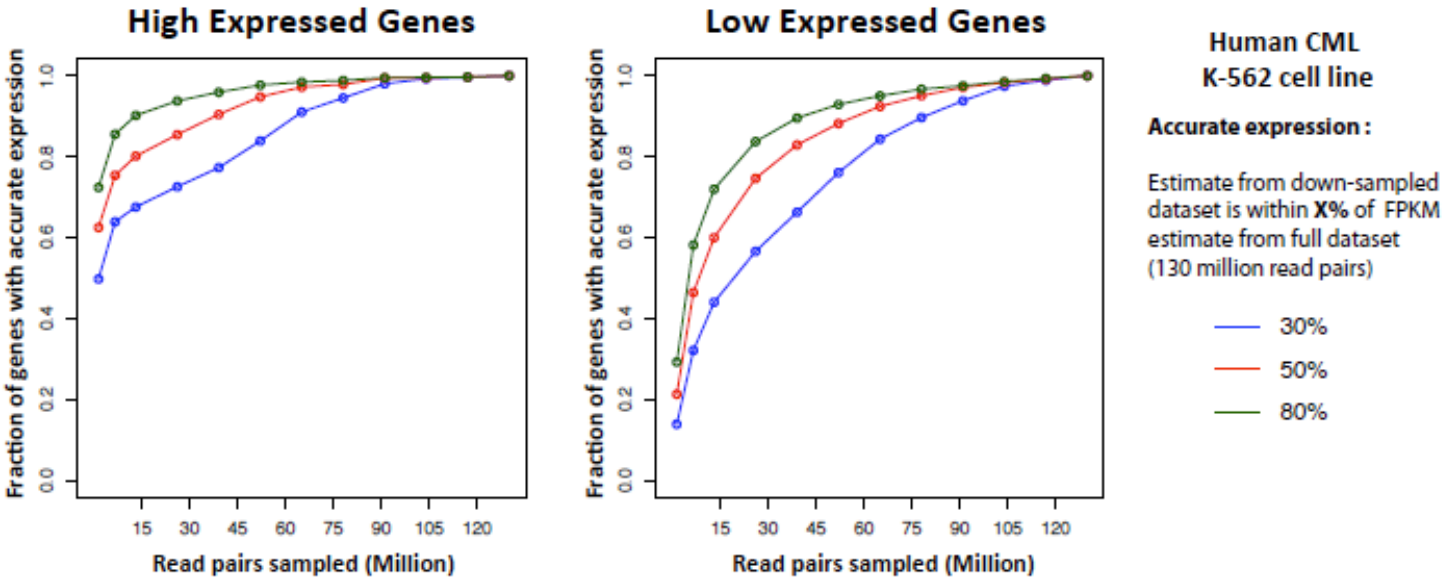
10) Digest dUTP strand

11) PCR

First cycle: only one primer anneals.

Second cycle: generates a product representing a single strand of what will be the final amplicon.

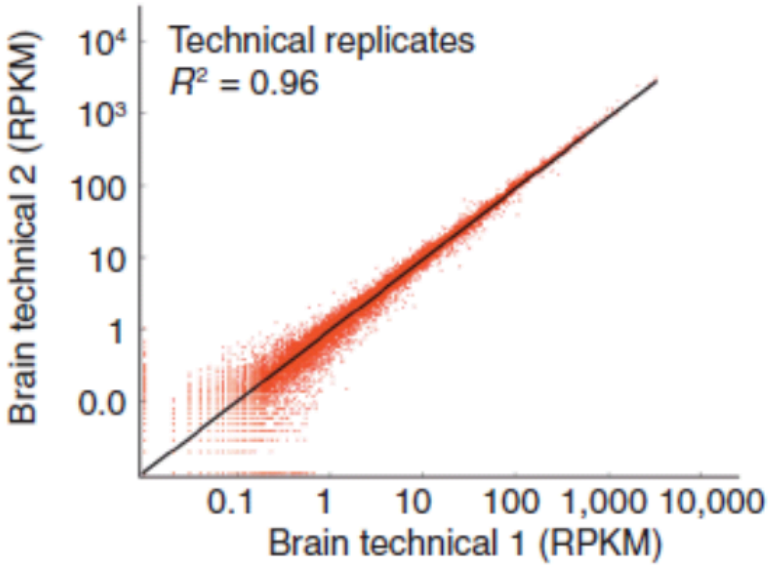
Considerations: sequencing depth



Source: Rahul Satija & Joshua Levin

- **More coverage needed to accurately measure levels for low expressed genes**
- **30M read pairs probably sufficient for expression levels (in this case)**
- **More needed for splicing isoform levels or allele-specific expression**

Considerations: biological and technical replicates



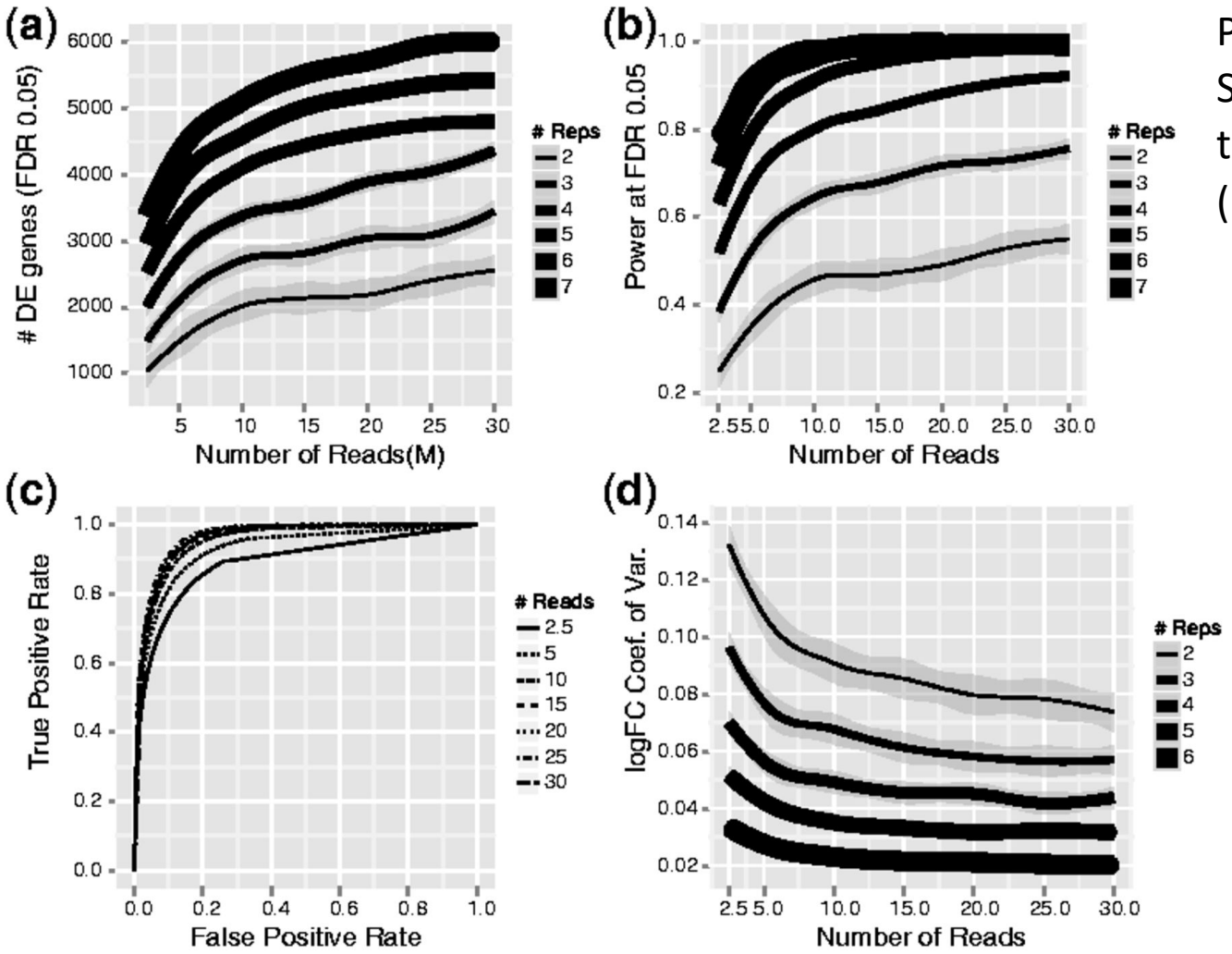
Example from Mortazavi et al. (2008) *Nature Methods* **5**:621

Technical replicates – sequencing 2 independent cDNA libraries from the same RNA – usually not necessary

Biological replicates – necessary

Additional reading: Auer & Doerge (2010) "Statistical design and analysis of RNA sequencing data" *Genetics* 185:405

(a) Increase in biological replication significantly increases the number of DE genes identified.



Power:
Sensitivity of test
to detect true effects
(probability)

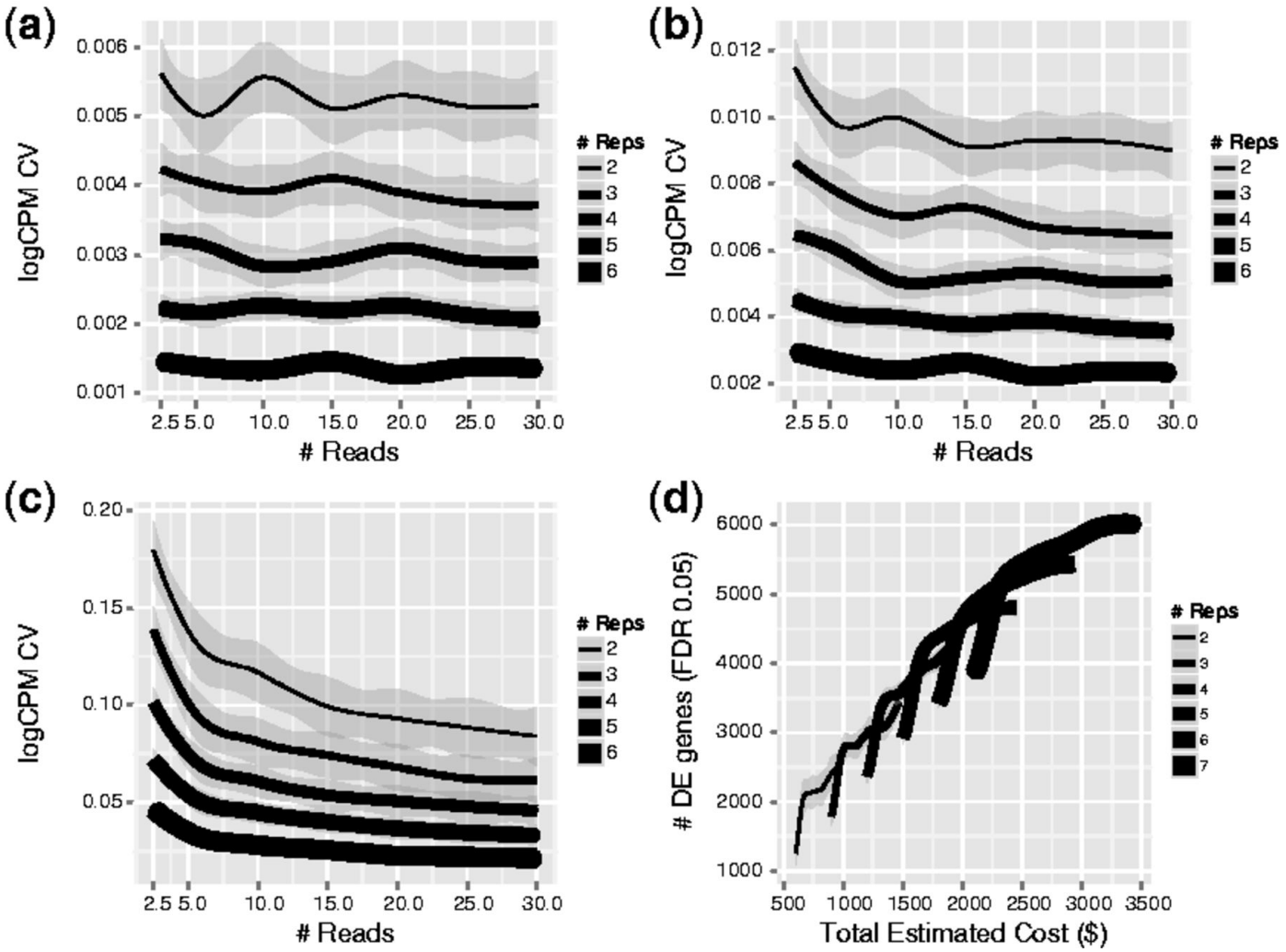
Coeff. of Var (CV):
(SD/mean)

Yuwen Liu et al. Bioinformatics 2014;30:301-304

© The Author 2013. Published by Oxford University Press. All rights reserved. For Permissions, please e-mail: journals.permissions@oup.com

Bioinformatics

(a–c) The CV of logCPM for high expression level genes (a), medium expression level genes (b) and low expression level genes (c) (see Section 2 for definition).



Yuwen Liu et al. Bioinformatics 2014;30:301-304

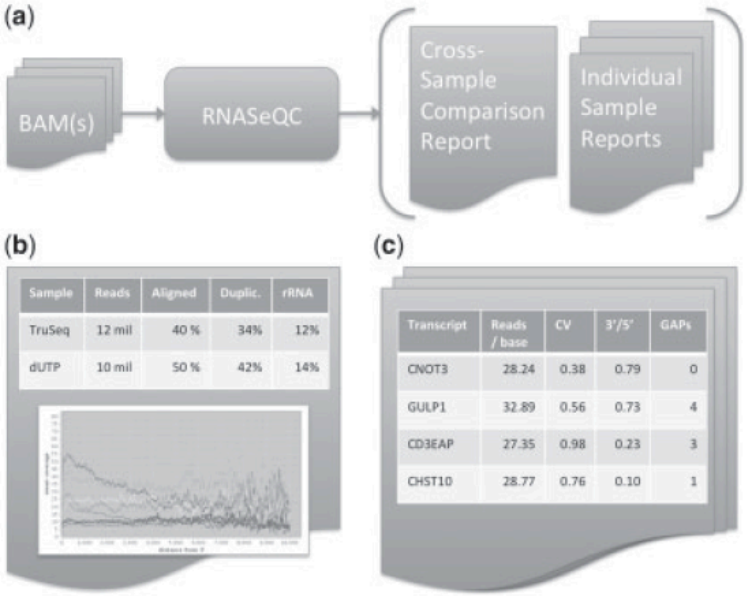
© The Author 2013. Published by Oxford University Press. All rights reserved. For Permissions, please e-mail: journals.permissions@oup.com

Bioinformatics

Post-sequencing QC and analysis

- After library prep:
  - RNA-seq Quality control
  - Processing pipelines
  - Normalization methods
  - Differential gene expression analysis

RNASeQC - quality control pipeline



Sample	Reads	Aligned	Duplic.	rRNA
TruSeq	12 mil	40 %	34%	12%
dUTP	10 mil	50 %	42%	14%

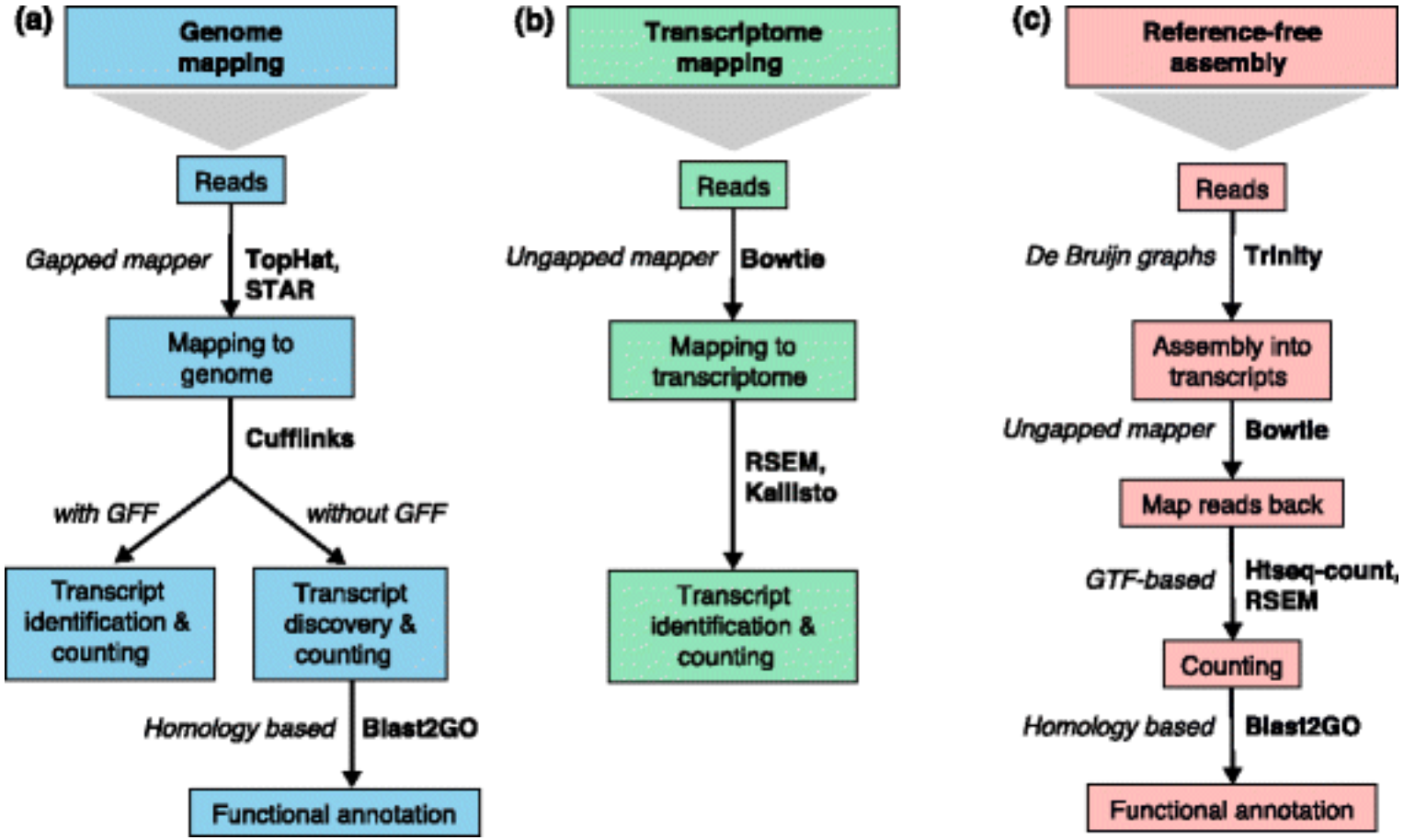
Transcript	Reads / base	CV	3'/5'	GAPs
CNOT3	28.24	0.38	0.79	0
GULP1	32.89	0.56	0.73	4
CD3EAP	27.35	0.98	0.23	3
CHST10	28.77	0.76	0.10	1

- Total, unique and duplicate reads
- Mapped reads and mapped unique reads
- rRNA reads
- Transcript-annotated reads:
- Expressed transcripts: count of transcripts with reads ≥1.
- Strand specificity

- Sample reports: calculates a number of metrics useful for assessing quality of libraries and depth of sequencing.
- Comparison of metrics between samples.

http://bioinformatics.oxfordjournals.org/content/28/11/1530.long

Variations of RNA-seq mapping strategies

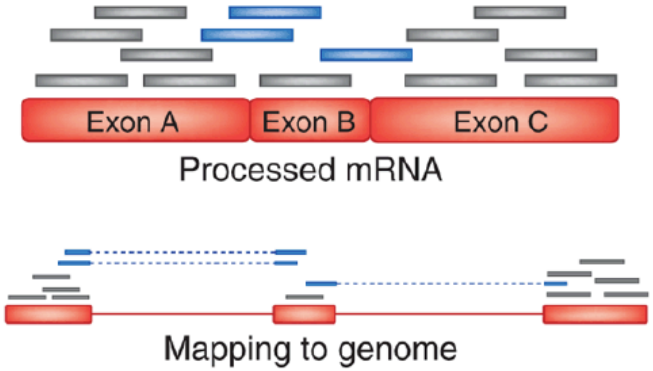


Mapping issues/options

Issues

- With RNA-seq we really want to align to the transcriptome.
- Splice junction reads will not align to the genome.
- The longer the reads, the more likely one will hit a junction.

- Alignment of genomic sequencing vs RNA-seq



Cole Trapnell & Steven L Salzberg, *Nature Biotechnology* **27**, 455 - 457 (2009)

Mapping issues/options

Issues

- With RNA-seq we really want to align to the transcriptome.
- Splice junction reads will not align to the genome.
- The longer the reads, the more likely one will hit a junction.

Options

- Don’t worry about it, align to the genome.
- Build a junction library, and align to that.
- Create your own transcriptome “de novo”.
- Combination of first two or all three

Tuxedo tools software suite

Bowtie (fast short-read alignment)

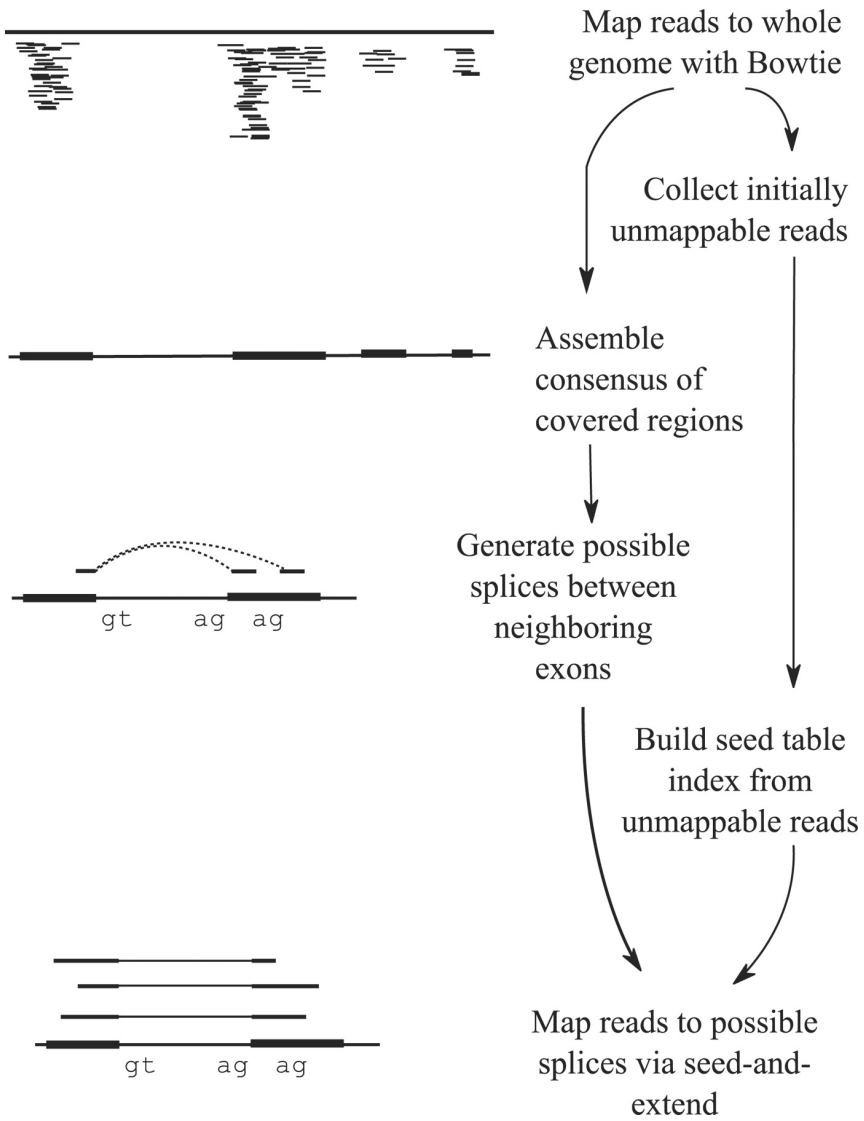
TopHat (spliced short-read alignment)

Cufflinks (transcript reconstruction from alignments)

Cuffdiff (differential expression analysis)

CummeRbund (visualization & analysis)

The TopHat pipeline for de novo splice junction discovery.



© 2009 The Author(s)

Cole Trapnell et al. Bioinformatics 2009;25:1105-1111

Mapping issues/options: Alignment options

BIOINFORMATICS ORIGINAL PAPER

Vol. 25 no. 9 2009, pages 1105–1111
doi:10.1093/bioinformatics/btp120

Sequence analysis

TopHat: discovering splice junctions with RNA-Seq

Cole Trapnell[1,*], Lior Pachter[2] and Steven L. Salzberg[1]

[1]Center for Bioinformatics and Computational Biology, University of Maryland, College Park, MD 20742 and [2]Department of Mathematics, University of California, Berkeley, CA 94720, USA

Received on October 23, 2008; revised on February 24, 2009; accepted on February 26, 2009

Advance Access publication March 16, 2009

Associate Editor: Ivo Hofacker

BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 1 2013, pages 15–21
doi:10.1093/bioinformatics/bts635

Sequence analysis

Advance Access publication October 25, 2012

STAR: ultrafast universal RNA-seq aligner

Alexander Dobin[1,*], Carrie A. Davis[1], Felix Schlesinger[1], Jorg Drenkow[1], Chris Zaleski[1], Sonali Jha[1], Philippe Batut[1], Mark Chaisson[2] and Thomas R. Gingeras[1]

[1]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, USA and [2]Pacific Biosciences, Menlo Park, CA, USA

Associate Editor: Inanc Birol

HISAT: a fast spliced aligner with low memory requirements

Daehwan Kim[1,2], Ben Langmead[1–3] & Steven L Salzberg[1–3]

[1]Center for Computational Biology, McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA. [2]Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, Maryland, USA. [3]Department of Computer Science, Johns Hopkins University, Baltimore, Maryland, USA. Correspondence should be addressed to D.K. (infphilo@gmail.com), B.L. (langmea@cs.jhu.edu) or S.L.S. (salzberg@jhu.edu).

RECEIVED 7 AUGUST 2014; ACCEPTED 16 JANUARY 2015; PUBLISHED ONLINE 9 MARCH 2015; DOI:10.1038/NMETH.3317

Mapping issues/options: Alignment options

Table 1. Mapping speed and RAM benchmarks on the experimental RNA-seq dataset.

Aligner	Mapping speed: Million read pairs / hour		Peak physical RAM, GB	
	6 threads	12 threads	6 threads	12 threads
STAR	309.2	549.9	27.0	28.4
STAR sparse	227.6	423.1	15.6	16.0
TopHat2	8.0	10.1	4.1	11.3
RUM	5.1	7.6	26.9	53.8
MapSplice	3.0	3.1	3.3	3.3
GSNAP	1.8	2.8	25.9	27.0

Star is much faster, but requires more resources (RAM)

HISAT2 was released after this benchmarking and it is my preferred RNA-seq mapping software

HISAT2:

HISAT: a fast spliced aligner with low memory requirements

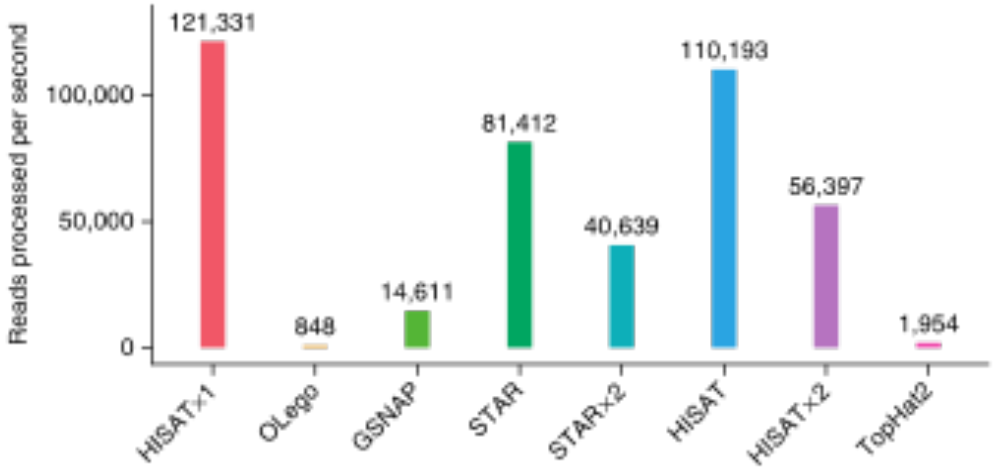
Daehwan Kim[1,2], Ben Langmead[1–3] & Steven L Salzberg[1–3]

[1]Center for Computational Biology, McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA. [2]Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, Maryland, USA. [3]Department of Computer Science, Johns Hopkins University, Baltimore, Maryland, USA. Correspondence should be addressed to D.K. (infphilo@gmail.com), B.L. (langmea@cs.jhu.edu) or S.L.S. (salzberg@jhu.edu).

RECEIVED 7 AUGUST 2014; ACCEPTED 16 JANUARY 2015; PUBLISHED ONLINE 9 MARCH 2015; DOI:10.1038/NMETH.3317

NATURE METHODS | VOL.12 NO.4 | APRIL 2015 | 357

Hierarchical Indexing for Spliced Alignment of Transcripts



HISAT2 is faster

Table 2 | Run times and memory usage for HISAT and other spliced aligners

Program	Run time (min)	Memory usage (GB)
HISATx1	22.7	4.3
HISATx2	47.7	4.3
HISAT	26.7	4.3
STAR	25	28
STARx2	50.5	28
GSNAP	291.9	20.2
OLego	989.5	3.7
TopHat2	1,170	4.3

Run times and memory usage for HISAT and other spliced aligners to align 109 million 101-bp RNA-seq reads from a lung fibroblast data set. We used three CPU cores to run the programs on a Mac Pro with a 3.7 GHz Quad-Core Intel Xeon E5 processor and 64 GB of RAM.

HISAT2 uses less memory space

HISAT2: workflow with new tuxedo tools

PROTOCOL

Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown

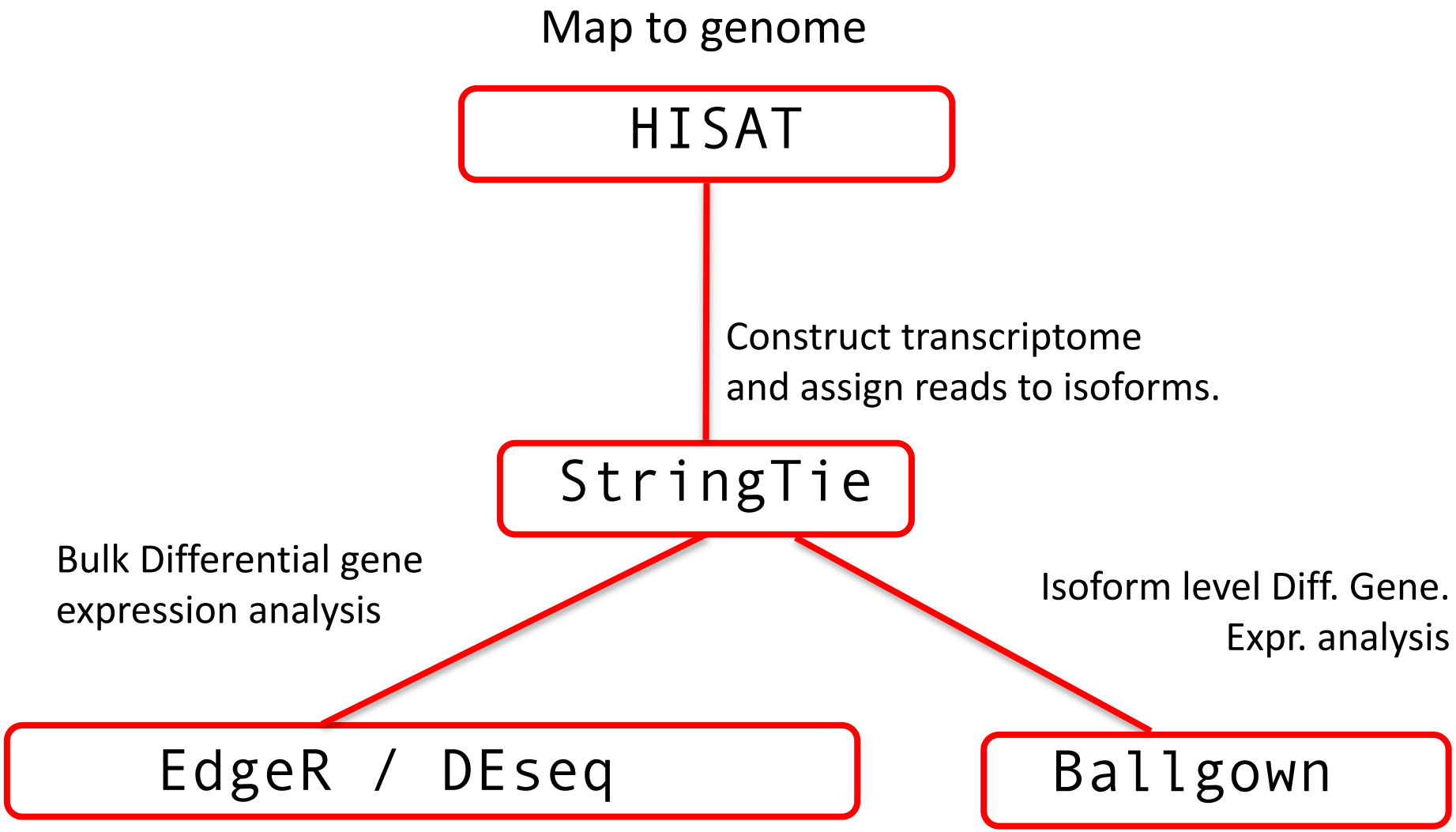
Mihaela Pertea[1,2], Daehwan Kim[1], Geo M Pertea[1], Jeffrey T Leek[3] & Steven L Salzberg[1–4]

StringTie enables improved reconstruction of a transcriptome from RNA-seq reads

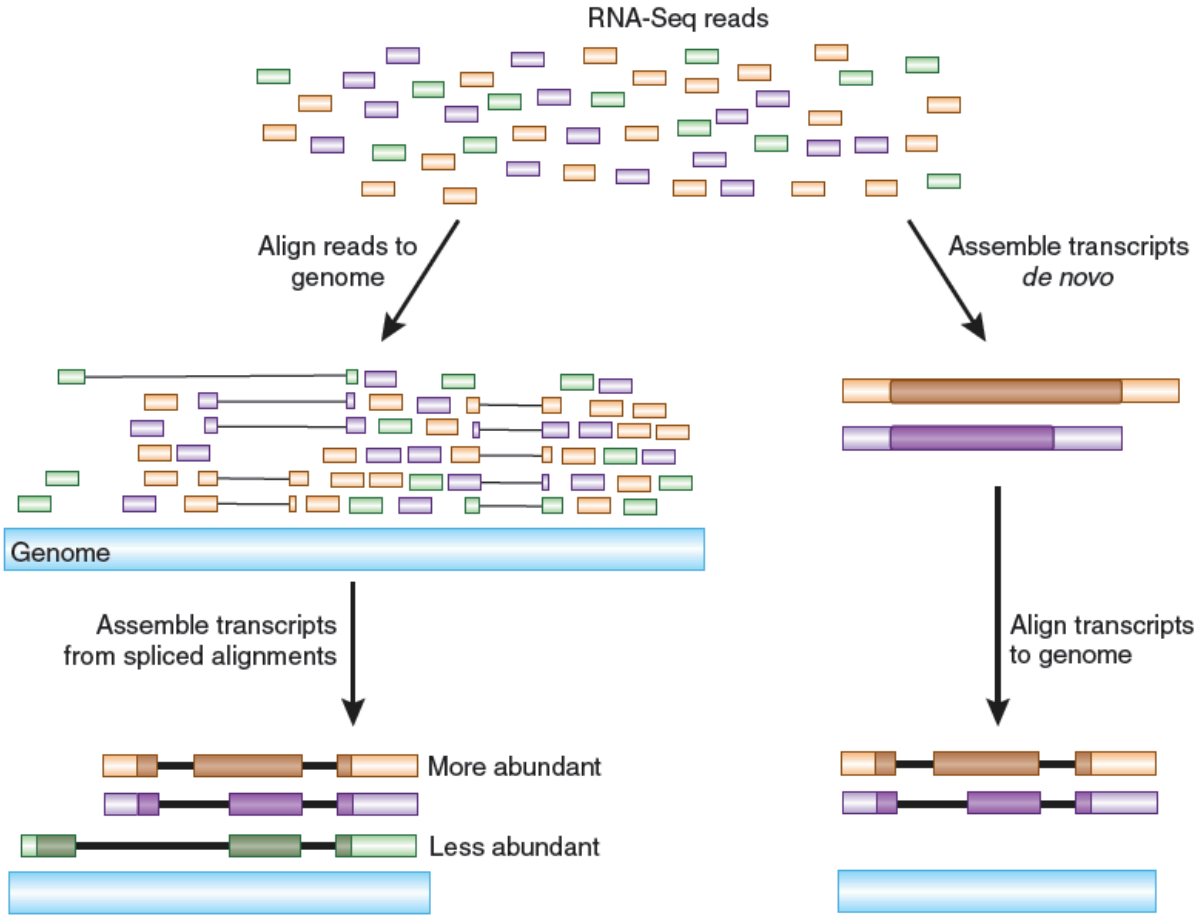
Mihaela Pertea[1,2], Geo M Pertea[1,2], Corina M Antonescu[1,2], Tsung-Cheng Chang[3,4], Joshua T Mendell[3–5] & Steven L Salzberg[1,2,6,7]

Ballgown bridges the gap between transcriptome assembly and expression analysis

HISAT2: workflow with new tuxedo tools



De novo transcript assembly with or without a genome



Haas and Zody, NBT, 2010

Trinity

Grabherr, Haas, & Yassour et al., Nature Biotechnology, 2011

Normalization methods

- Total count Normalization (FPKM, RPKM, TPM)
  - By total mapped reads
    - F= unique Fragments
    - R= Reads
    - T = Transcripts
- Upper quartile normalization
  - Read count of genes in upper quartile
- Housekeeping genes
- Trimmed mean (TMM) normalization

Added level:

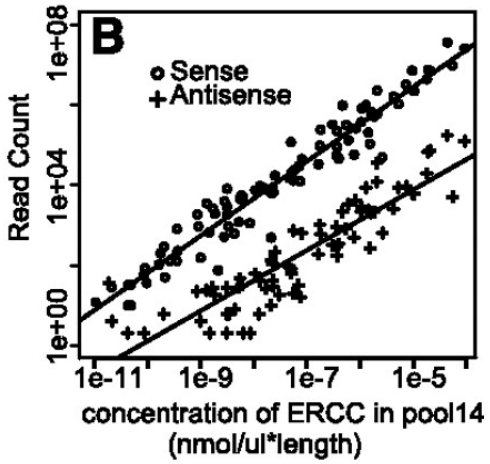
- Spike-in controls

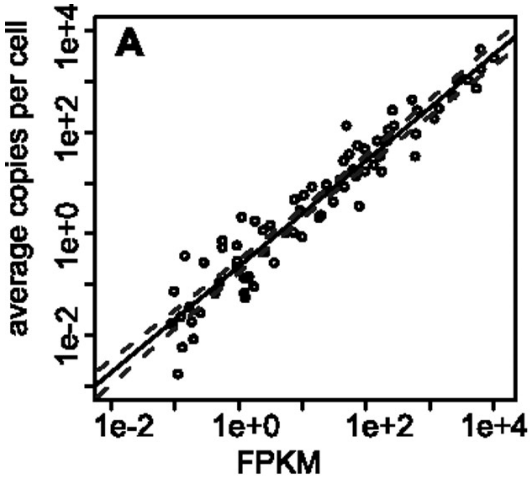
Library characteristics, ERCC quantification, and coverage, transcript counting.

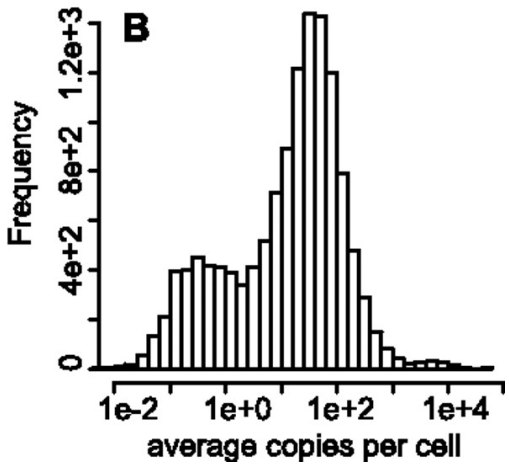
ERCC = External RNA Control Consortium

Linear standards

Allow calculating average transcripts /cell



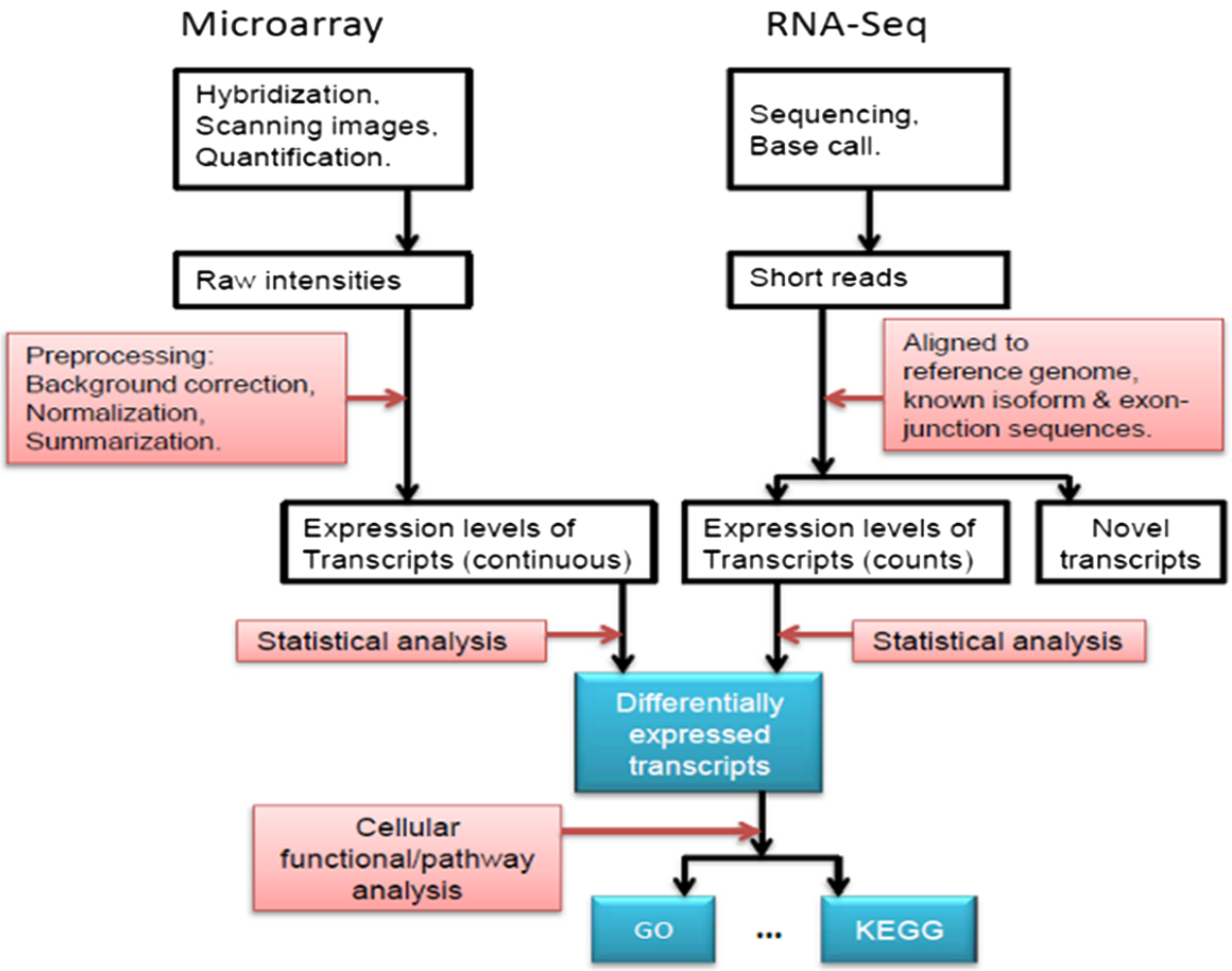




Lichun Jiang et al. Genome Res. 2011;21:1543-1551

©2011 by Cold Spring Harbor Laboratory Press

Gene expression analysis workflows



http://www.labome.com/method/RNA-seq-Using-Next-Generation-Sequencing.html

Differential gene expression analysis: 3 popular methods

Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks

Cole Trapnell[1,2], Adam Roberts[3], Loyal Goff[1,2,4], Geo Pertea[5,6], Daehwan Kim[5,7], David R Kelley[1,2], Harold Pimentel[3], Steven L Salzberg[5,6], John L Rinn[1,2] & Lior Pachter[3,8,9]

Love et al. Genome Biology (2014) 15:550
DOI 10.1186/s13059-014-0550-8

Genome Biology

METHOD Open Access

Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2

Michael I Love[1,2,3], Wolfgang Huber[2] and Simon Anders[2*]

BIOINFORMATICS APPLICATIONS NOTE
Vol. 26 no. 1 2010, pages 139–140
doi:10.1093/bioinformatics/btp616

Gene expression

edgeR: a Bioconductor package for differential expression analysis of digital gene expression data

Mark D. Robinson[1,2,*,†], Davis J. McCarthy[2,†] and Gordon K. Smyth[2]

DE Analysis: Options and trade offs

Table 1 Number of false differential expression genes predicted by each method at adjusted *P* values (or false discovery rate) ≤0.05 separated by gene read count quantiles.

Expression quantile	Cuffdiff	DESeq	edgeR	limmaQN	limmaVoom	PoissonSeq	baySeq
100% (high expression)	28	5	3	0	0	7	1
75%	76	6	0	0	0	0	0
50%	84	27	1	2	0	0	0
25% (low expression)	5	9	0	87	0	0	0
Total	193	47	4	89	0	7	1

Table 2 Comparison of methods.

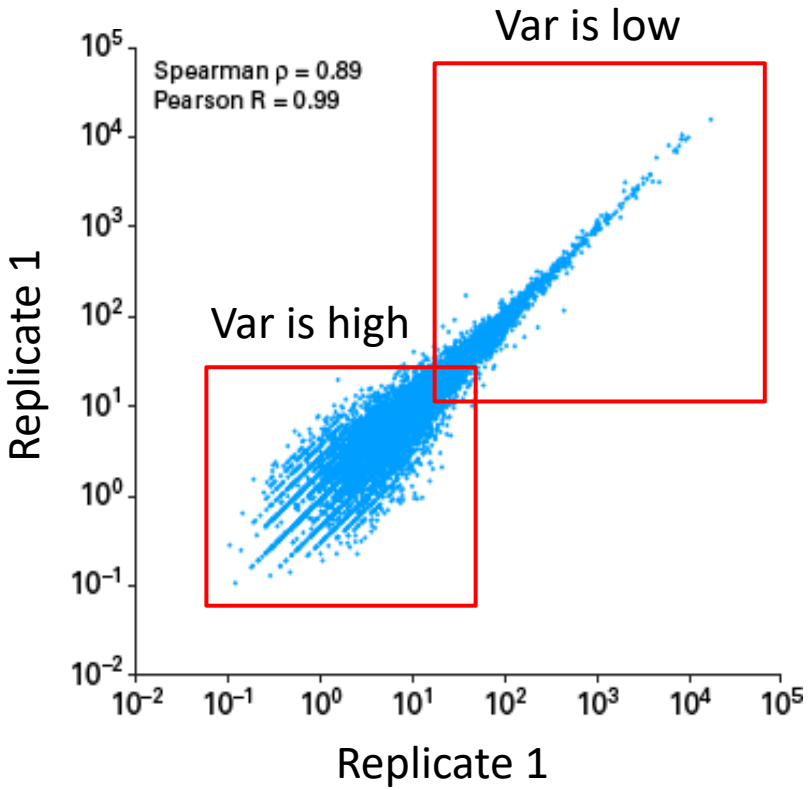
Evaluation	Cuffdiff	DESeq	edgeR	limmaVoom	PoissonSeq	baySeq
Normalization and clustering	All methods performed equally well					
DE detection accuracy measured by AUC at increasing qRT-PCR cutoff	Decreasing	Consistent	Consistent	Decreasing	Increases up to log expression change ≤ 2.0	Consistent
Null model type I error	High number of FPs	Low number of FPs	Low number of FPs	Low Number of FPs	Low number of FPs	Low number of FPs
Signal-to-noise vs *P* value correlation for genes detected in one condition	Poor	Poor	Poor	Good	Moderate	Good
Support for multi-factored experiments	No	Yes	Yes	Yes	No	No
Support DE detection without replicated samples	Yes	Yes	Yes	No	Yes	No
Detection of differential isoforms	Yes	No	No	No	No	No
Runtime for experiments with three to five replicates on a 12 dual-core 3.33 GHz, 100 G RAM server	Hours	Minutes	Minutes	Minutes	Seconds	Hours

AUC, area under curve; DE, differential expression; FP, false positive.

Rapaport F et al. Genome Biology, 2013 14:R95

Read depth and DE expression calling

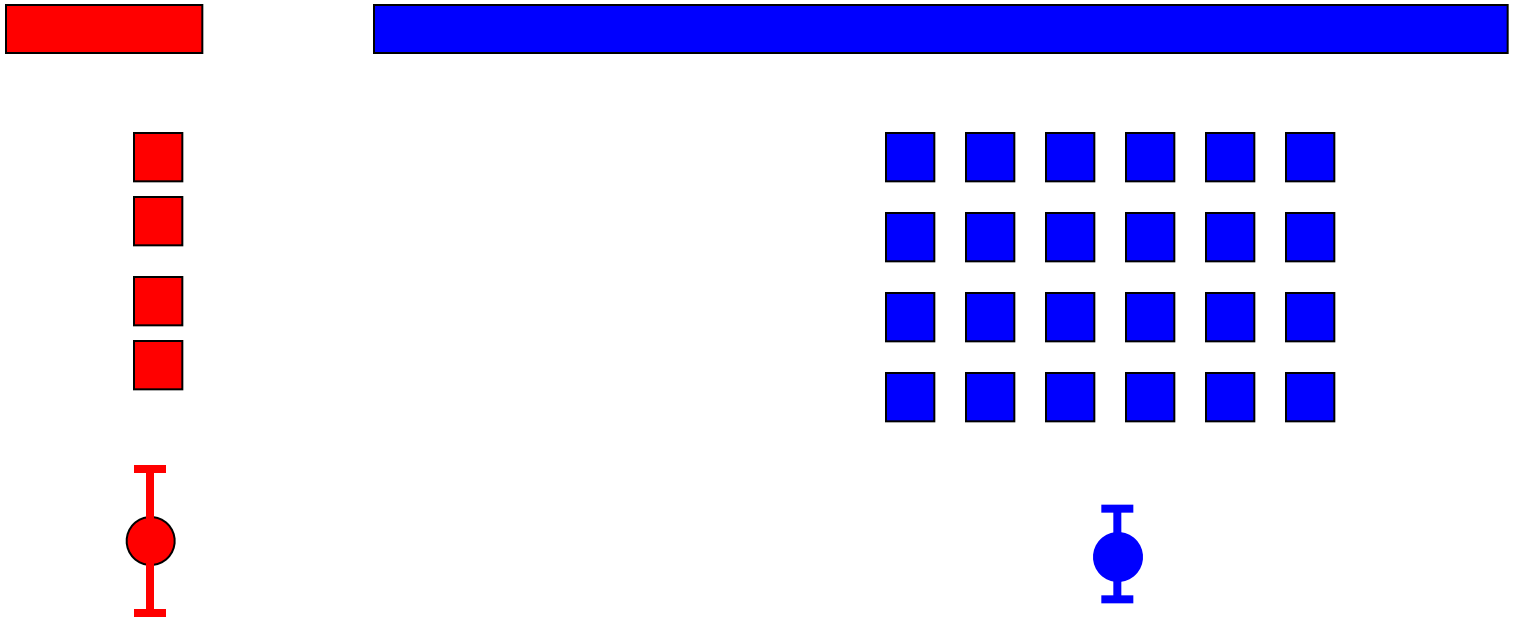
More variation in genes with lower counts



Fewer reads = higher variance

Fewer reads = less power to call DE

Length bias in RNA-seq



For genes of the same expression level longer transcripts will have more reads

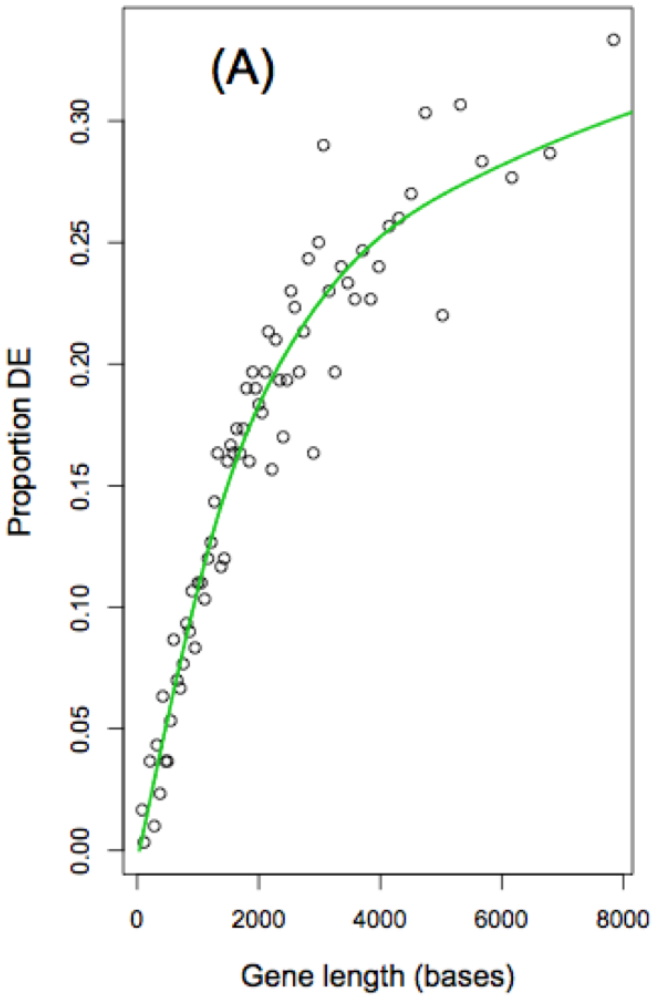
- There is more information for longer transcripts than shorter ones.
- Higher power to detect DE in longer transcripts.
- This length bias is not present in microarray gene expression data.

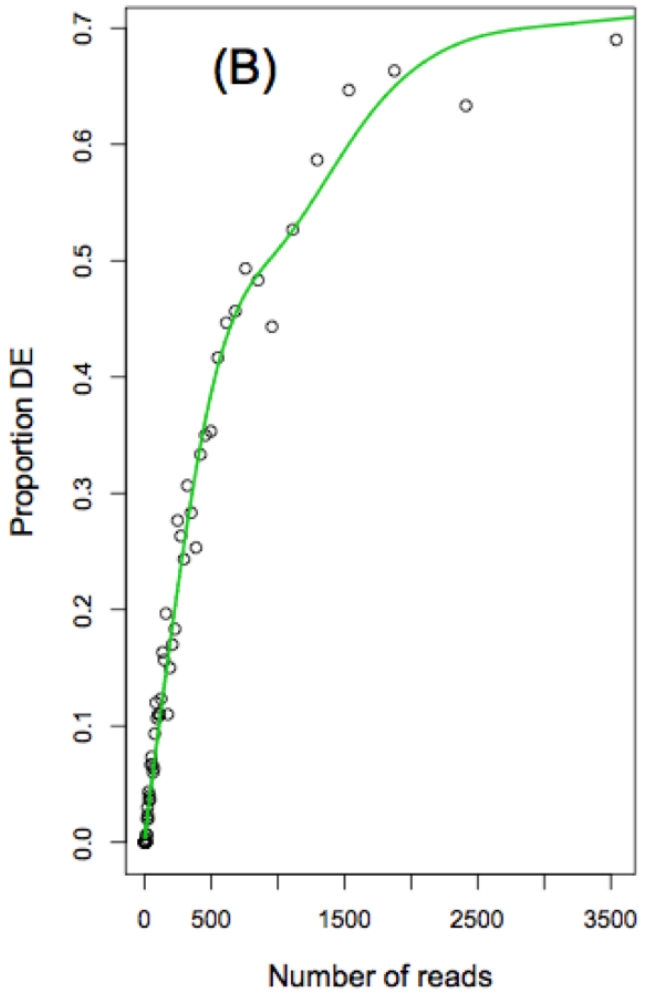
Proportion of DE genes v gene length, # reads

3 lanes treated vs
4 untreated.

LNCaP cells

Fisher exact
test, FDR ≤ 10^{-4}.





Analysis: Young *et al*, Gen. Biol. 2010. Data : Lovci *et al*, PNAS 2008

Dealing with the length bias

- It is not easy to do anything about this bias without throwing away data.
- It has the capacity to bias downstream gene set and Gene Ontology analyses.
- Young et al, Gen. Biol. 2010 and others have shown using p-values weighted on curves similar to previous slides (based on gene lengths), can help to alleviate some of the bias.
- The problem is still largely ignored by most of the standard DE expression software. However, DESeq2 recently added including gene length as an optional parameter.

Useful links for transcriptome and DE analysis

Video:

WATCH THIS: https://www.youtube.com/watch?v=5NiFibnbE8o#action=share

Papers:

TopHat/cufflinks:

http://www.nature.com/nprot/journal/v7/n3/pdf/nprot.2012.016.pdf

HISAT/StringTie/Ballgown – papers and tutorial

http://www.nature.com/nmeth/journal/v12/n4/full/nmeth.3317.html

http://www.nature.com/nbt/journal/v33/n3/full/nbt.3122.html

http://www.nature.com/nbt/journal/v33/n3/pdf/nbt.3172.pdf

http://www.nature.com/nprot/journal/v11/n9/pdf/nprot.2016.095.pdf

edgeR:

http://bioinformatics.oxfordjournals.org/content/26/1/139.full.pdf+html

DEseq2:

http://genomebiology.com/content/pdf/s13059-014-0550-8.pdf

Tutorials:

Tophat:

http://www.nature.com/nprot/journal/v7/n3/pdf/nprot.2012.016.pdf

EdgeR:

http://www.bioconductor.org/packages/release/bioc/vignettes/edgeR/inst/doc/edgeRUsersGuide.pdf

http://cgrlucb.wikispaces.com/file/view/edgeR_Tutorial.pdf

DEseq2:

http://www.bioconductor.org/packages/release/bioc/vignettes/DESeq2/inst/doc/DESeq2.pdf

http://dwheelerau.com/2014/02/17/how-to-use-deseq2-to-analyse-rnaseq-data/

After the next R lecture:

Exercises:

Alignment, summarization and normalization of RNA-seq data