מכון ויצמן למדע
WEIZMANN INSTITUTE OF SCIENCE

RNA-Seq Analysis Introduction

Dena Leshkowitz,

Course: Introduction to Deep-Sequencing Data Analysis 2015

Bioinformatics Unit, WIS

Main Topics

- Introduction
- Experimental design issues
- Analysing RNA-Seq data
  - RNA-Seq pipeline: Tophat-Cufflinks-Cuffdiff
- Challenges

RNA-Seq Potential

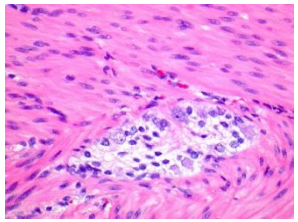
RNA-Seq: a revolutionary tool for transcriptomics

It allows to deeply sequence a transcriptome by random sampling

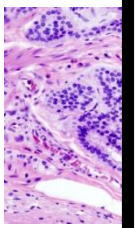
In theory RNA-Seq can be used to built a complete map of the transcriptome across all cell types, perturbations and states (Trapnell C. et al, Nature methods 6 469-477(2011))

RNA-Seq Workflow

Samples of inter

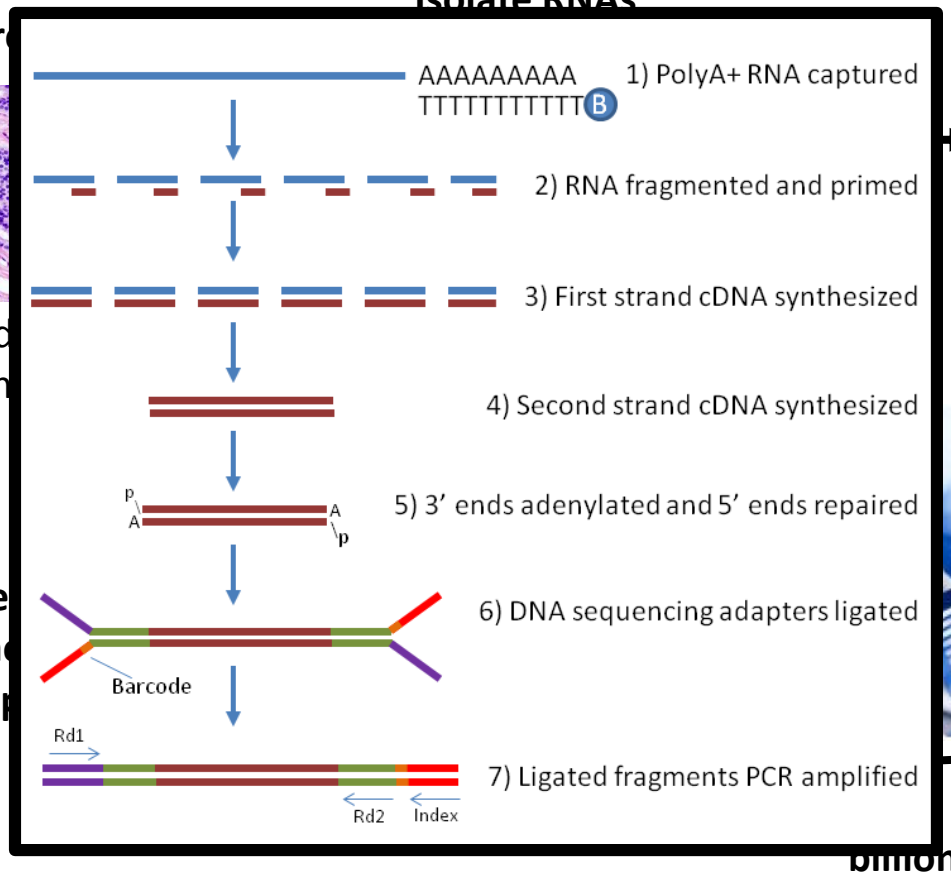


Condition 1
(normal colon)

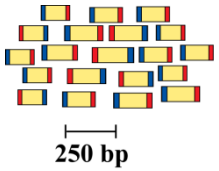


Cond
(colon

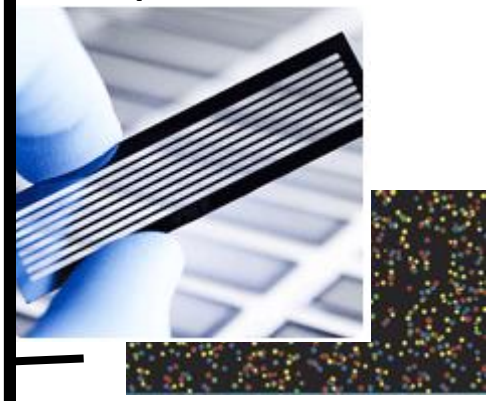
Isolate RNAs



Converted to amplified small cDNA fragments with linkers at the ends



Sequence ends



billions of fragments are sequenced

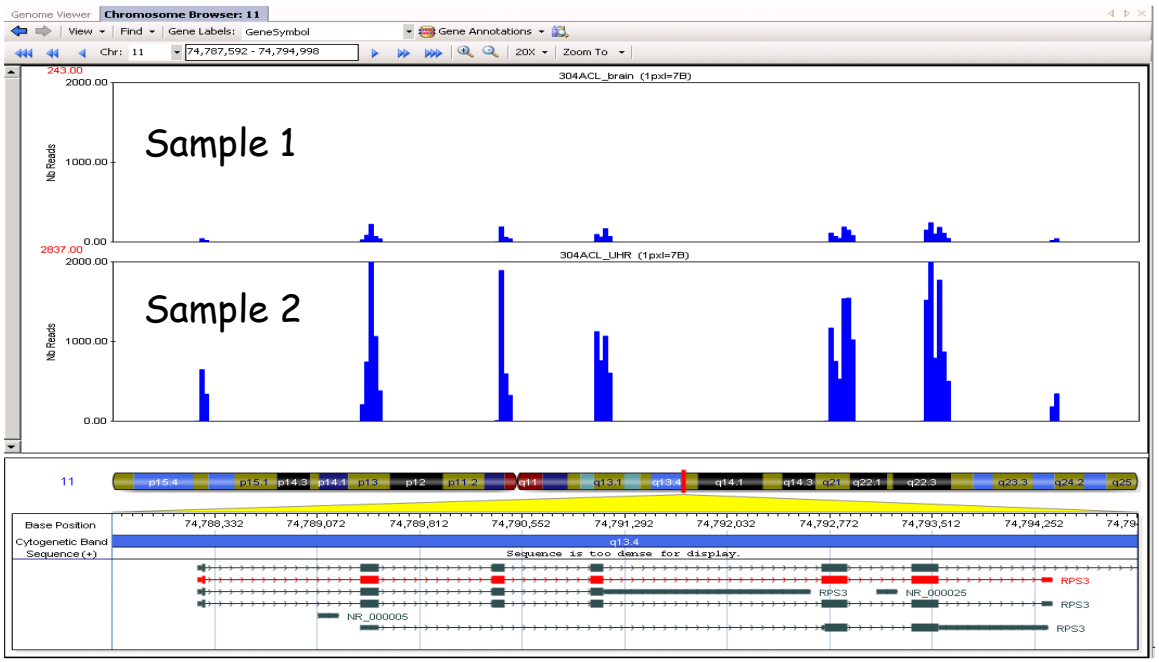
Assemble to genom transcri

Quantify and detect differential expression

Modified - www.med.nyu.edu/rcr/rcr/course/**RNA-seq**.pptx

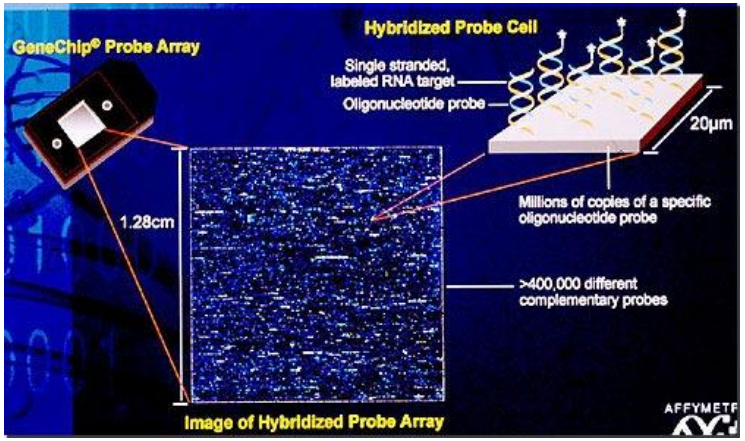
RNA-Seq Aims

- Estimate relative abundance of transcripts or genes (from their depth of coverage)
- Detect differential expressed genes and transcripts
- ✓ Identify novel genes or transcripts



Is the gene expressed differently expressed in the two samples?

High Throughput Genomics



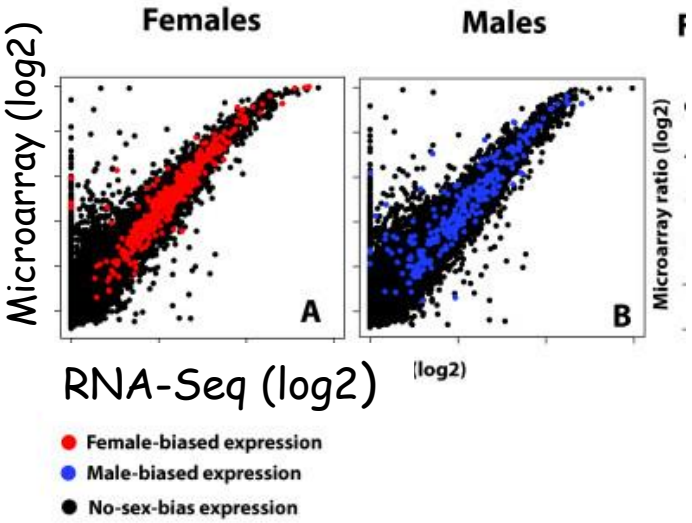
DNA Microarrays

Illumina HiSeq2500

Microarrays vs RNA-Seq

Malone et al. BMC Biol. 2011; 9: 34.

- Both high throughput methods can profile the genes with similar performance
- Microarrays suffer from compression (saturation) at the high end
- Low expression is problematic in both platforms



Microarray & RNA-Seq Pros and Cons

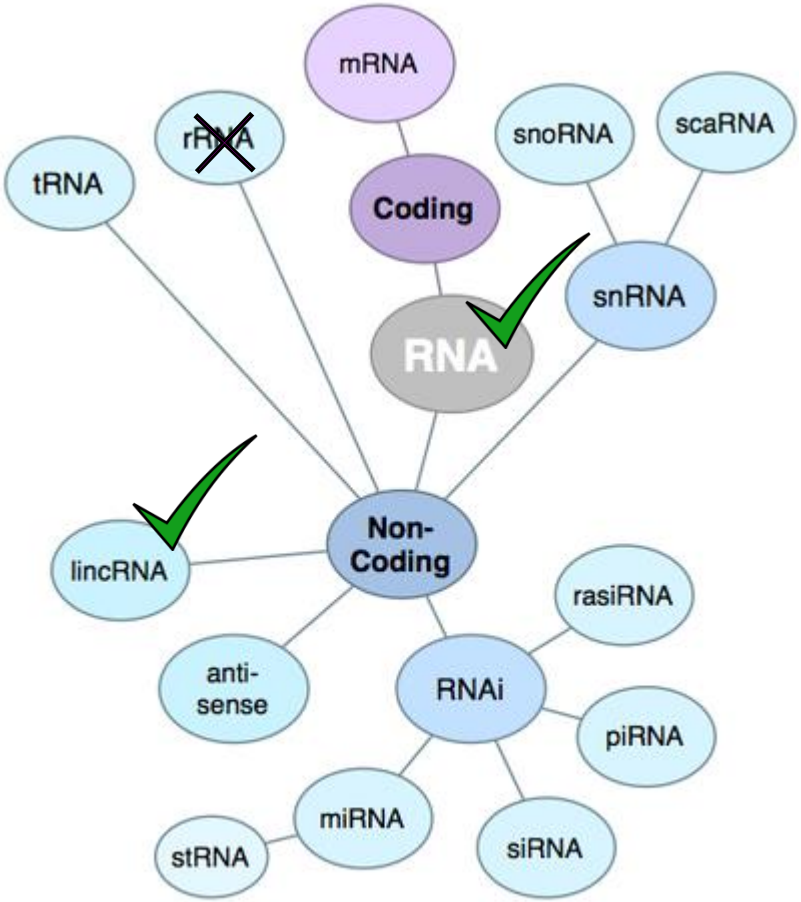
	Microarrays	RNA-Seq
Cost	$$	$$ (gene profiling) or $$$
Biases	Decade of research and solutions	Understanding is evolving
Data sizes	Mb -images	Gb- sequence data
Dynamic range	10^2	10^5
Transcript discovery , isoform identification & Transcript-chimeras	No	Yes
Genome required	Yes	No
Allele specific expression	No	Yes

Main Topics

- Introduction
- **Experimental design issues**
- Analysing RNA-Seq data
  - RNA-Seq pipeline: Tophat-Cufflinks-Cuffdiff
- Challenges

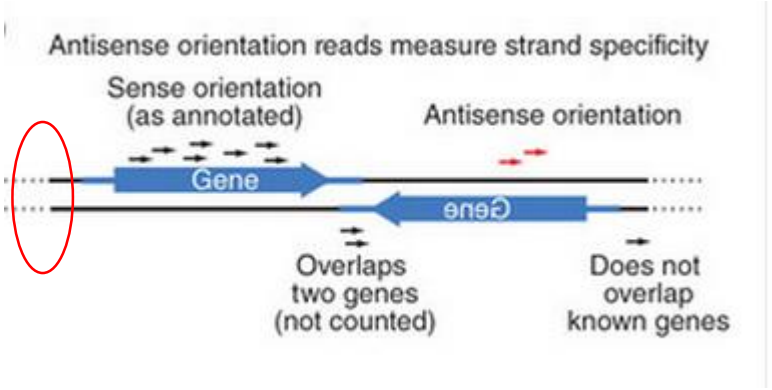
The Diverse RNA World

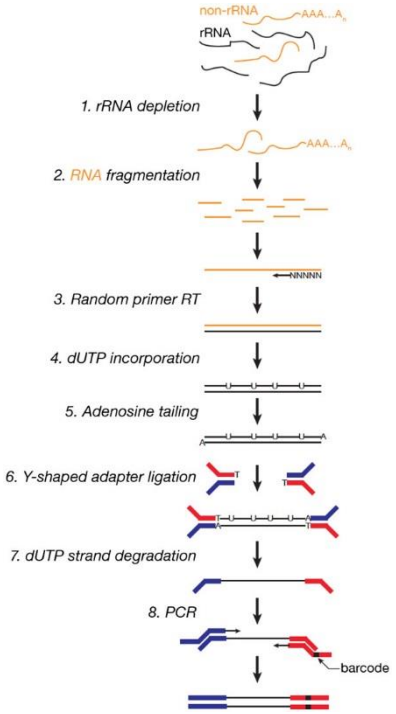
- Most abundant RNA is rRNA – 98%
- Illumina standard protocol enriches for mRNA by:
  - oligo(dT)-based affinity matrices
  - Size : hybridization-based rRNA depletion (Duplex--specific Nuclease (DSN))
  - Sequence: rRNA capture beads (Ribo-Zero)



Experiment Design
Strand specific protocol

- Why is strand information important?
- How is the stranded library made?





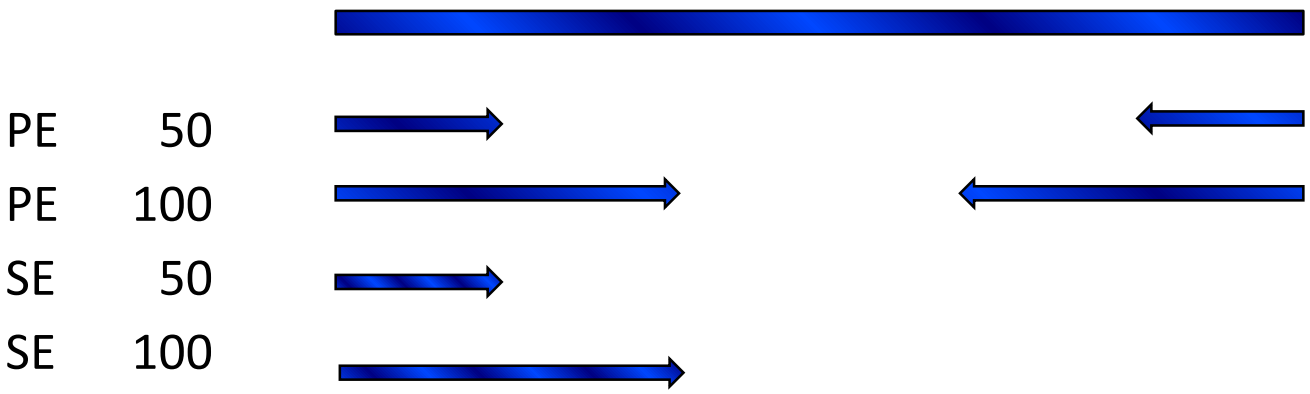
Nature Methods 7, 709–715 (2010)

Experiment Design
Sequencing Options

Sequencing options:

- **Length of sequence (50/100bases)**
- **Paired-end (PE) or single-end (SE)**

Both PE and length increase the sensitivity and specificity of the detection of the alternative splicing and chimeras



Sequencing Depth

How many sequences do I need per sample?

Should I divide the lane between samples (using multiplexing)?

- ENCODE consortium's *Standards, Guidelines and Best Practices for RNA-Seq*
  - Gene profiling with a mammalian genome - 30M pair-end reads of length > 30 bp
  - Novel and isoforms depth 100-200M, >= 2 x 76bp
- Wang et al. BMC Bioinformatics 2011

  10M (75 bp) reads could detect about 80% of annotated chicken genes
- http://eh.uc.edu/genomics/files/Illumina_Whitepaper_RNASeq_to_arrays_comparison.pdf

BIOINFORMATICS DISCOVERY NOTE

Vol. 30 no. 3 2014, pages 301–304
doi:10.1093/bioinformatics/btt688

Gene expression

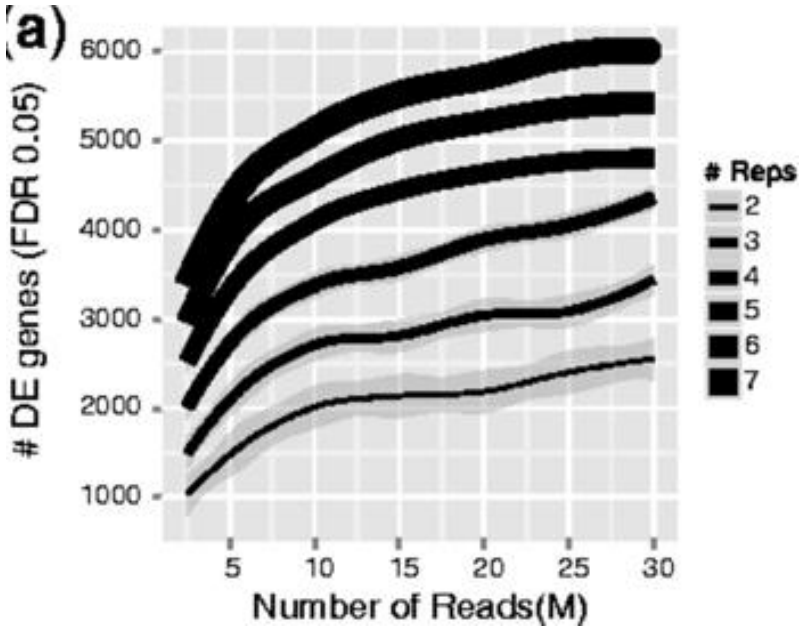
Advance Access publication December 6, 2013

RNA-seq differential expression studies: more sequence or more replication?

Yuwen Liu[1,2], Jie Zhou[1,3] and Kevin P. White[1,2,3,*]

[1]Institute of Genomics and Systems Biology, [2]Committee on Development, Regeneration, and Stem Cell Biology and [3]Department of Human Genetics, University of Chicago, Chicago, IL 60637, USA

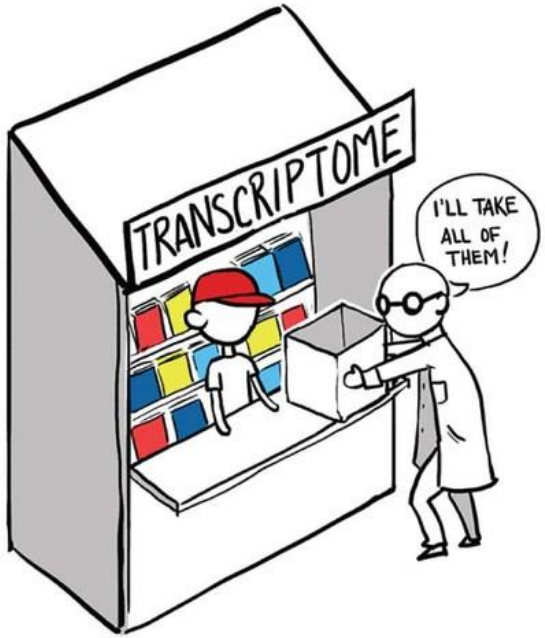
Associate Editor: Janet Kelso

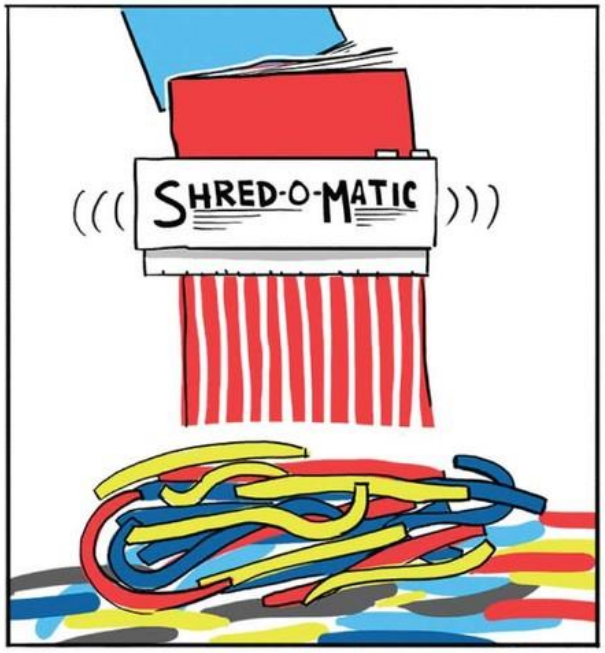


Experimental Design

- Library protocol
- Sequencing depth (# reads/fragments per sample)
- Sequencing options (length, SE, PE)
- Assessing biological variation requires biological replicates
  - Duplicates (2X2) are a minimum, yet more recommended

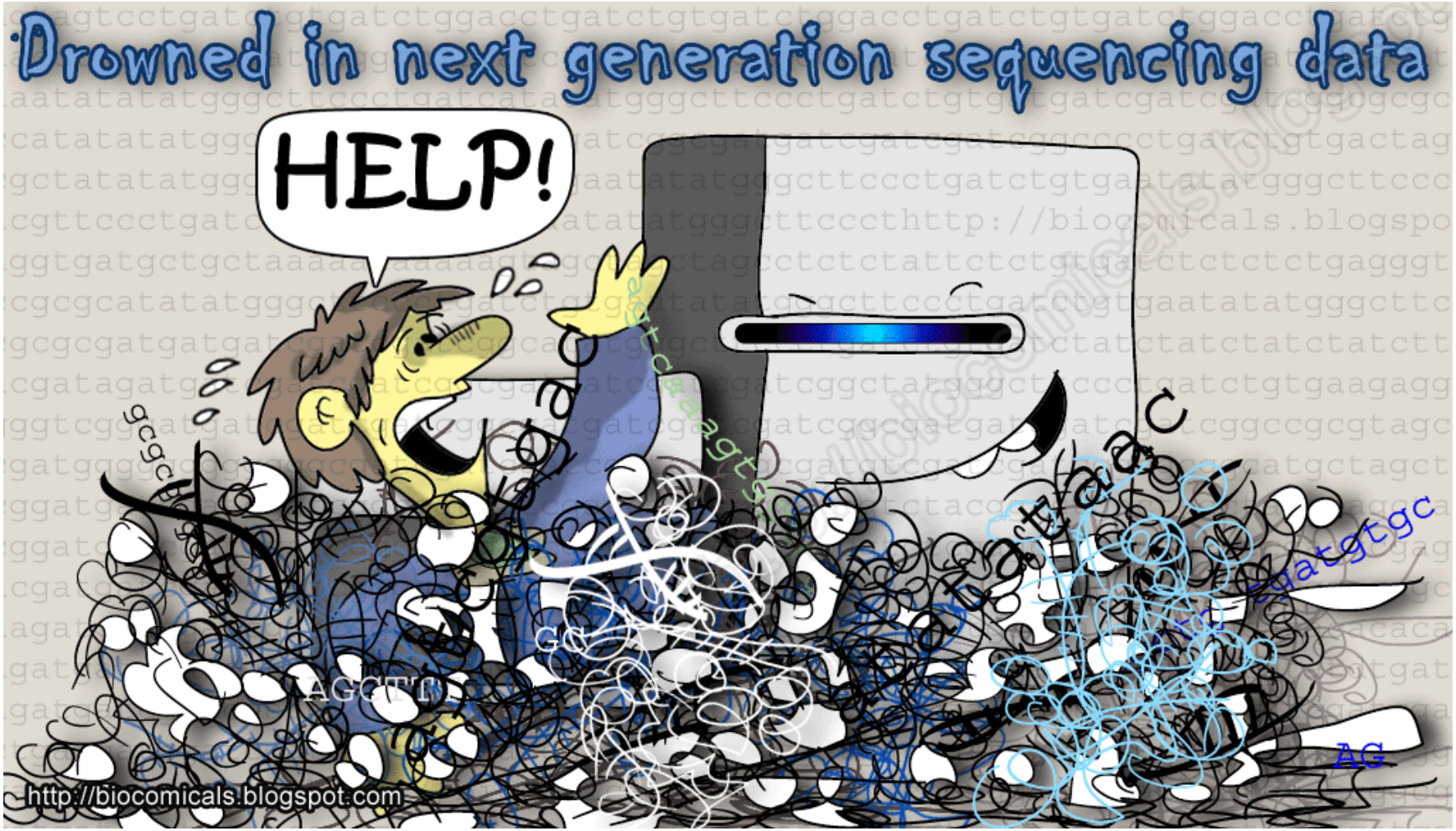
    (pooling, avoid batch effect)
- Consult with the person which will analyse the data before performing the experiment – Kick-off meeting





RNA-Seq is a straightforward process: you isolate RNA, sequence it with a high-throughput sequencer, and put it all back together. What is the problem?

Abigail Yu Nature Methods 10, 1165–1166 (2013)



HELP !!!!

I just got sequence data from two lanes of HiSeq run each with 200 million pieces…

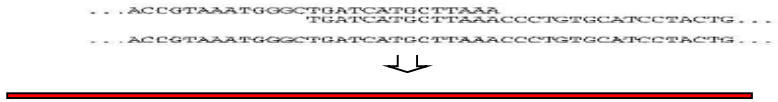
Main Topics

- Introduction
- Experimental design issues
- **Analysing RNA-Seq data**
  - **RNA-Seq pipeline: Tophat-Cufflinks-Cuffdiff**
- Challenges

Pre-processing

- Current run of Illumina produces 150M clusters-fragments per lane
- Do we use all the data?
- Recommendation is to use the high quality sequence data (more important in de novo assembly):
  - Filter low quality reads
  - Check the amount of read duplication (too much PCR amplification)
  - Trim sequences if 3 end is of low quality
  - Remove adaptor & Remove spiked-in sequences

From Sequences to Transcriptome Quantification



No Genome or Transcriptome

De novo assembly

RNA-Seq

Mapping

Genome

Transcriptome

Computational Steps

- **Assembly (?)**

  A necessary step in defining novel transcripts. Either with or without genome information.

- **Quantification**

  Given RNA-Seq reads and transcripts, estimate their relative abundance

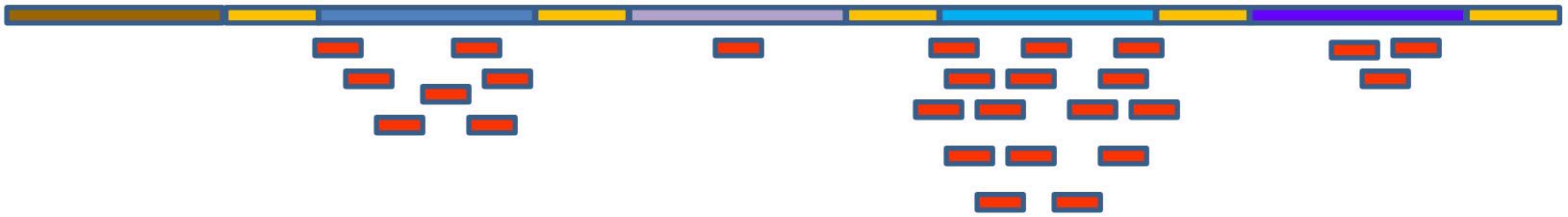
- **Differential Expression**

  Determine for a given gene or transcript, if the observed difference in read counts is significant, that is, greater than would be expected from just natural random variation

Basic Quantification Step

- Align (map) reads against a set of reference transcript sequences

Annotated genome



- Count the number of reads aligned to each transcript
- Convert read counts into relative expression levels

Normalization

- Between samples
  - Need to account for the different number of reads/fragments sequenced for each sample
- Between genes/transcripts
  - Need to account for genes having different length,since sequencing is done on short fragments, longer genes have a higher chance of being sequenced

Expression Values

Fragments (Reads) Per Kilobase of exon per Million mapped fragments

Nat Methods. 2008, Mapping and quantifying mammalian transcriptomes by RNA-Seq. Mortazavi A et al.

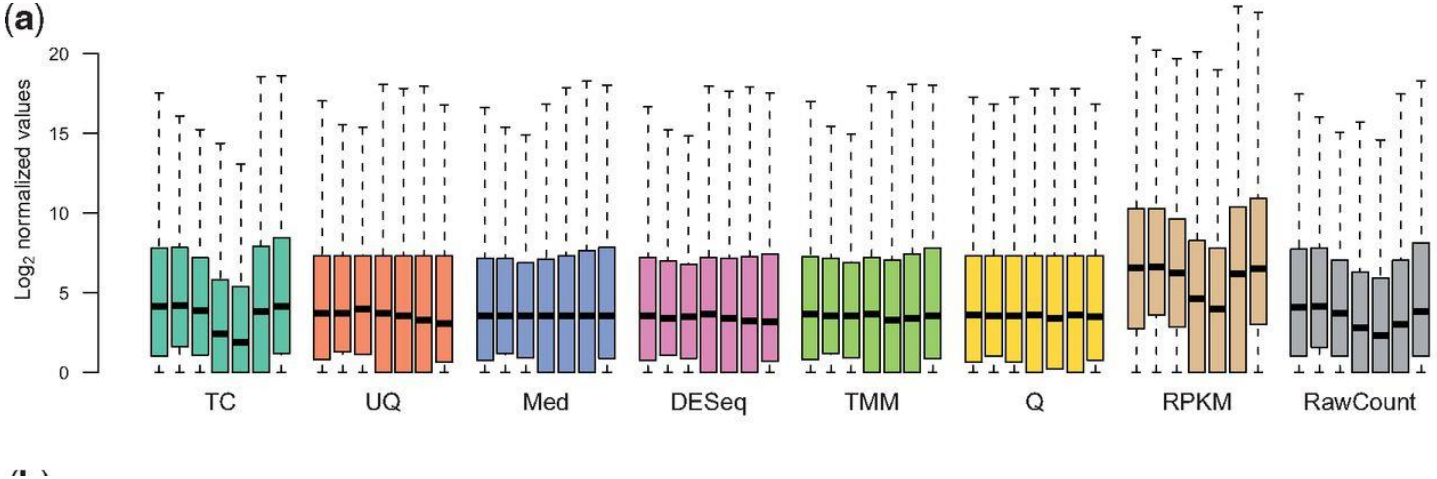
$$FPKMi = 10^6 \times 10^3 \times \frac{Ci}{NLi}$$

C= the number of fragments mapped onto the gene's exons

N= total number of (mapped) fragments in the experiment

L= the length of the transcript (sum of exons)

Comparison of Normalization Methods for Real Data



Dillies M et al. Brief Bioinform 2012;bib.bbs046

Note: There are other means of normalization/scaling which perform better (such as DESeq)

FPKM AND TPM

Theory Biosci. (2012) 131:281–285
DOI 10.1007/s12064-012-0162-3

SHORT COMMUNICATION

Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples

Günter P. Wagner · Koryu Kin · Vincent J. Lynch

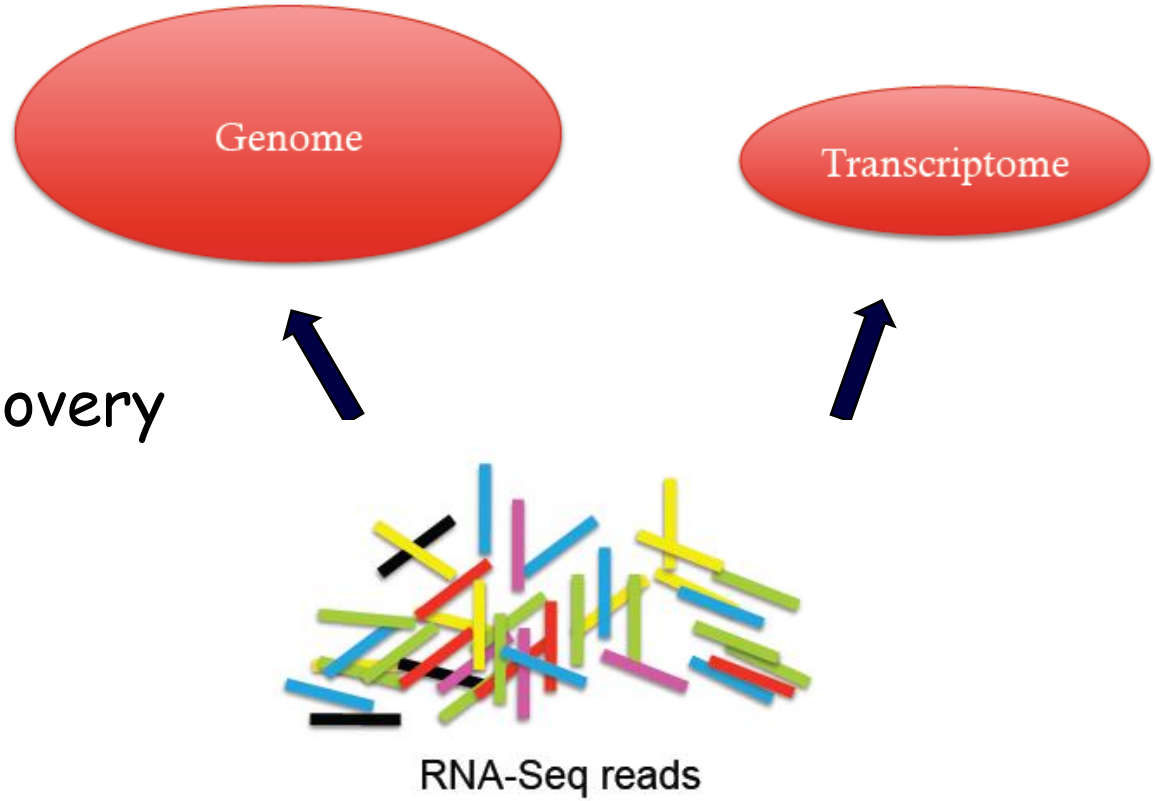
Table 1 Comparison of average TPM and RPKM among different cells types and samples (see supplementary material and Wang et al. 2011)

Species	Tissue/cell type	Replicate	AvTPM	AvRPKM	Scaling *f*
Human	Differentiated decidual cells	1	46.518	15.94	2.92
		2	46.518	16.13	2.83
Human	Un-differentiated dec. cells	1	46.518	15.27	3.05
		2	46.518	15.22	3.06
Human	Myofibroblast cells	1	46.518	17.66	2.62
		2	46.518	17.65	2.62
Human	Chondrocyte cells	1	46.518	16.57	2.81
		2	46.518	16.57	2.81
Human	Myometrial cells	1	46.518	17.77	2.62
		2	46.518	17.79	2.61
Chicken	Forelimb digit 1 stage 28–29	–	65.527	28.35	2.31
Chicken	Forelimb digit 1 stage 31	–	65.527	28.56	2.29

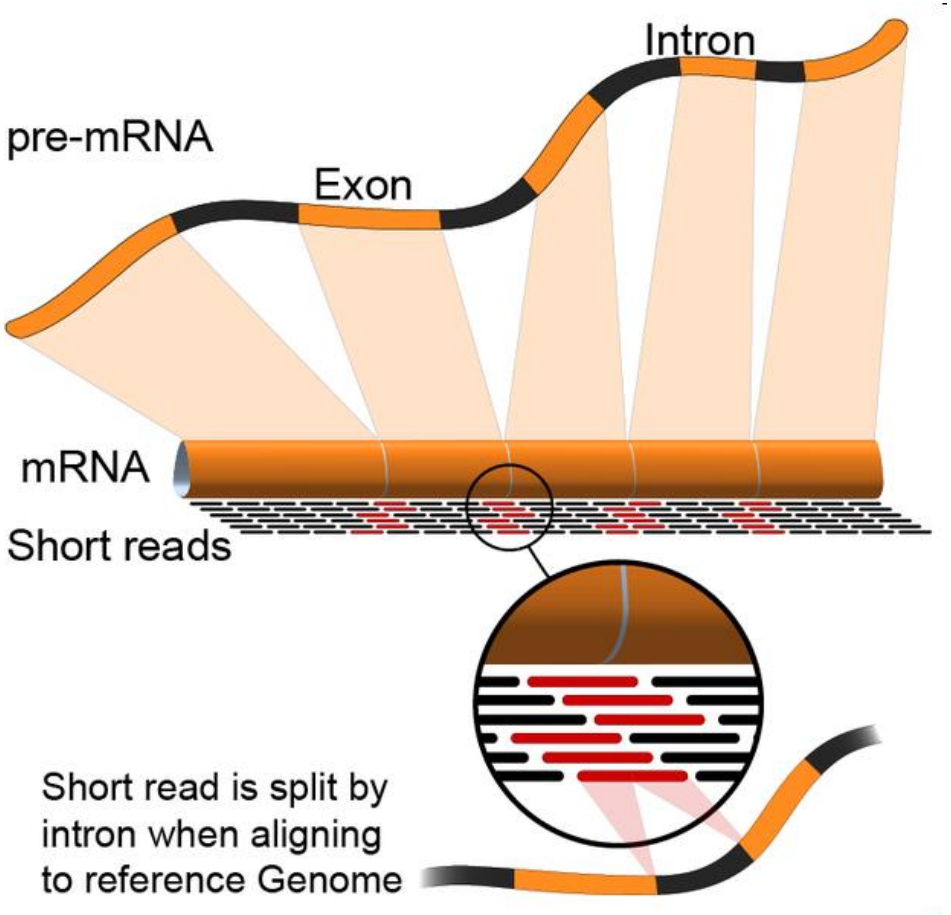
TPM = (**FPKM** for gene / (**sum** of all **FPKM** for all genes)) * 10^6

Mapping Short RNA-Seq Reads

Do I align the reads to the genome or to the transcriptome?



Mapping to Genome
How to detect Spliced Reads?



http://en.wikipedia.org/wiki/RNA-Seq

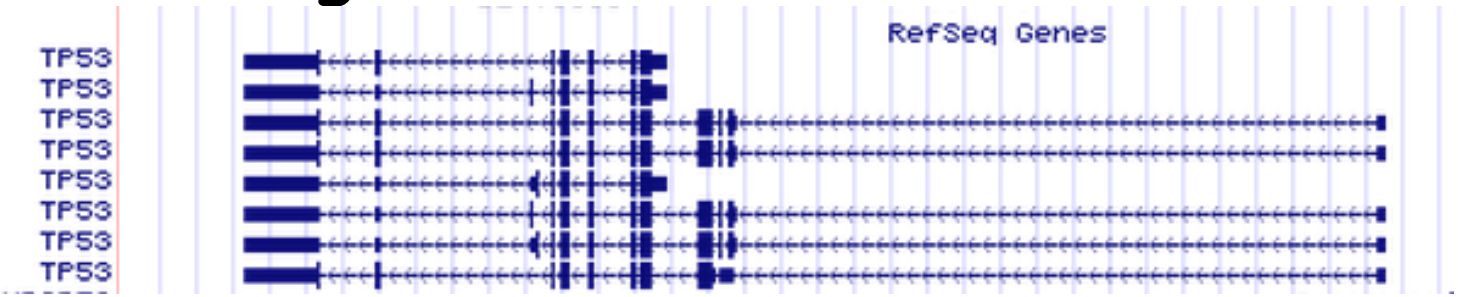
Advantages of the Methods

- Alignment to genome
  - Allows detection of new transcripts and isoforms
- Alignment to transcriptome
  - Computationally inexpensive
  - Does not require a genome sequence
  - Spliced (exon junction) reads map correctly
  - No mis-mapping to a pseudogene
- Assembly
  - The only alternative when there is no genome or good transcriptome
  - Allows detection of chimera transcripts and resolution of 'breakpoints'
  - Problem: transcripts with low coverage, transcript variants

Align to Transcriptome Quantification Problem

- Count the number of sequences that map uniquely to the genes or transcripts
- Problem - Results in false estimates of alternatively spliced transcripts which share exons and in gene families

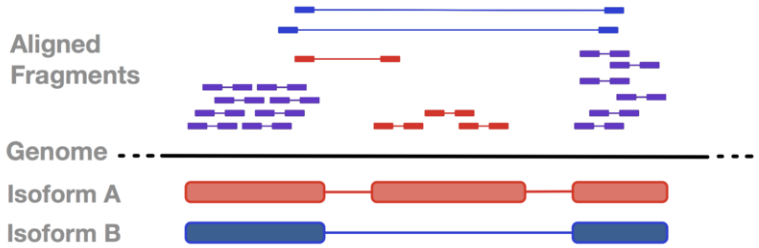
Human p53 locus



- Computational challenge- use reads that map ambiguously between isoforms and genes (EM algorithm)

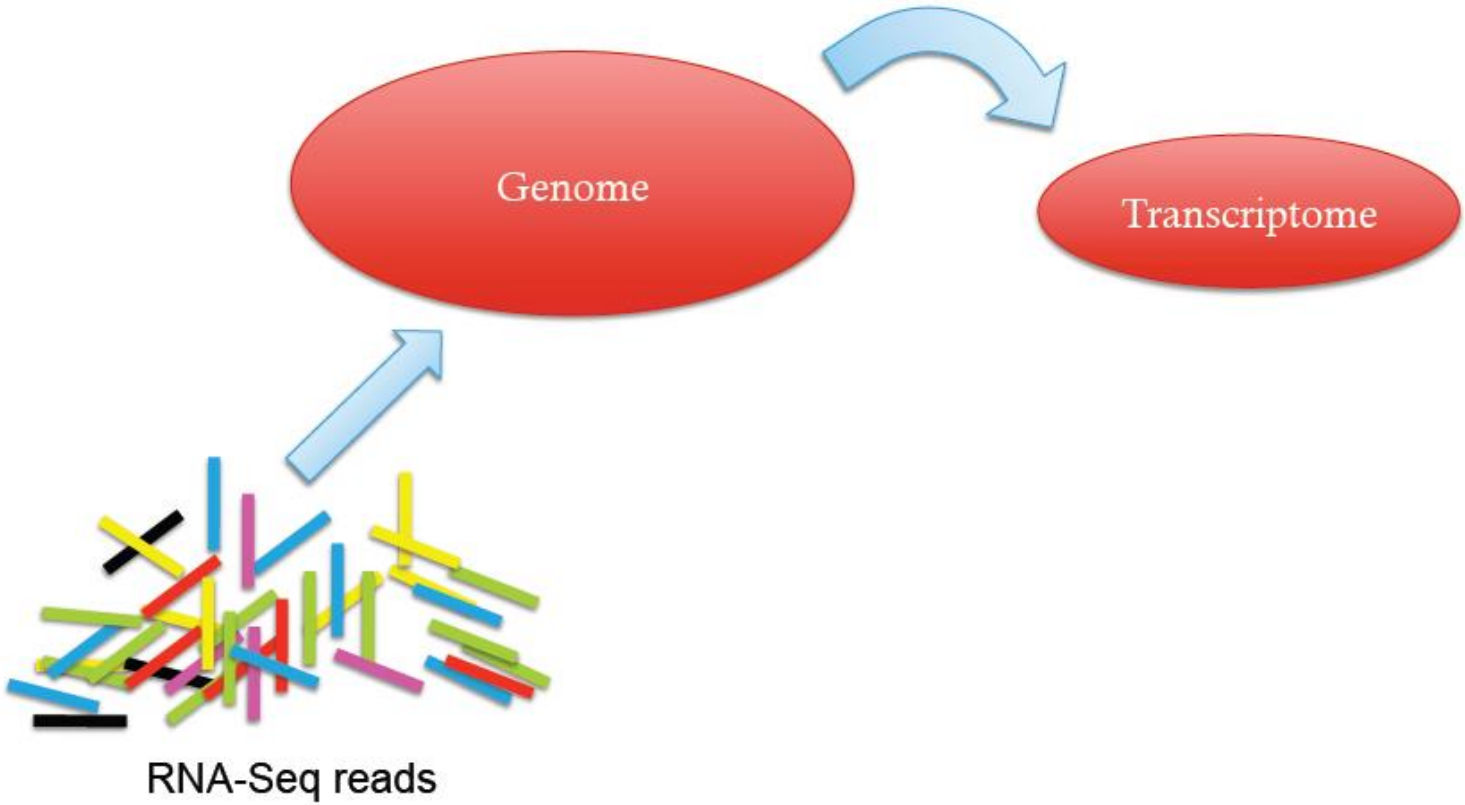
The Estimation Problem

How to distribute the purple reads among the two isoform transcripts?

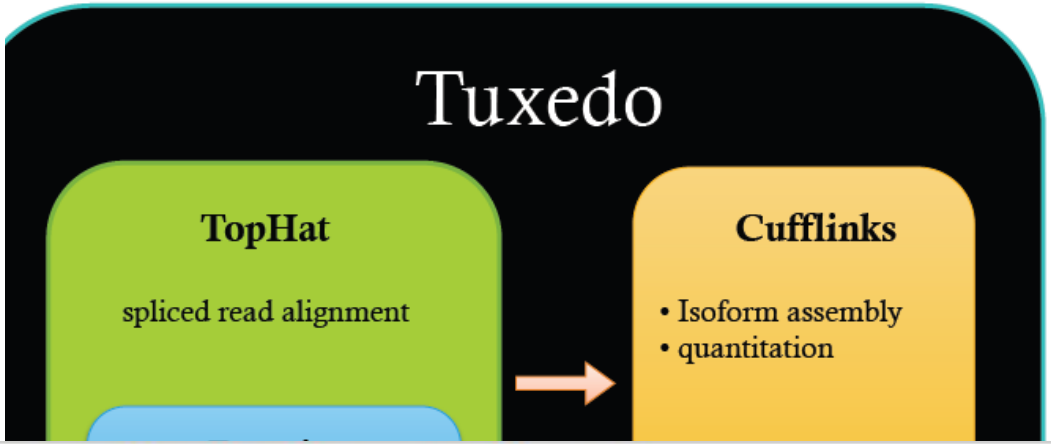


RNA-Seq mapping with TopHat

Goal: **identify** all transcripts and estimate relative amounts from RNA-Seq data



The Tuxedo Tools



Bioinformatics. 2009 May 1; 25(9): 1105–1111.
PMCID: PMC2672628
Published online 2009 March 16. doi: 10.1093/bioinformatics/btp120

TopHat: discovering splice junctions with RNA-Seq

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

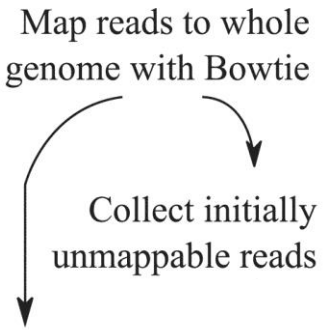
[1]Center for Bioinformatics and Computational Biology, University of Maryland, College Park, MD 20742 and [2]Department of Mathematics, University of California, Berkeley, CA 94720, USA
*To whom correspondence should be addressed.
Associate Editor: Ivo Hofacker

The Challenge –Identifying Novel Junctions

- Reads are short and contain errors
- Rarely transcribed genes have few reads spanning the junctions
- We are interested in discovering novel junctions i.e. we are not relying on annotation of known genes (?)
- Need to perform the task in a timely manner

Tophat: Exon first two step approach

- Mapping to the genome is done with Bowtie
- Extracting unmapped reads (not including low complexity)

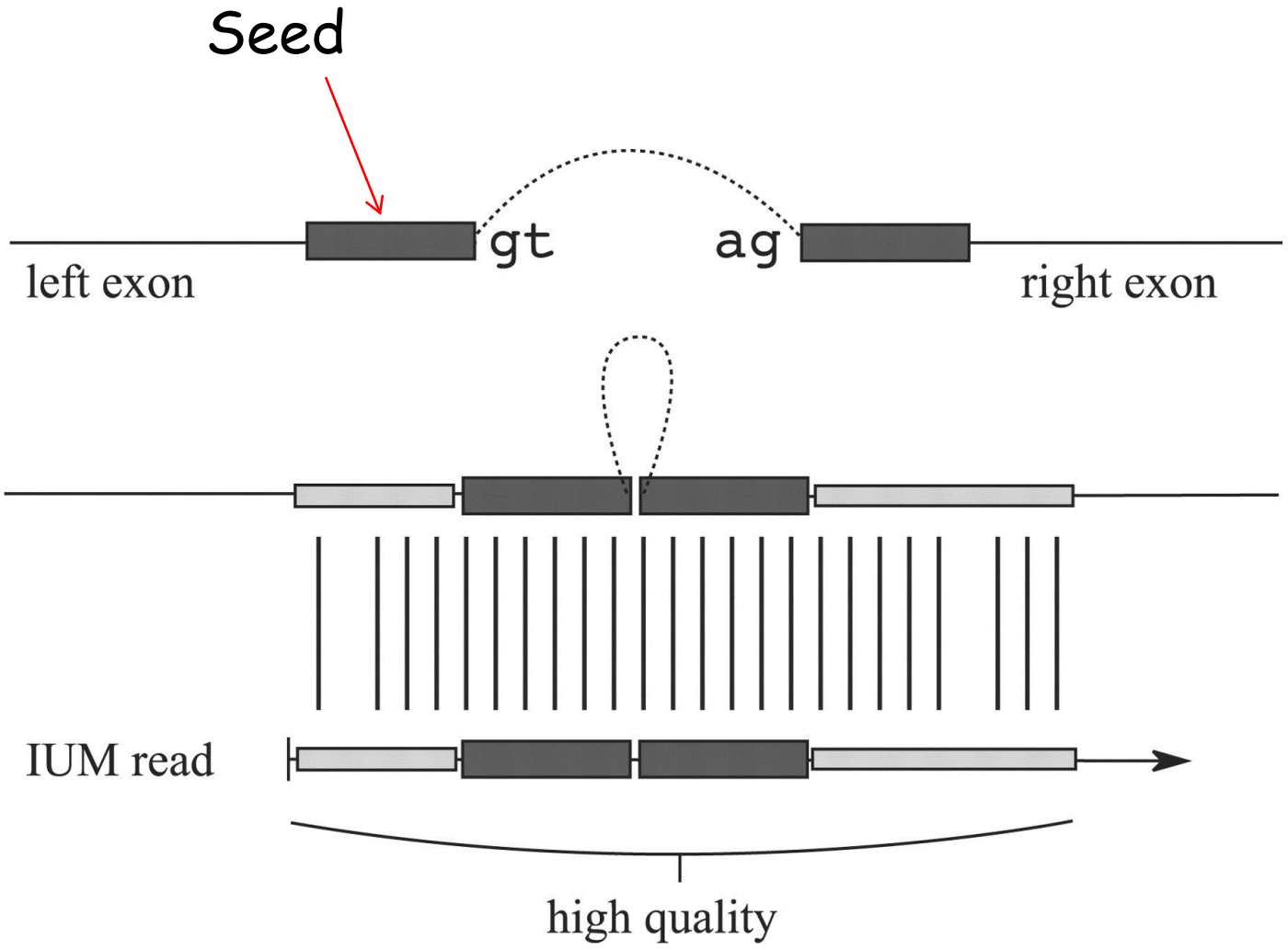


Identifying the transcriptome

identify candidate exons via genomic mapping

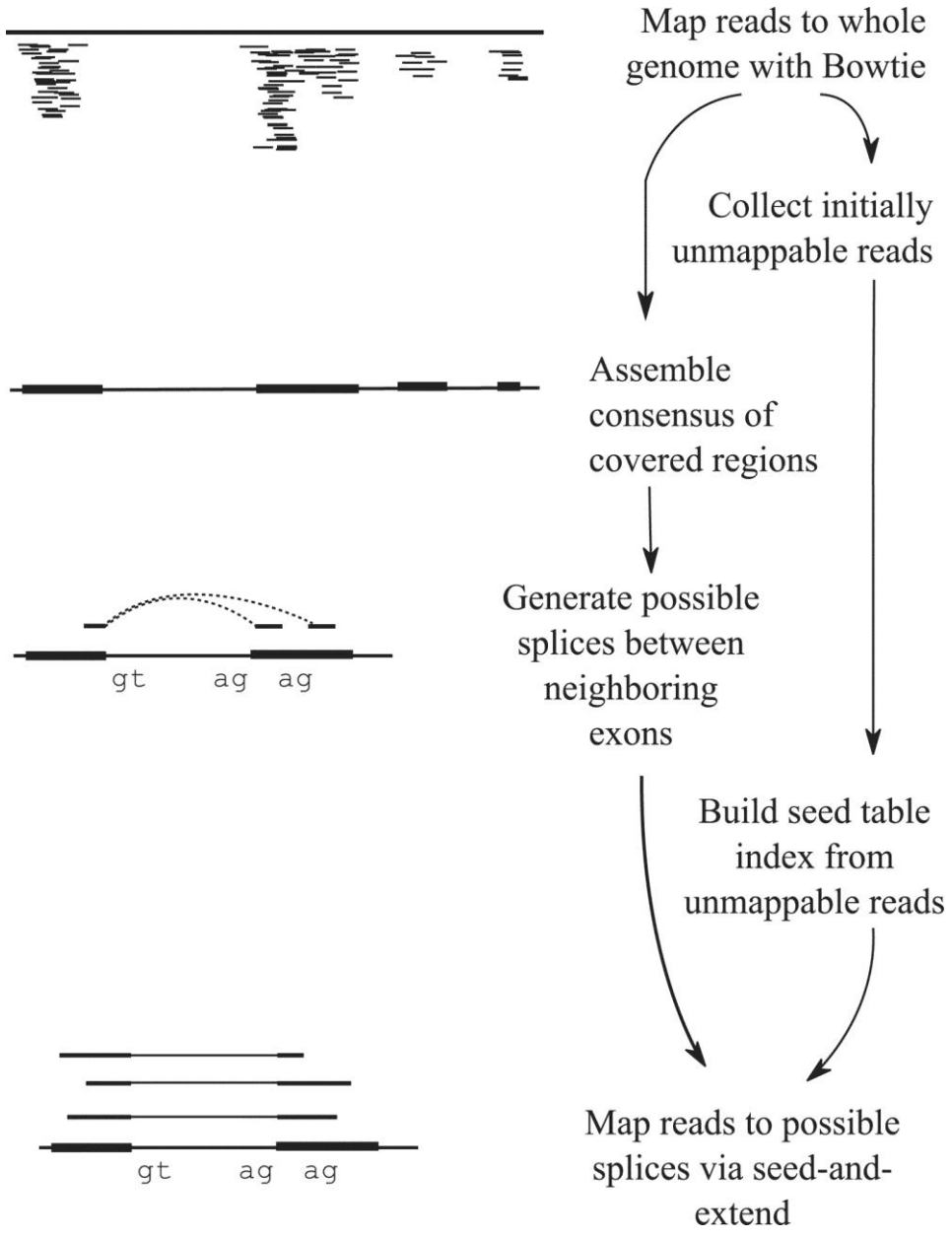
Generate possible pairings of exons

Align “unmappable” reads to possible junctions



Trapnell, C. et al. Bioinformatics 2009 25:1105-1111; doi:10.1093/bioinformatics/btp120

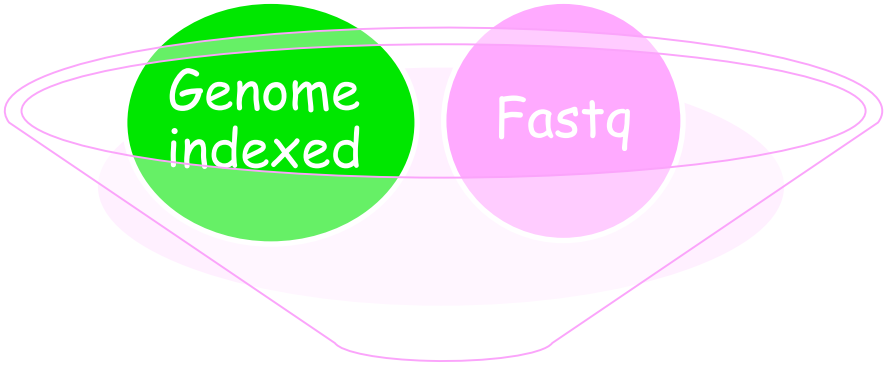
Bioinformatics



Trapnell, C. et al. Bioinformatics 2009 25:1105-1111; doi:10.1093/bioinformatics/btp120

Bioinformatics

Tophat Outputs



Tophat

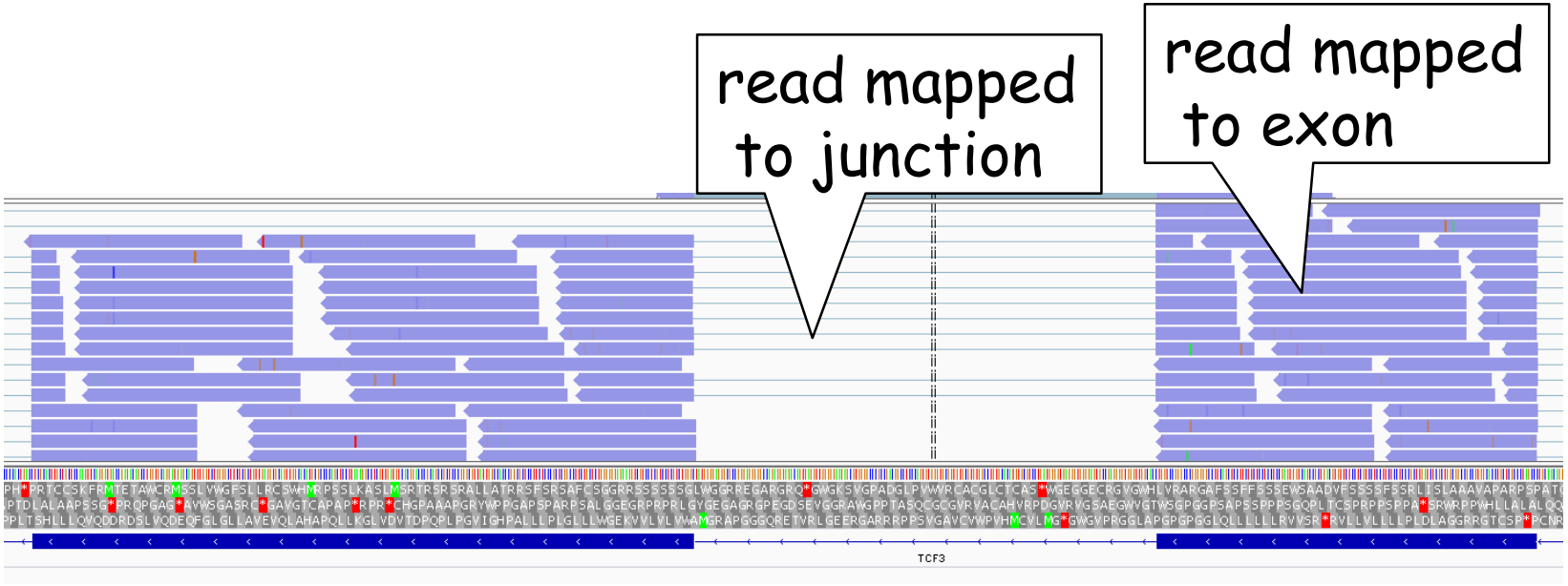
Accepted_hits.bam + Insertion.bed Deletion.bed + Junctions.bed

Genome Browser

Junction.bed

5	chr1	94935734	94936354	JUNC00022394	457	-	11,33	0,587
6	chr1	94962730	94963662	JUNC00022414	456	+	32,34	0,898
7	chr10	28642081	28642574	JUNC00005781	38	-	18,35	0,458
8	chr10	76067101	76067623	JUNC00006282	74	+	30,34	0,488
9	chr10	118650871	118652271	JUNC00007170	20	-	16,22	0,1378
10	chr10	126684651	126685114	JUNC00007305	87	-	26,35	0,428
11	chr10	127986002	127986178	JUNC00007509	96	-	25,35	0,141
12	chr10	128006919	128007139	JUNC00007511	63	+	27,34	0,186
13	chr11	23326211	23326669	JUNC00002682	31	+	34,35	0,423
14	chr11	59637921	59639825	JUNC00003176	20	-	35,28	0,1876
15	chr11	69161988	69162244	JUNC00003397	49	-	35,33	0,223
16	chr11	69652041	69652451	JUNC00003476	24	+	34,30	0,380
17	chr11	97048737	97048877	JUNC00004353	22	+	18,24	0,116
18	chr12	113031876	113032578	JUNC00009692	45	+	34,31	0,671
19	chr14	53158445	53159099	JUNC00012157	20	-	31,29	0,625
20	chr14	53160268	53160487	JUNC00012158	30	-	32,32	0,187
21	chr15	76454881	76455044	JUNC00010523	21	-	30,27	0,136
22	chr15	89289222	89289365	JUNC00011107	27	-	16,29	0,114

Visualization of Tophat outputs in a Genome Browser (IGV)



The Tuxedo Tools

NATURE BIOTECHNOLOGY | RESEARCH | LETTER

Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation

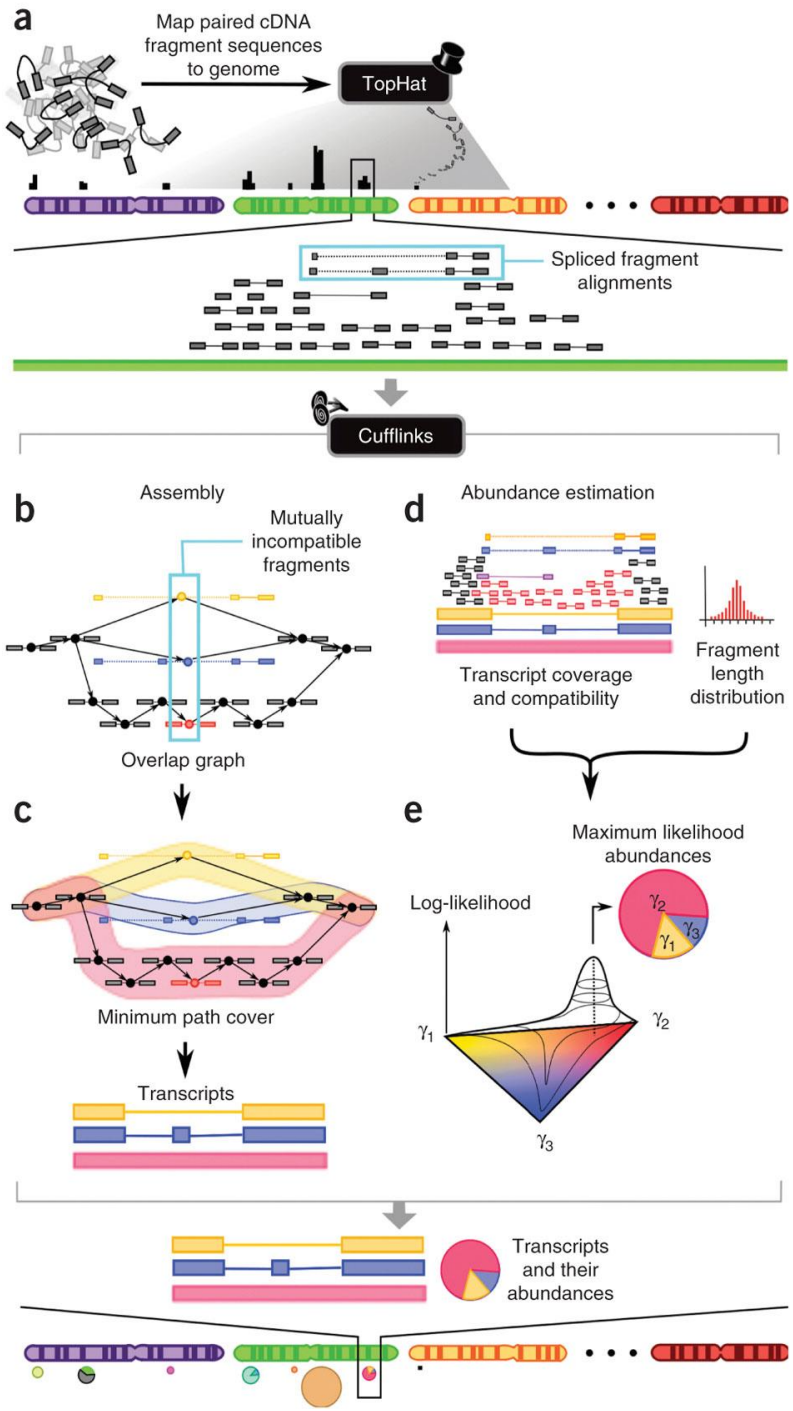
Cole Trapnell, Brian A Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J van Baren, Steven L Salzberg, Barbara J Wold & Lior Pachter

Affiliations | **Contributions** | **Corresponding author**

Nature Biotechnology **28**, 511–515 (2010) | doi:10.1038/nbt.1621
Received 02 February 2010 | Accepted 22 March 2010 | Published online 02 May 2010

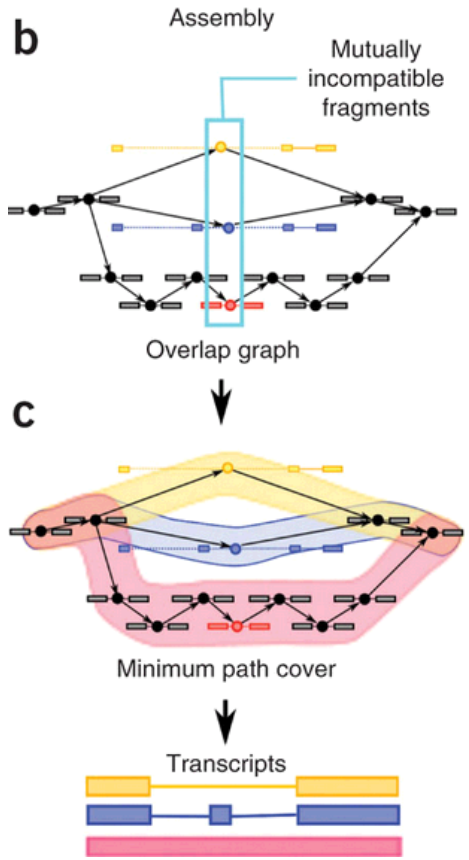
Cufflinks Detects Novel and Known Transcripts

- "To test Cufflinks, we sequenced and analyzed >430 million paired 75-bp RNA-Seq reads from a mouse myoblast cell line over a differentiation time series. We detected 13,692 known transcripts and 3,724 previously unannotated ones, 62% of which are supported by independent expression data or by homologous genes in other species."



Nature Biotechnology 28, 511-515 (2010)

Overview of Cufflinks

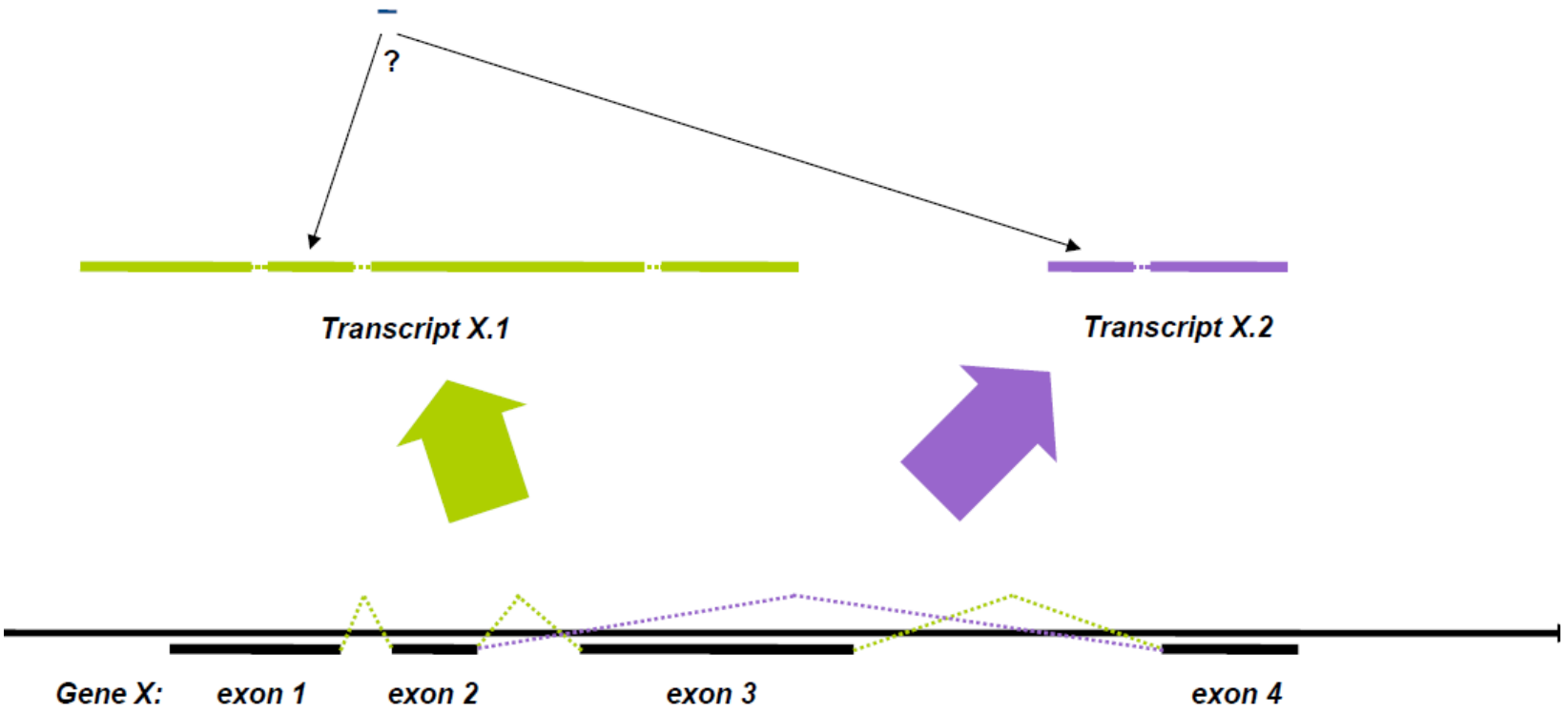


- Identify pairs of 'incompatible' fragments that must have originated from distinct spliced mRNA isoforms
- Fragments are connected in an 'overlap graph' when they are compatible and their alignments overlap in the genome
- Find minimum number of transcripts needed to 'explain' all the fragments

Transcripts and their abundances

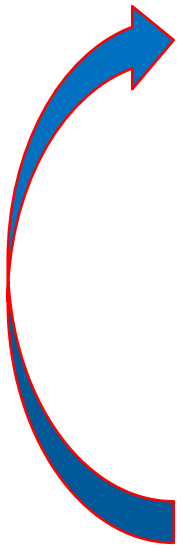
Trapnell et al. Nature Biotechnology 28, 511-515 (2010)

Align to Transcriptome



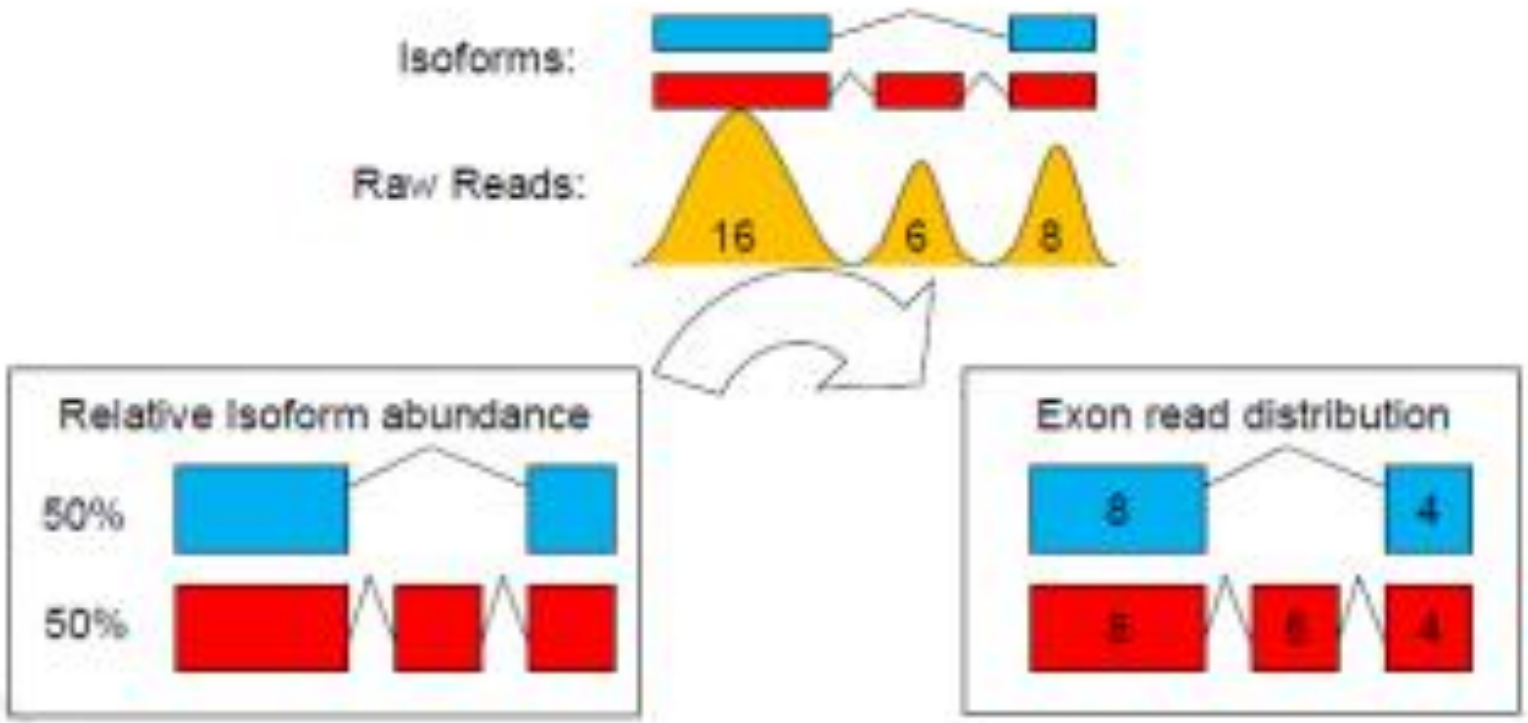
Taken from Joe Fass presentation

Isoform Expression Quantification Expectation Maximization Algorithm

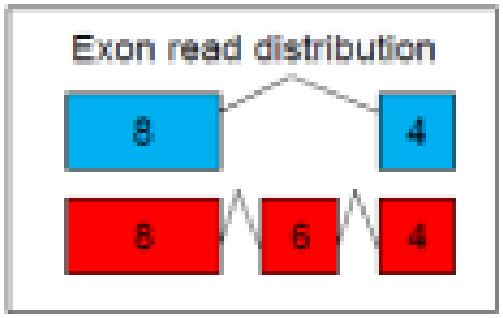


- Step 1- Assume isoforms are equally abundant
- Step 2 - Distribute the reads to the isoforms based on the abundance
- Step 3 - Recalculate the isoforms abundance based on the reads counts and isoforms length
- Step 4- If abundance has changed go back to step 2 otherwise stop

1st step E/M algorithm



Calculating Abundance after 1st EM Cycle



Total reads	Length
12	300
18	400

Exon length (bp) 200 100 100

The red transcript abundance after the first cycle:

$$p_{red} = \frac{counts_{red} / length_{red}}{counts_{red} / length_{red} + counts_{blue} / length_{blue}}$$

p_{red} = 18/400 / (12/300 + 18/400) = 0.53
p_{blue} = 12/300 / (12/300 + 18/400) = 0.47

EM Calculation: 100 Iterations

	starting relative proportion (p)	read counts	New proportion after iteration (p)		starting relative proportion (p)	read counts	New proportion after iteration (p)		Iteration #
Blue	0.5	12	0.470588	Red	0.5	18	0.529412		1
		11.29412	0.445993			18.70588	0.554007		2
		10.70383	0.425161			19.29617	0.574839		3
		10.20386	0.407324			19.79614	0.592676		4
		9.775778	0.39191			20.22422	0.60809		5
		9.405837	0.378482			20.59416	0.621518		6
		9.083574	0.366704			20.91643	0.633296		7
		8.800885	0.356308			21.19911	0.643692		8
		8.551391	0.347084			21.44861	0.652916		9
		8.330004	0.338859			21.67	0.661141		10
		6.00743	0.25029			23.99257	0.74971		94
		6.006965	0.250272			23.99303	0.749728		95
		6.006529	0.250255			23.99347	0.749745		96
		6.006121	0.250239			23.99388	0.749761		97
		6.005738	0.250224			23.99426	0.749776		98
		6.005379	0.25021			23.99462	0.74979		99
		6.005042	**0.2502**			23.99496	**0.7498**		100

Blue 25%

Red 75%

The Benefit of Longer and PE Reads

- Reads mapping to junctions
  - With longer reads we will have more of these reads

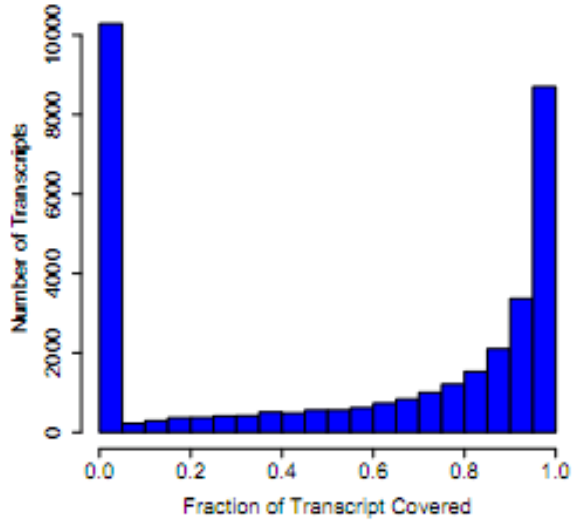
- Paired end reads

  Knowing both ends of a fragment it is easier to determine from which isoform this fragment originated

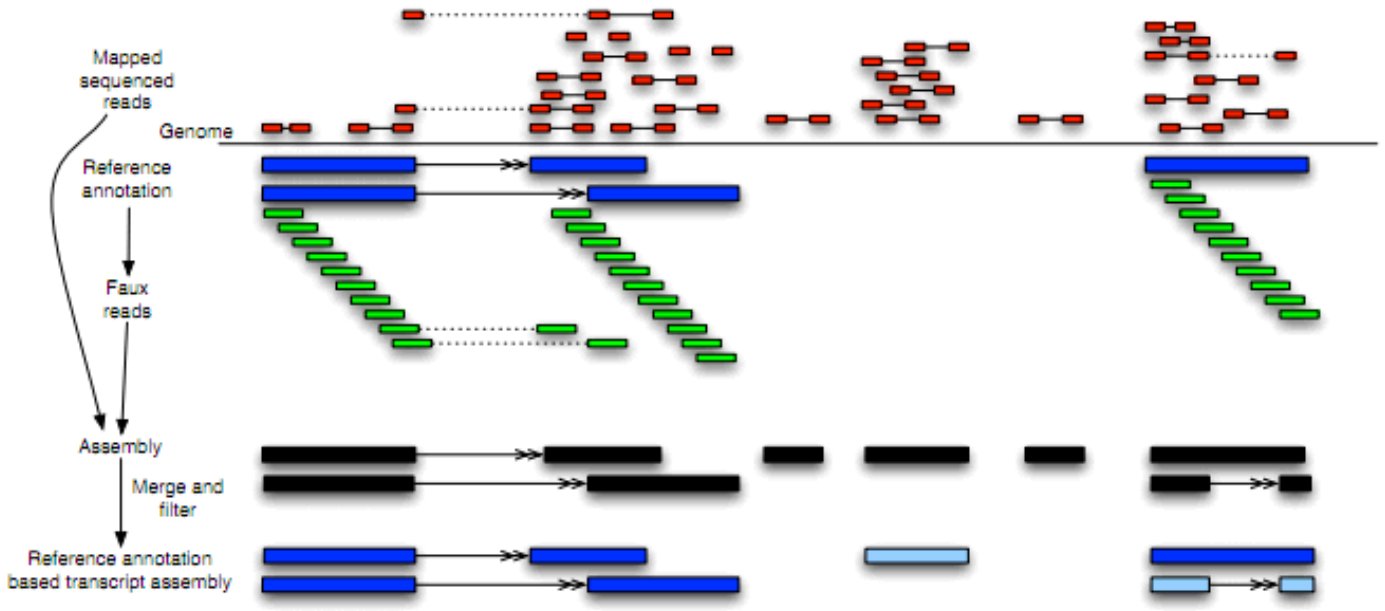
Cufflinks -RABT

- Transcripts that are expressed in low level are represented by few reads and therefore only partially covered (64%).
- That means that naive assembly methods will fail to construct the majority of the transcripts

Roberts et al. Bioinformatics. 2011 Sep 1;27(17):2325-9.



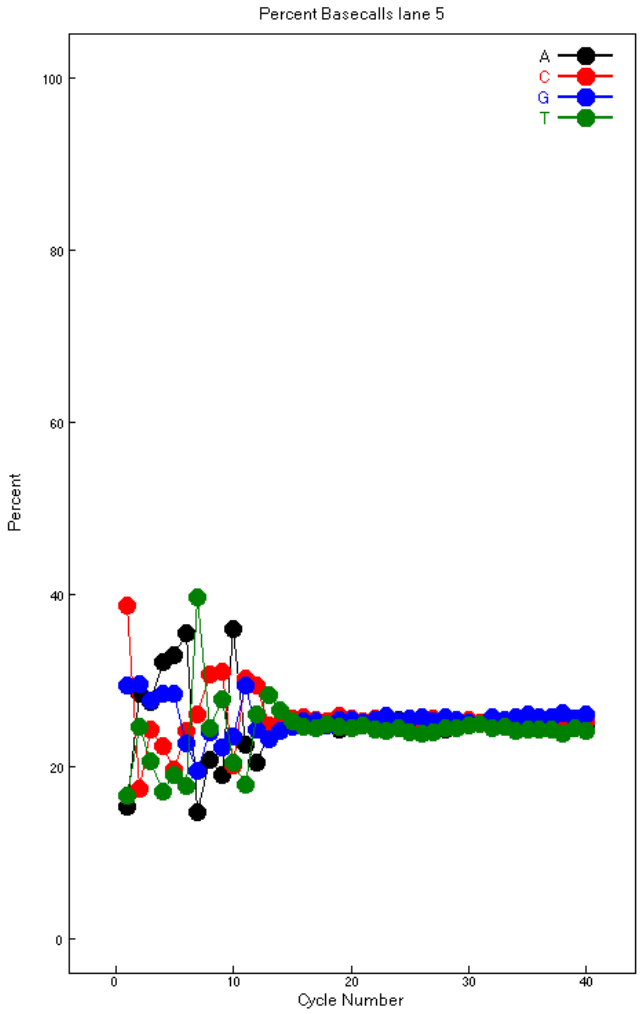
RABT: Reference Annotation Based Assembler (-g)



Faux reads tiling the transcripts are added to the real reads by cufflinks algorithm in the process of assembly

Cufflinks Bias Correction

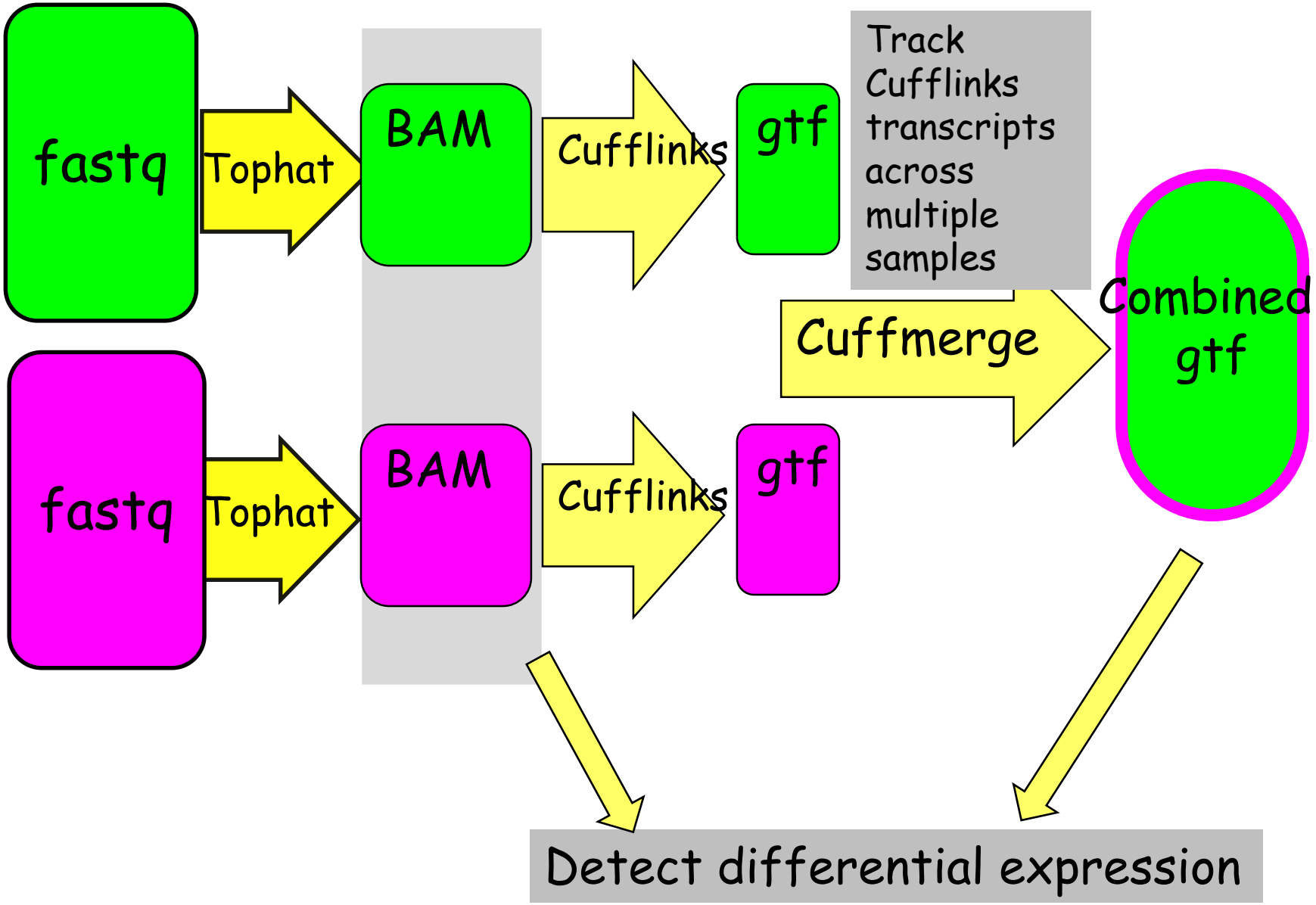
The random priming in the process of cDNA creation causes a positional preferred location for sequencing at the beginning of the transcript



Cuffcompare - Cuffmerge

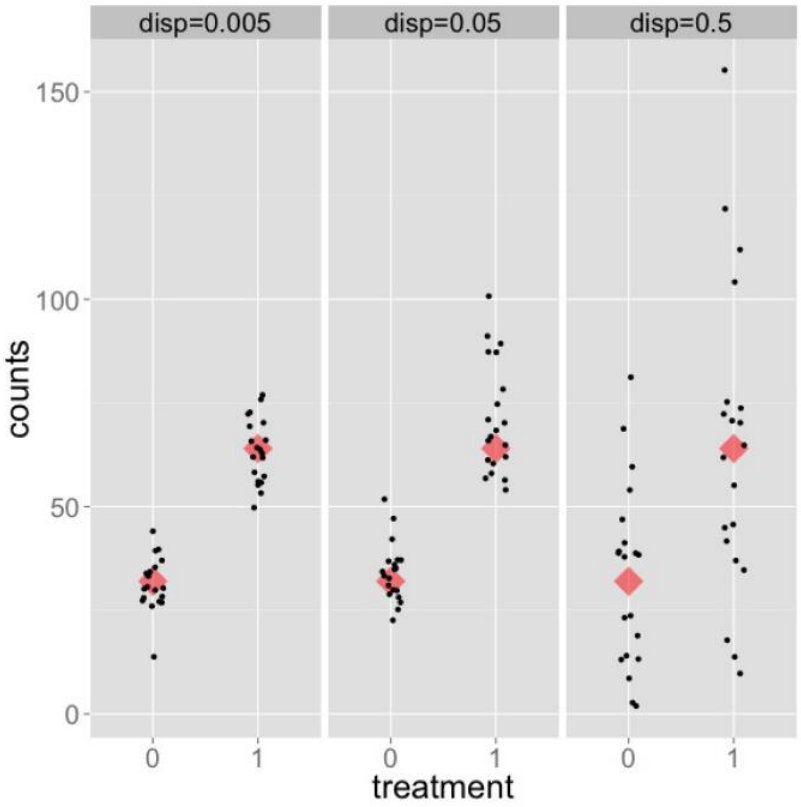
- Compare your assembled transcripts to a reference annotation
- Track Cufflinks transcripts across multiple experiments - samples

Tophat → Cufflinks → Cuffmerge → Cuffdiff



Determining Differentially Expressed Genes and Transcripts

Discover transcripts showing different average expression levels across two groups



http://www.molgen.mpg.de/1242892/rnaseq.pdf

The statistical model for finding differential expressed transcripts or genes depends on whether we have biological replicates. The advantage of having many replicates allows to learn about the biological variation within the conditions tested.

Model Used Negative Binomial

- Count data follows a Poisson distribution. Poisson assumes that the mean equals the variance . However, in RNA-Seq data, genes with larger mean cou er variance, which is de overdispersion probl
- **Negative binomial m DESeq, cuffdiff) ac term in the model**

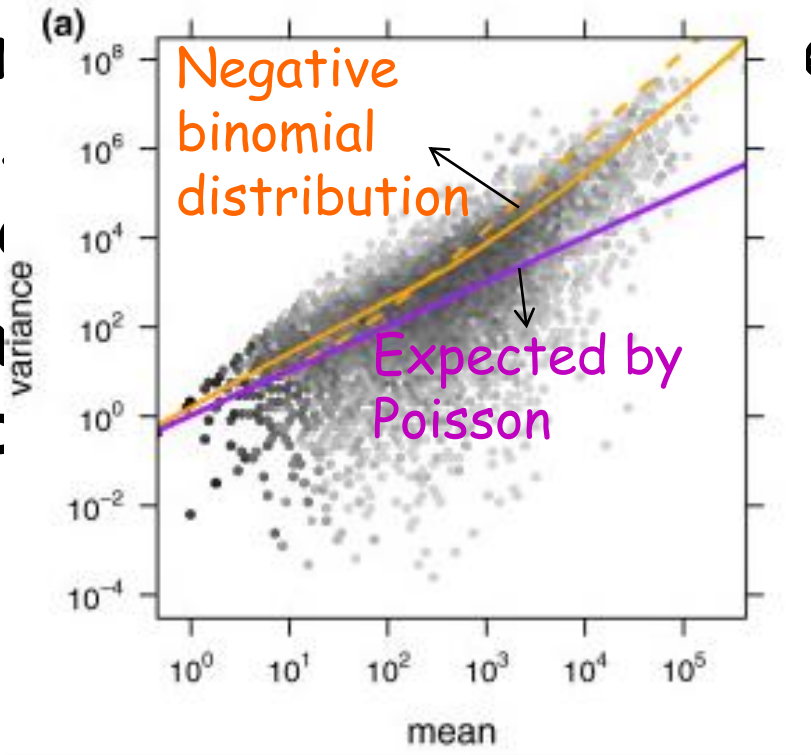


Fig. 1 from Anders & Huber, 2010: Dependence

RNA-Seq Challenges

- NGS challenge
  - Short sequences are produced
  - There is a PCR amplification step which inserts biases
- In RNA-Seq we sample the transcripts and the short and rare transcripts have a reduced ability to be sampled
- Transcriptomes are dominated by few highly abundant transcripts
- Multiple isoforms and gene families are problematic to quantify
- RNA preparations may contain incompletely processed RNAs or transcriptional noise

Assessment of transcript reconstruction methods for RNA-seq

Tamara Steijger, Josep F Abril, Pär G Engström, Felix Kokocinski, The RGASP Consortium, Tim J Hubbard, Roderic Guigó, Jennifer Harrow & Paul Bertone

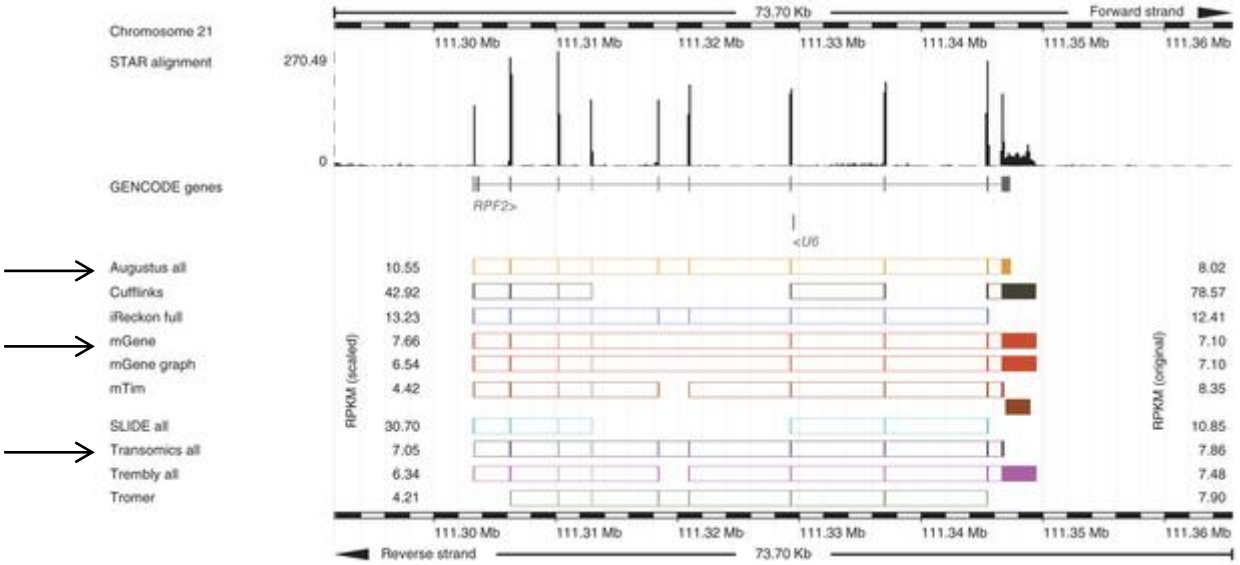
Affiliations | Contributions

Nature Methods **10**, 1177–1184 (2013) | doi:10.1038/nmeth.2714

Received 31 March 2013 | Accepted 23 September 2013 | Published online 03 November 2013

Examples of transcript calls and expression-level estimates.

Use a model of gene structure



No method achieved even 60% accuracy for transcript reconstruction in human

RNA-Seq Analysis Approaches

- *Align the reads to the genome*
- *Annotate-then-identify*
  - Use the known gene structure database to quantify the genes and transcripts
- *Assemble-then-identify*
  - Allow the aligned reads to identify novel exons and gene structures

Main Topics

- Introduction
- Experimental design issues
- Analysing RNA-Seq data
  - RNA-Seq pipeline: Tophat-Cufflinks-Cuffdiff
- Challenges

RNA-Seq Exercises

- Observe the Tuxedo outputs
- Use Chipster to perform RNA-Seq analysis
- Use Genome Browser IGV to analyse outputs

- THANKS
- Questions???