

Assembly and quantification of transcripts from RNA-Seq data

Dena Leshkowitz,
Introduction to Deep-Sequencing Data Analysis
for Biologists 2021

RNA-Seq Applications



- Transcript level analysis (to be discussed Dec 13th) :
 - Discover novel transcripts
 - Determine transcript structure
 - Measure transcripts expression
 - Detect differentially expressed transcripts/isoforms between conditions, treatments...
- Gene expression analysis for Model Organism:
 - Gene expression quantification
 - Detect differentially expressed genes between conditions, treatments... **based on known gene structures**

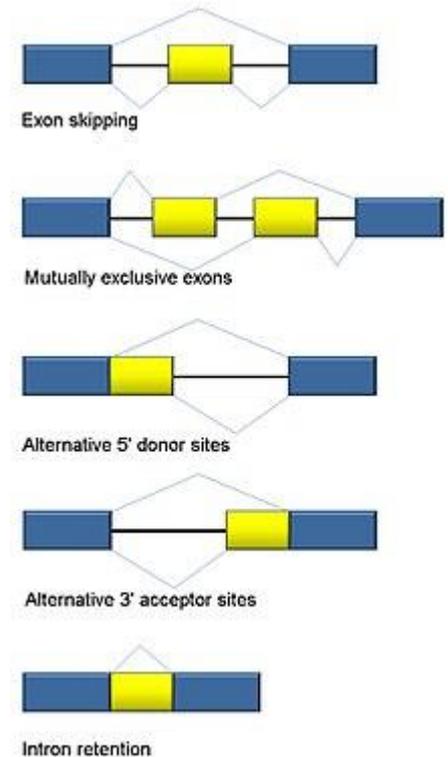
Main Topics

RNA-Seq transcript level analysis

- Introduction to alternative splicing
- Transcript level analysis - alignment based pipelines
- RNA-Seq alignment free quantification
- How to report the analysis

Why Transcript Level Analysis?

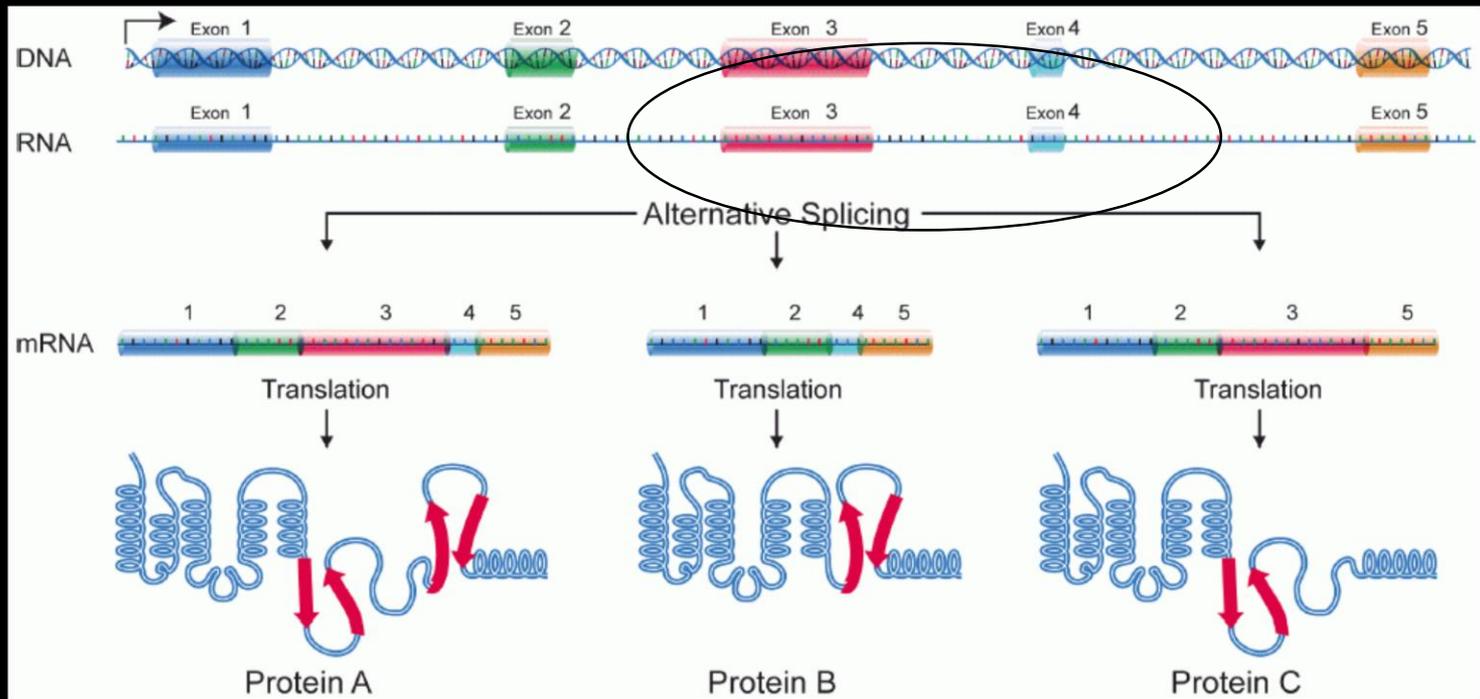
- The repertoire of transcripts is much larger than the genes and this also affects the proteins that are expressed
- Alternative RNA splicing is a process during gene expression that allows a single gene to produce many transcripts



GENCODE 2021			Protein-coding
Human	GENCODE 35	Genes	19954
		Transcripts	154580
Mouse	GENCODE M25	Genes	21859
		Transcripts	102241

Alternative Splicing

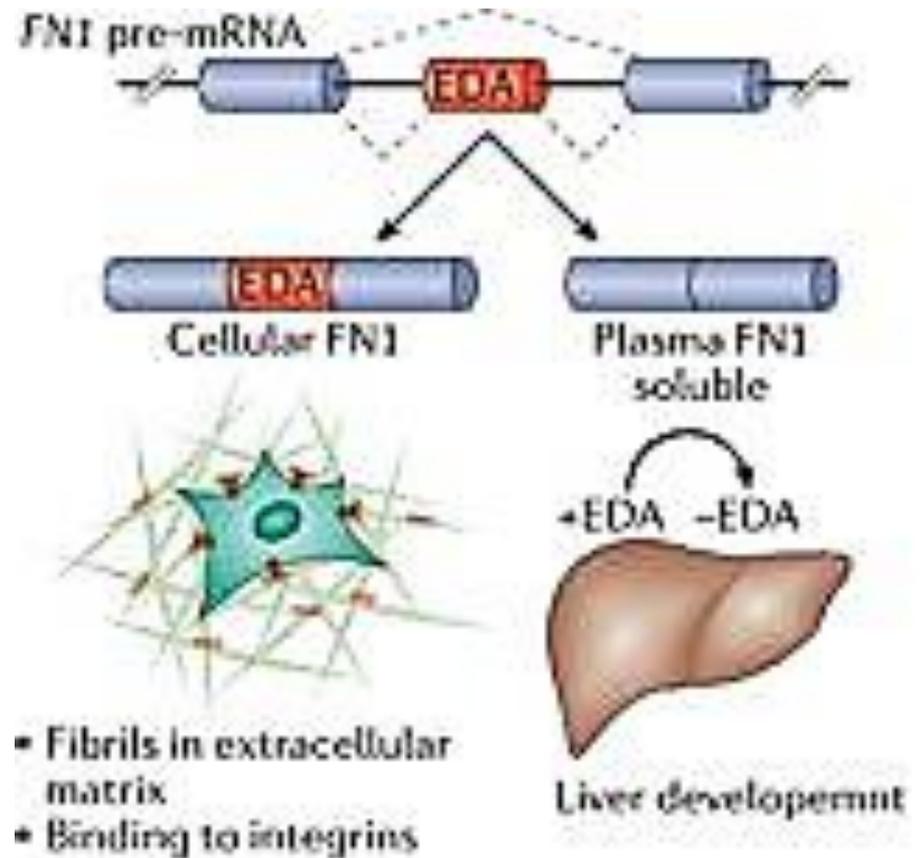
- Alternative splicing increases the biodiversity of proteins that can be encoded by the genome.
- In humans, ~95% of multi-exonic genes are alternatively spliced



Alternatively spliced exons yield three different protein isoforms

Example - Alternative Splicing

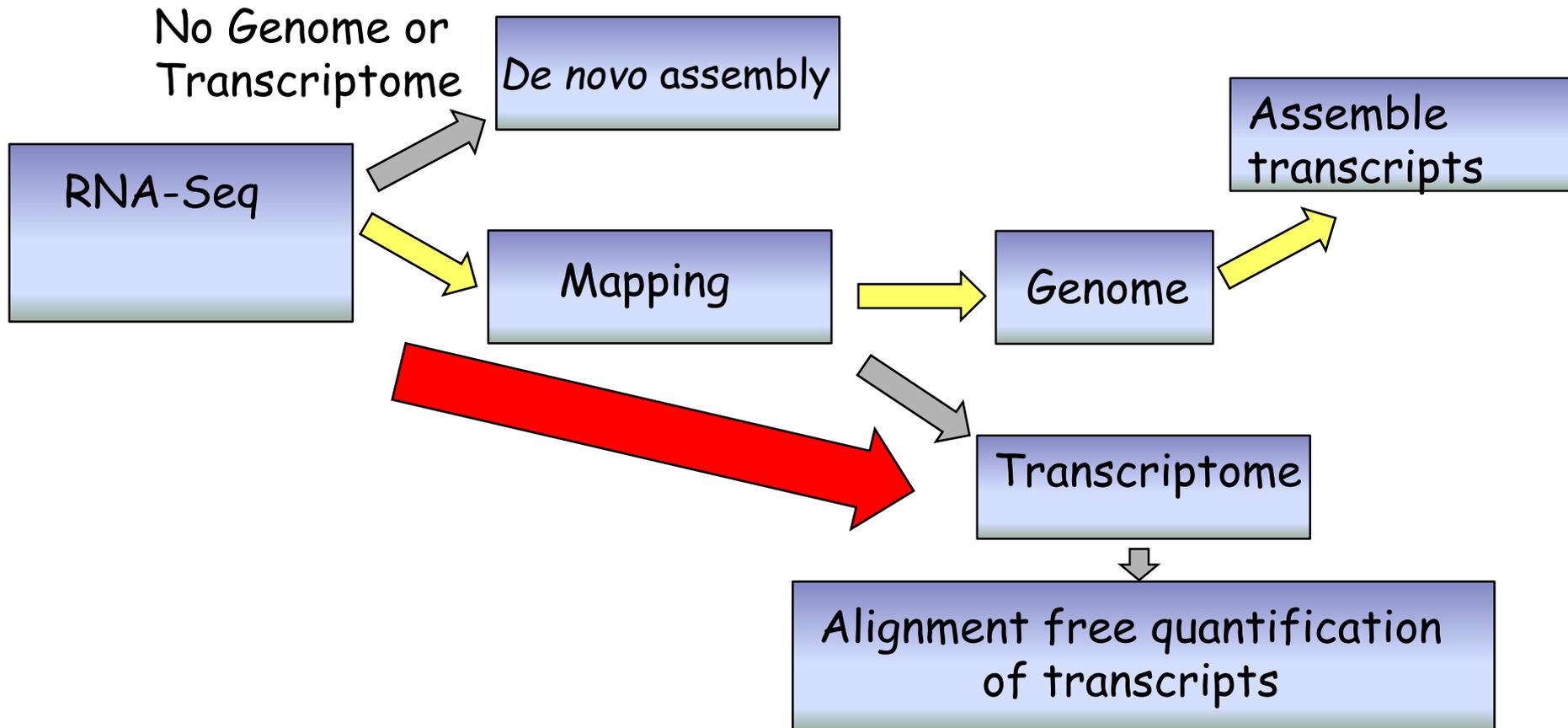
- Fibronectin (FN1) was one of the first genes for which alternative splicing was described and its mechanism studied.
- The figure illustrates the retention or exclusion of the extradomain-A exon (EDA; dashed lines), which controls FN1 tissue localization.



Nature Reviews Molecular Cell Biology volume 18 (2017)

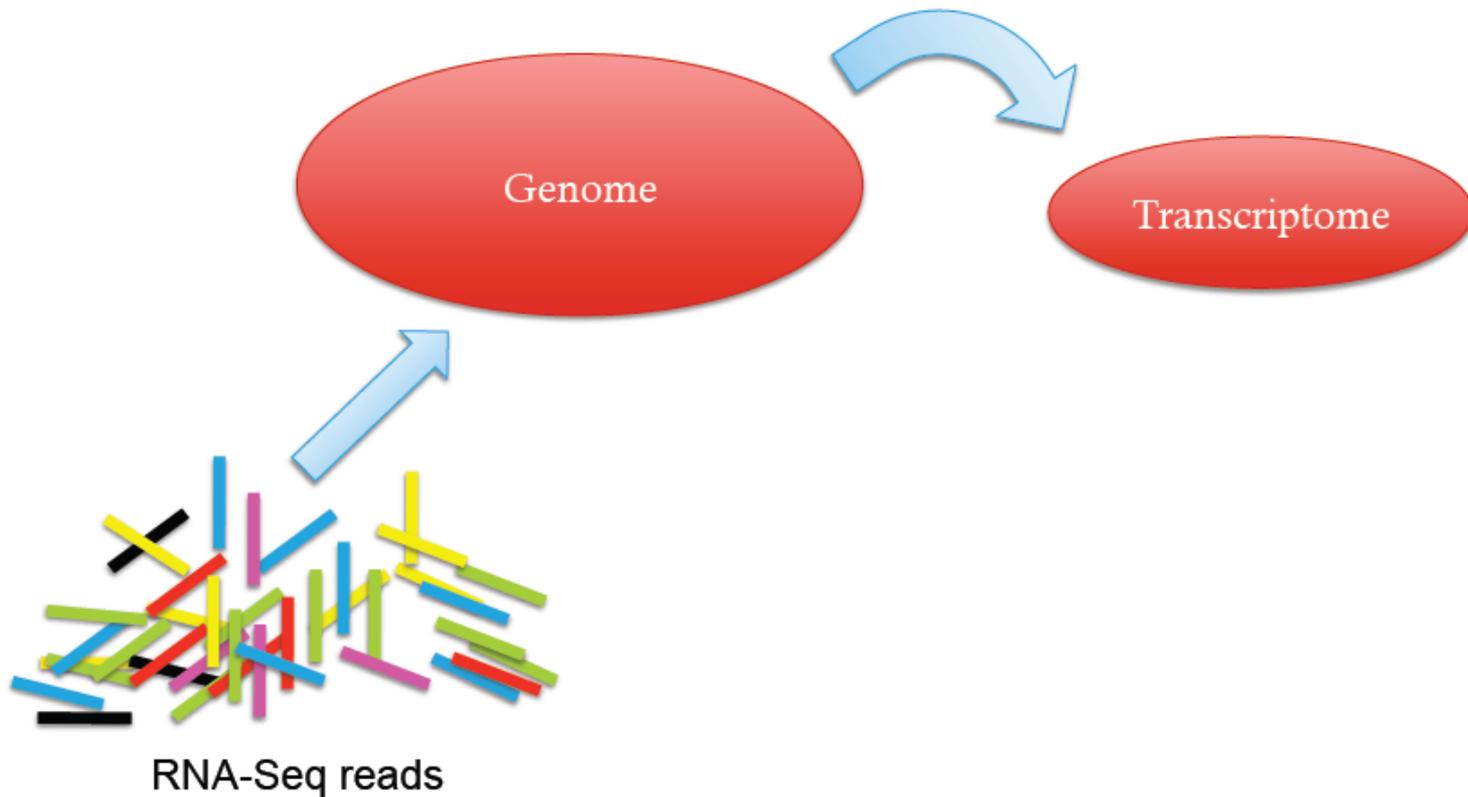
From Sequences to Transcriptome Analysis

```
... ACCGTAATGGGCTGATCATGCTTAA  
TGATCATGCTTAAACCCCTGEGCATCCTACTG ...  
... ACCGTAATGGGCTGATCATGCTTAAACCCCTGEGCATCCTACTG ...
```



Mapping Reads to the Genome

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





Tuxedo

TopHat

spliced read alignment

Bowtie

Unspliced alignment



Cufflinks

- Isoform assembly
- quantitation

NATURE BIOTECHNOLOGY | RESEARCH | LETTER

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

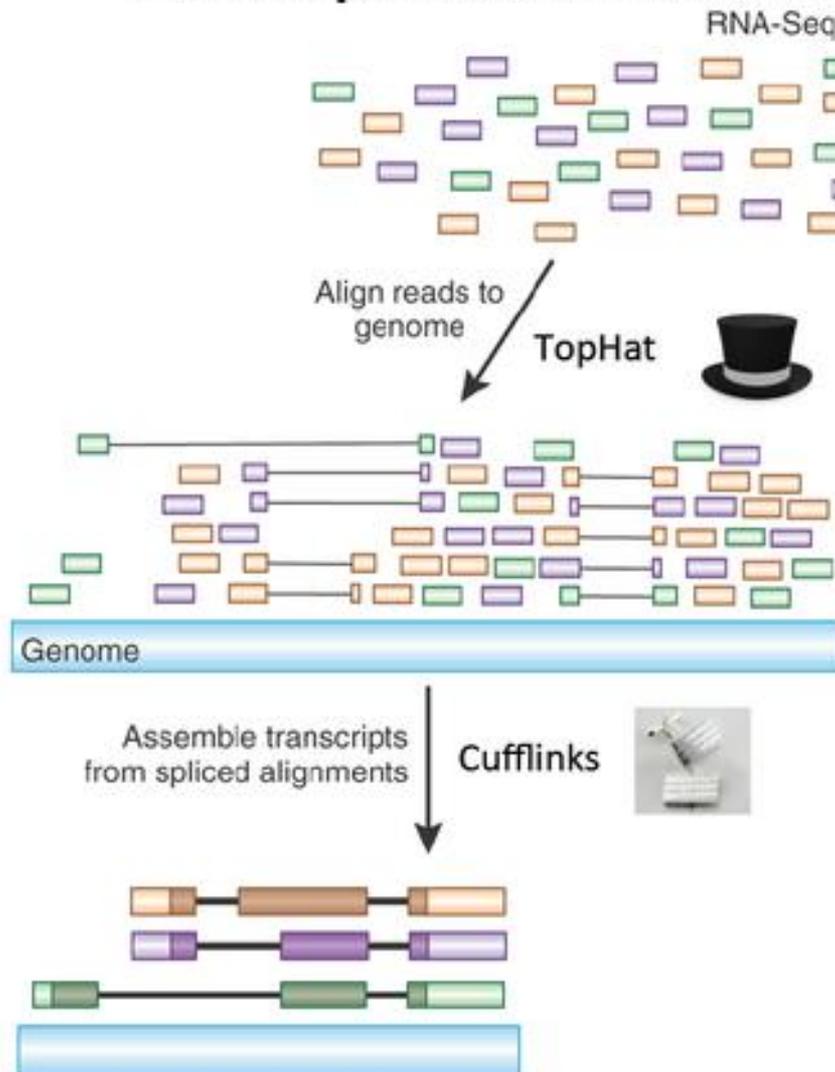
Cole Trapnell, Brian A Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J van Baren, Steven L Salzberg, Barbara J Wold & Lior Pachter

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

Transcript Reconstruction from RNA-Seq Reads



The Tuxedo Suite: End-to-end Genome-based RNA-Seq Analysis Software Package

NATURE PROTOCOLS | PROTOCOL

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

Cole Trapnell, Adam Roberts, Loyal Goff, Geo Pertes, Daehwan Kim, David R Kelley, Harold Pimentel, Steven L Salzberg, John L Rinn & Lior Pachter

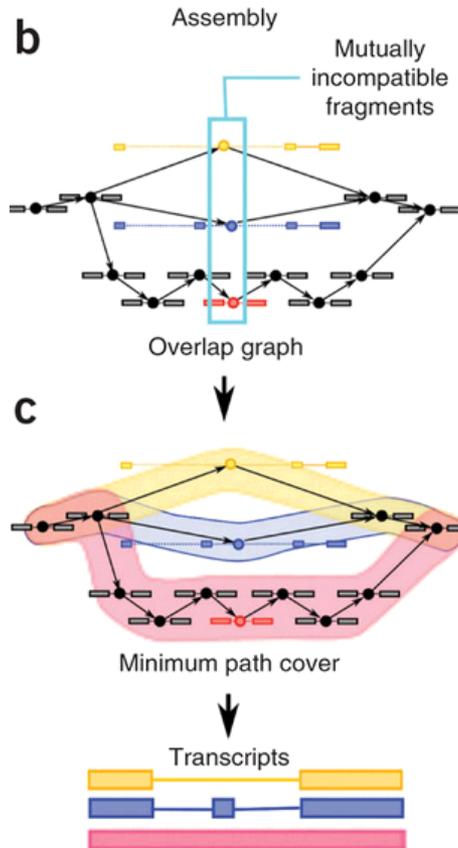
Affiliations | Contributions | Corresponding author

Nature Protocols 7, 562–578 (2012) | doi:10.1038/nprot.2012.016
Published online: 01 March 2012

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."

Cufflinks Assembly



- 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



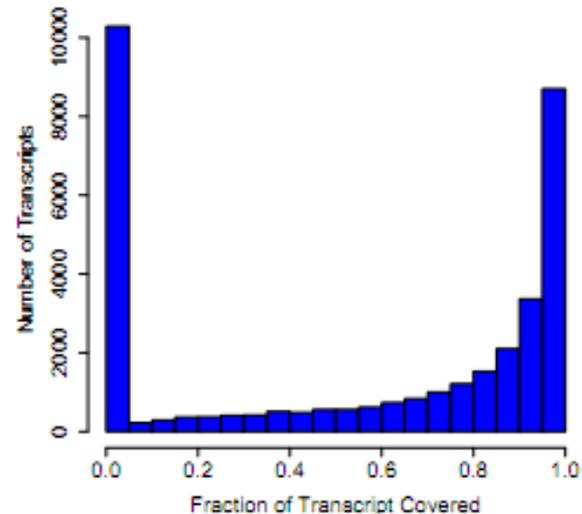
Transcripts
and their
abundances

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

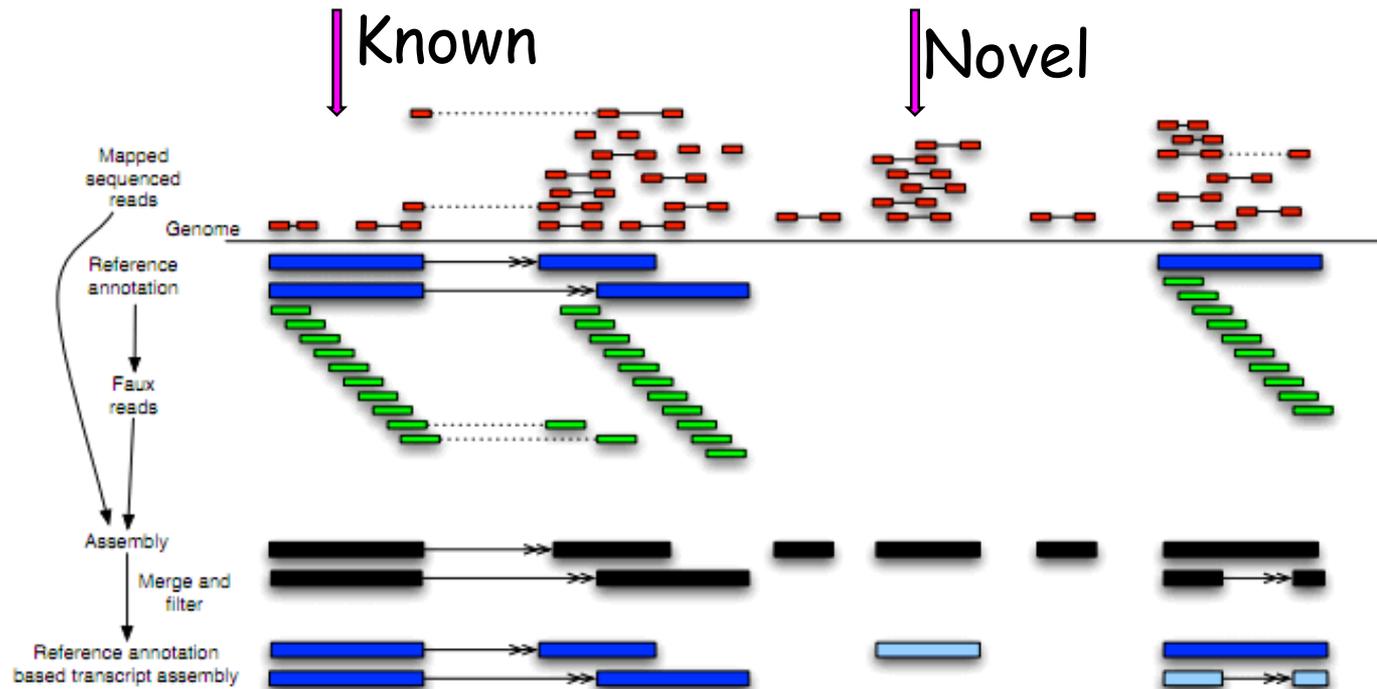
Cufflinks -RABT

- 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

Gene transfer format (GTF)

GTF file is used to hold information about gene structure. It is a tab-delimited text format

```
chr1 unknown exon 3214482 3216968 . - . gene_id "Xkr4"; gene_name "Xkr4"; p_id "P15391"; transcript_id "NM_001011874"; tss_id "TSS27105";
chr1 unknown stop_codon 3216022 3216024 . - . gene_id "Xkr4"; gene_name "Xkr4"; p_id "P15391"; transcript_id "NM_001011874"; tss_id "TSS27105";
chr1 unknown CDS 3216025 3216968 . - 2 gene_id "Xkr4"; gene_name "Xkr4"; p_id "P15391"; transcript_id "NM_001011874"; tss_id "TSS27105";
chr1 unknown CDS 3421702 3421901 . - 1 gene_id "Xkr4"; gene_name "Xkr4"; p_id "P15391"; transcript_id "NM_001011874"; tss_id "TSS27105";
chr1 unknown exon 3421702 3421901 . - . gene_id "Xkr4"; gene_name "Xkr4"; p_id "P15391"; transcript_id "NM_001011874"; tss_id "TSS27105";
chr1 unknown CDS 3670552 3671348 . - 0 gene_id "Xkr4"; gene_name "Xkr4"; p_id "P15391"; transcript_id "NM_001011874"; tss_id "TSS27105";
chr1 unknown exon 3670552 3671498 . - . gene_id "Xkr4"; gene_name "Xkr4"; p_id "P15391"; transcript_id "NM_001011874"; tss_id "TSS27105";
chr1 unknown start_codon 3671346 3671348 . - . gene_id "Xkr4"; gene_name "Xkr4"; p_id "P15391"; transcript_id "NM_001011874"; tss_id "TSS27105";
chr1 unknown exon 4290846 4293012 . - . gene_id "Rp1"; gene_name "Rp1"; p_id "P17361"; transcript_id "NM_001195662"; tss_id "TSS6138";
chr1 unknown stop_codon 4292981 4292983 . - . gene_id "Rp1"; gene_name "Rp1"; p_id "P17361"; transcript_id "NM_001195662"; tss_id "TSS6138";
```

GTF format

GTF (Gene Transfer Format) is a refinement to GFF that tightens the specification. The first eight GTF fields are the same as GFF. The *group* field has been expanded into a list of *attribute*. Each attribute consists of a type/value pair. Attributes must end in a semi-colon, and be separated from any following attribute by exactly one space.

The attribute list must begin with the two mandatory attributes:

- **gene_id value** - A globally unique identifier for the genomic source of the sequence.
- **transcript_id value** - A globally unique identifier for the predicted transcript.

Example:

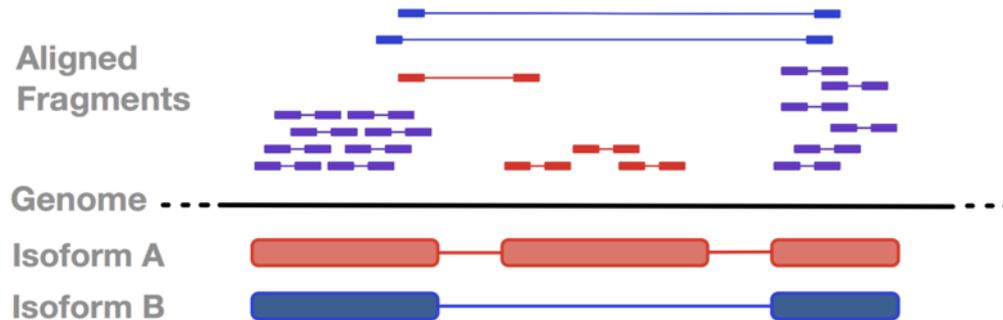
Here is an example of the ninth field in a GTF data line:

```
gene_id "Em:U62317.C22.6.mRNA"; transcript_id "Em:U62317.C22.6.mRNA"; exon_number 1
```

The Genome Browser groups together GTF lines that have the same *transcript_id* value. It only looks at features of type *exon* and *CDS*.

For more information on this format, see <http://mblab.wustl.edu/GTF2.html>. If you would like to obtain browser data in GTF format, please refer to [Genes in gtf or gff format](#) on the wiki.

Estimate Transcripts Abundance

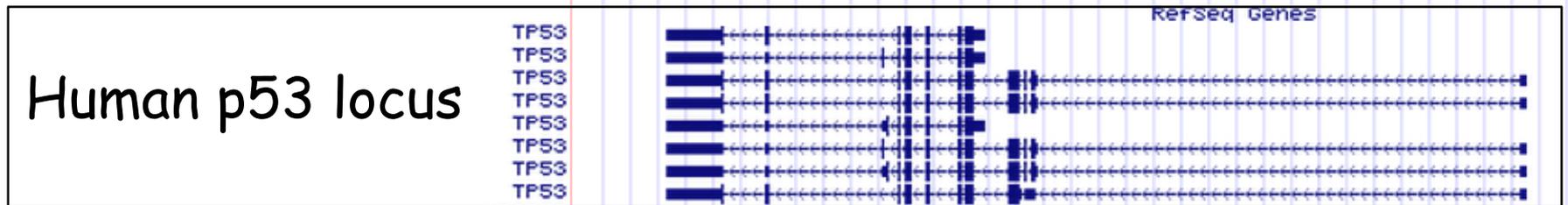


Trapnell et al. Nat Biotechnol. 2010 May;28(5):511-5.

From which transcript did the **purple reads** originate?

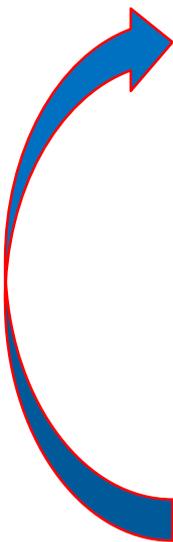
Quantification Problem

- We encounter the same problem when we align directly to a transcriptome
- Counting the number of sequences that map uniquely to transcripts results in false estimates of alternatively spliced transcripts, why?
- Computational challenge: How to use reads that map ambiguously between isoforms and genes
- Solution: EM algorithm

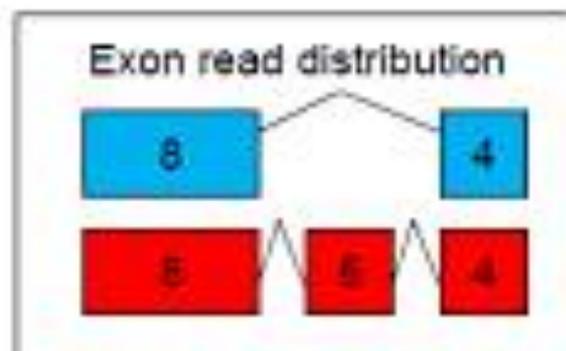
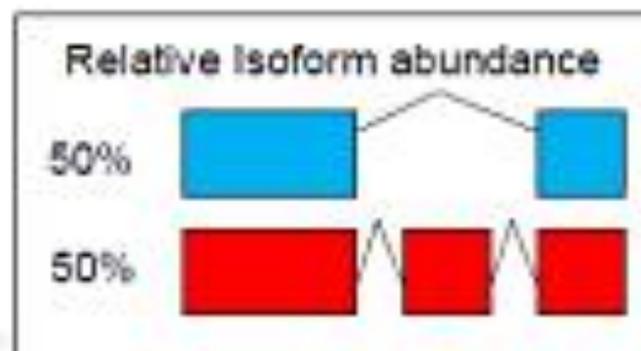
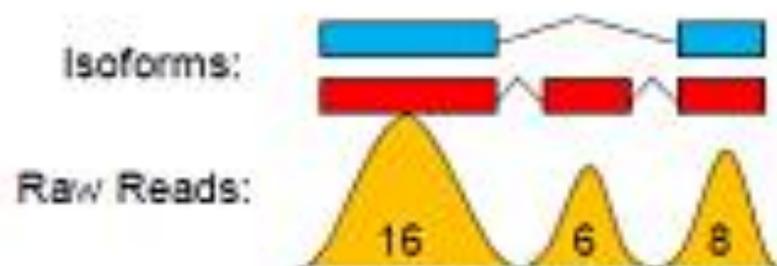


Isoform Expression Quantification Expectation Maximization Algorithm

Computing relative abundance of transcripts:

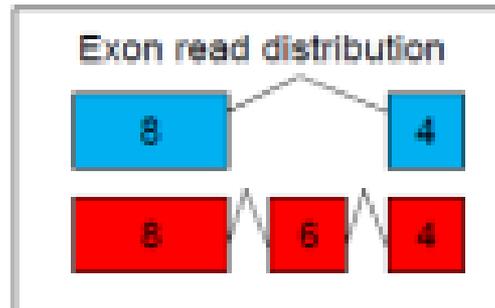
- Step 1- Assume isoforms are equally abundant (in case of 2 transcripts abundance is: $\frac{1}{2}$ and $\frac{1}{2}$)
 - Step 2 - Distribute the reads to the isoforms based on the abundance
 - Step 3 - Recalculate the isoforms abundance based on the reads counts and isoforms length
 - Step 4- If abundance has changed go back to step 2 otherwise stop
- 

1st step E/M algorithm



Calculating Abundance after 1st EM Cycle

In the calculation of transcript relative abundance we need to normalize for the length, why?



Total reads	Length
12	300
18	400

Exon length (bp) 200 100 100

The red transcript relative abundance (proportion) after the first cycle:

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

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

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

Relative Abundance Calculation Using 100 Iterations of EM

	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 0.25

Red 0.75

Normalization

Normalization at the transcript level requires to account for:

- Between samples accounting for number of reads/fragments
- Between genes/transcripts accounting for different length
 - The RNA-seq technology is **more sensitive to longer transcripts** because it generates more reads from longer transcripts.

Normalized Expression Values

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

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

C= The number of fragments mapped to the transcript exons

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

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

$$FPKM_i = 10^6 \times 10^3 \times \frac{C_i}{NL_i}$$

- When are the RPKM and the FPKM values the same?
- Problem the sum of FPKM for different samples is not necessarily the same

TPM (Transcripts per million)

- Divide the fragment counts by the length of each transcript in kilobases. This gives you fragments per kilobase (FPK).
- Scaling factor - Count up all the FPK values in a sample and divide this number by 1,000,000. This is your "per million" scaling factor.
- Divide the FPK values by the "per million" scaling factor. This gives you TPM.

TPM sum in all samples is a million, better way to normalize.

Metrics for quantifying gene expression levels

➤ RPKM

- Reads Per Kilobase per Million mapped reads
- Normalize relative to sequencing depth and gene length

➤ FPKM

- Similar to RPKM but count **DNA fragments** instead of reads
- Used in paired end RNA-Seq experiments to avoid bias

➤ TPM

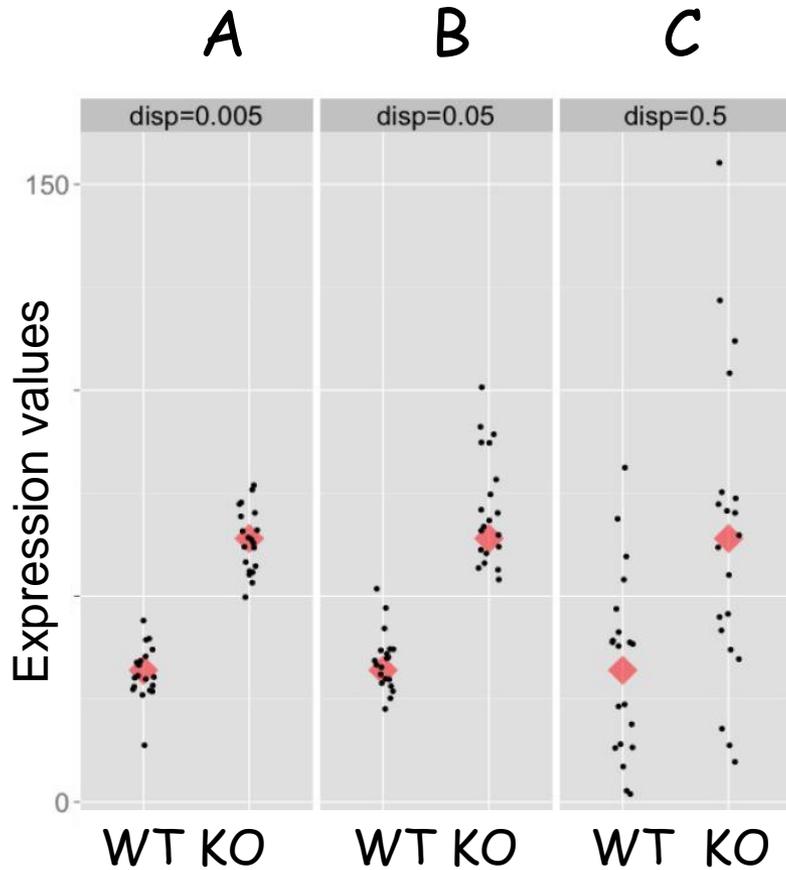
- Transcripts Per Million
- Normalize for transcript length, then normalize by sequencing depth

$$TPM = \frac{RPKM}{\sum(RPKM)} \times 10^6$$

More reading

https://www.reneshbedre.com/blog/expression_units.html

Determining Differentially Expressed Transcripts & Genes



- Aim: finding genes & transcripts which have a significant difference between the groups which is larger than the "noise" - variation within the groups
- The advantage of having many replicates - allows us to learn about the biological variation within the groups tested

<http://www.molgen.mpg.de/1242892/rna-seq.pdf>

Negative Binomial Distribution

- Cuffdiff tests for differential expression by the use of negative binomial generalized linear models

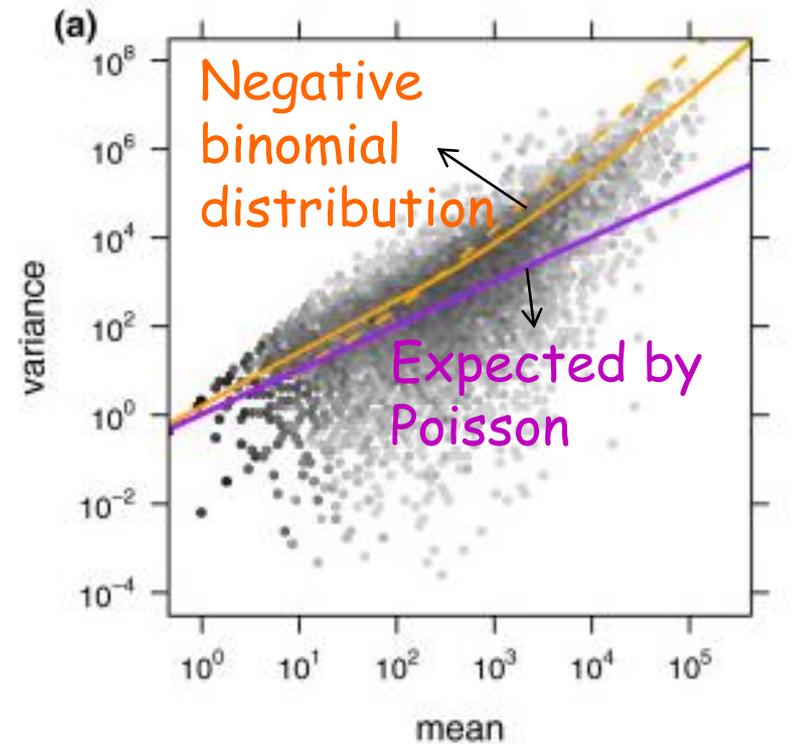
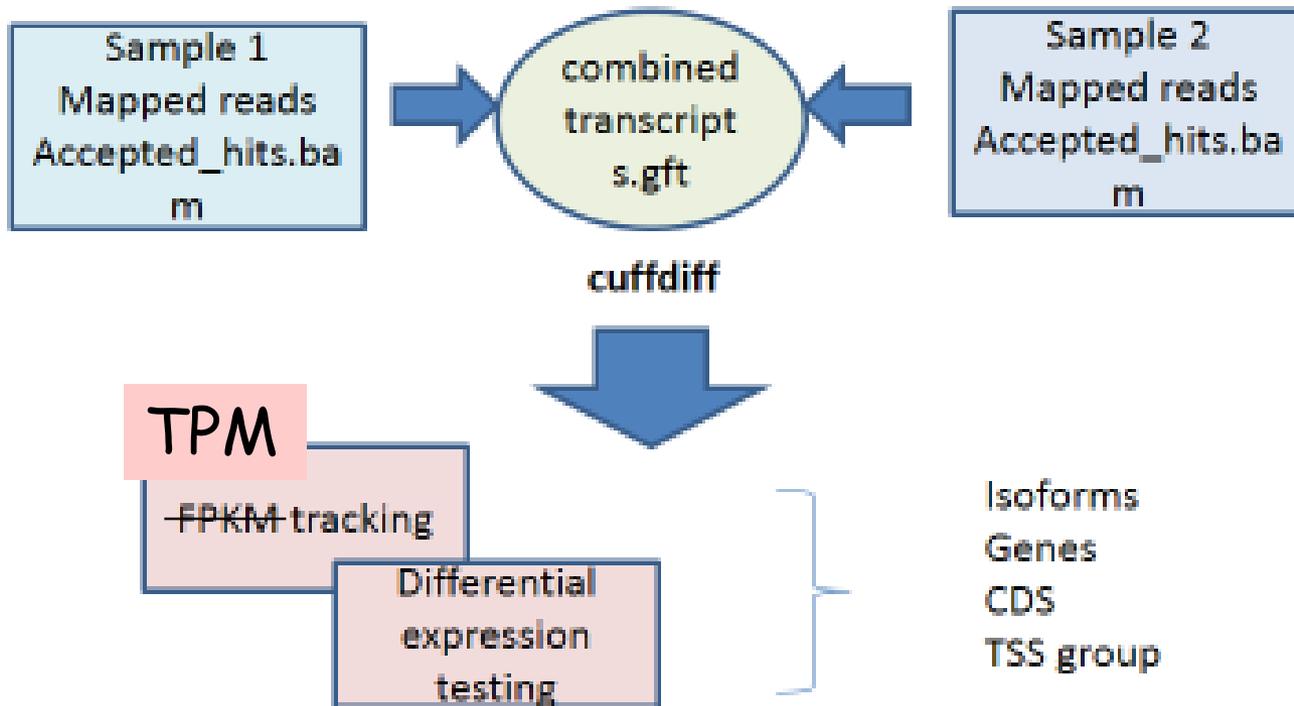


Fig. 1 from Anders & Huber, 2010

Cuffdiff



Genes expression is calculated as the sum of its isoform expression levels

RNA-Seq analysis on the transcript level

- Per sample map Reads to a Genome (using known annotation) (TopHat2)
- Per sample assemble transcripts (Cufflinks)
- Merge assembled transcripts built for the various samples into one "combined transcripts.gtf"
(Cuffmerge)
- Per sample quantify merged pool of transcripts (Cuffdiff)
- Normalize counts (Cuffdiff)
- Detect differentially expressed transcripts (Cuffdiff)

What is the Advantage of Longer and PE Reads?



- Reads mapping to junctions
 - With longer reads we will have more reads spanning exons



- Paired end reads
 - Knowing both ends of a fragment and an approximation of fragment size helps to determine the transcript from which it was derived.

Experimental Design

Mammalian tissue

Liu Y. et al., 2014; ENCODE 2011 RNA-Seq

Differential gene expression profiling	10-25M	50 base single-end
Alternative splicing	50-100M	100 base paired-end
Allele specific expression	50-100M	100 base paired-end
De novo assembly	>100M	100 base paired-end

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

Example of transcript calls and expression-level estimates

Results were evaluated from methods based on genome alignments (Augustus, Cufflinks, Exonerate, GSTRUCT, iReckon, mGene, mTim, NextGeneid, SLIDE, Transomics, Trembly and Tromer) as well as *de novo* assembly (Oases and Velvet).

Programs were run without genome annotation, aside from iReckon and SLIDE

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

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

Mihaela Pertea, Geo M Pertea, Corina M Antonescu, Tsung-Cheng Chang, Joshua T Mendell & Steven L Salzberg

[Affiliations](#) | [Contributions](#) | [Corresponding author](#)

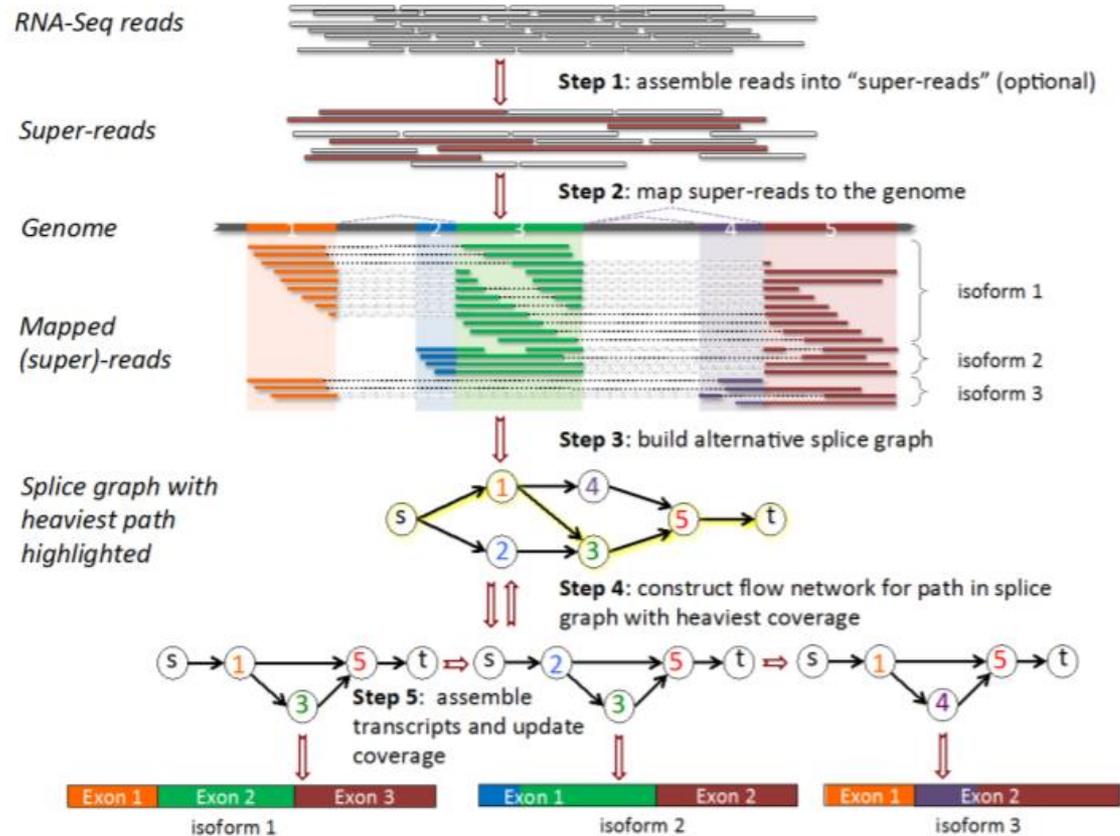
Nature Biotechnology 33, 290–295 (2015) | doi:10.1038/nbt.3122

Received 15 April 2014 | Accepted 09 December 2014 | Published online 18 February 2015

- **Faster**
- **More complete and accurate reconstructions and better estimates of expression levels**
- **Implements the maximum flow algorithm**

Stringtie Algorithm Steps:

1. Pre-assembles paired reads - super reads
2. Uses mapping of reads to the reference genome to build an alternative overlap graph
3. The path with the **heaviest coverage** is used to build a transcript
4. The assembly of transcripts and estimation of expression level is done simultaneously
5. The assembled transcript is removed from the splice graph and step 3-4 repeated until there is no more transcripts



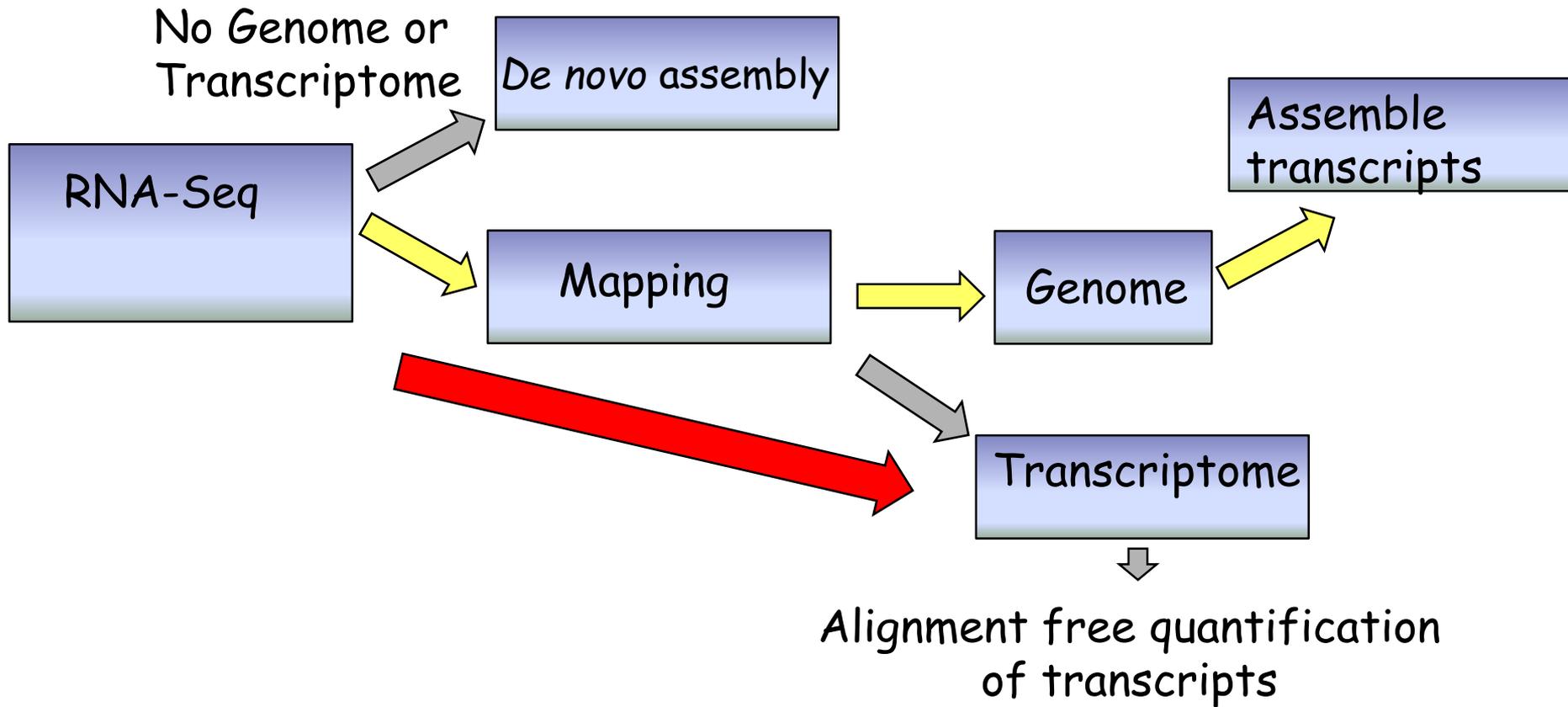
Supplementary Figure 12. The StringTie algorithm: RNA-seq reads are assembled into super-reads (Step 1) and then super-reads plus un-assembled reads are mapped to the genome (Step 2). In Step 3, mapped reads and super-reads are used to build an alternative splice graph. We use the path from source (s) to sink (t) with the heaviest coverage to build a flow network corresponding to the transcript represented by that path (Step 4). The maximum flow in this network represents the coverage of one assembled transcript, which is removed from the splice graph (Step 5). Steps 4 and 5 are repeated until no more transcripts can be assembled.

Main Topics

RNA-Seq transcript level analysis

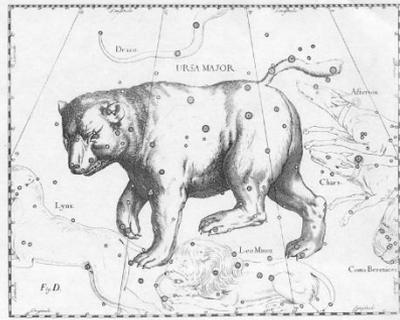
- Introduction to alternative splicing
- Transcript level analysis - alignment based pipelines
- **RNA-Seq alignment free quantification**
- How to report the analysis

Basic Strategies for Transcript Analysis from RNA-Seq Data

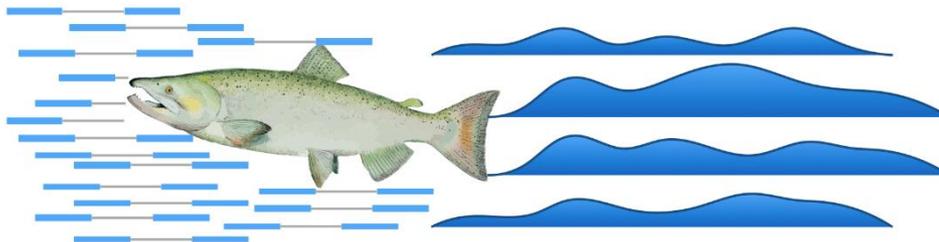


Alignment Free Quantification

- Kallisto: Bray, N.L., Pimentel, H. Melsted, P. & Lior Pachter (2016) Near-optimal probabilistic RNA-seq quantification Nature Biotechnology 34, 525-527

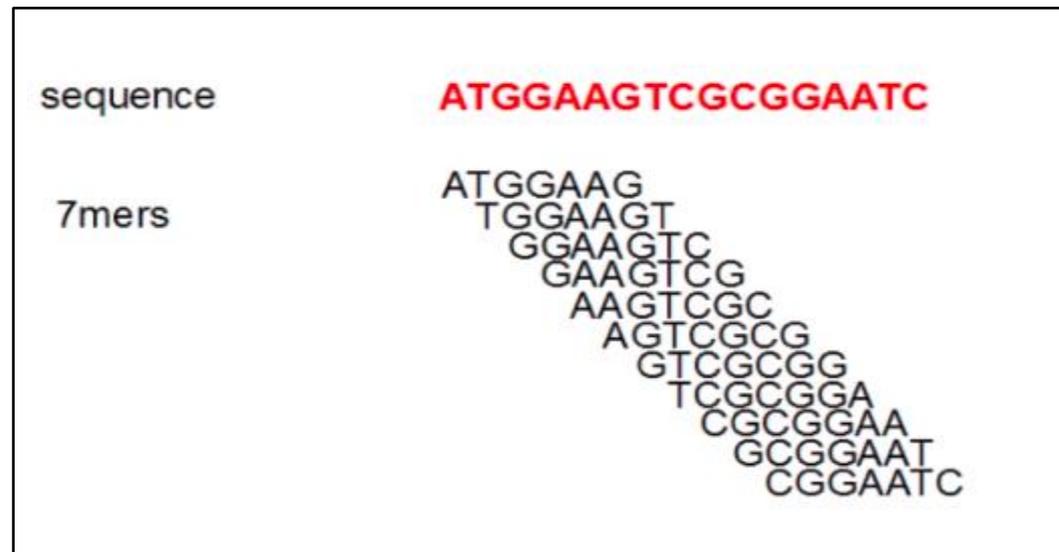


- Salmon : Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., & Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. Nature Methods.



Pseudo Alignment - for quantification purposes

- Identifying the transcripts from which the reads/fragments could have originated
- Does not try to pinpoint exactly how the sequences of the reads and transcripts align
- Pseudo alignment uses kmers -short sequences



Kallisto Method

The input :

- Reference transcriptome
- Reads from an RNA-seq experiment

Method:

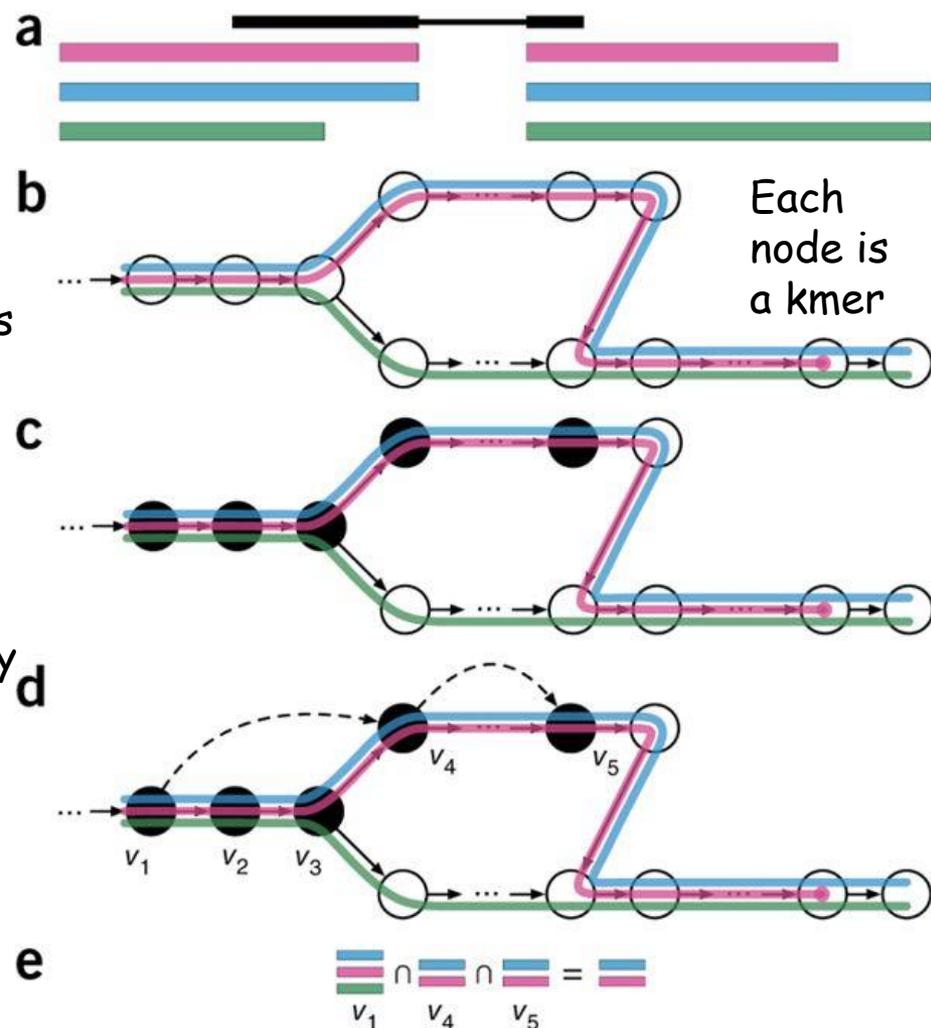
(a) An example of a read (in black) and three overlapping transcripts with exonic regions as shown.

(b) Construct a de-Bruijn graph from the transcriptome (T-DBG), where nodes (v_1, v_2, v_3, \dots) are k -mers, each transcript corresponds to a colored path.

(c) Evaluate k -mers of reads for compatibility with t-DBG.

(d) Can skip redundant k -mers (have the same k -compatibility class).

(e) The k -compatibility class of the read is determined by taking the intersection of the k -compatibility classes of its constituent k -mers.



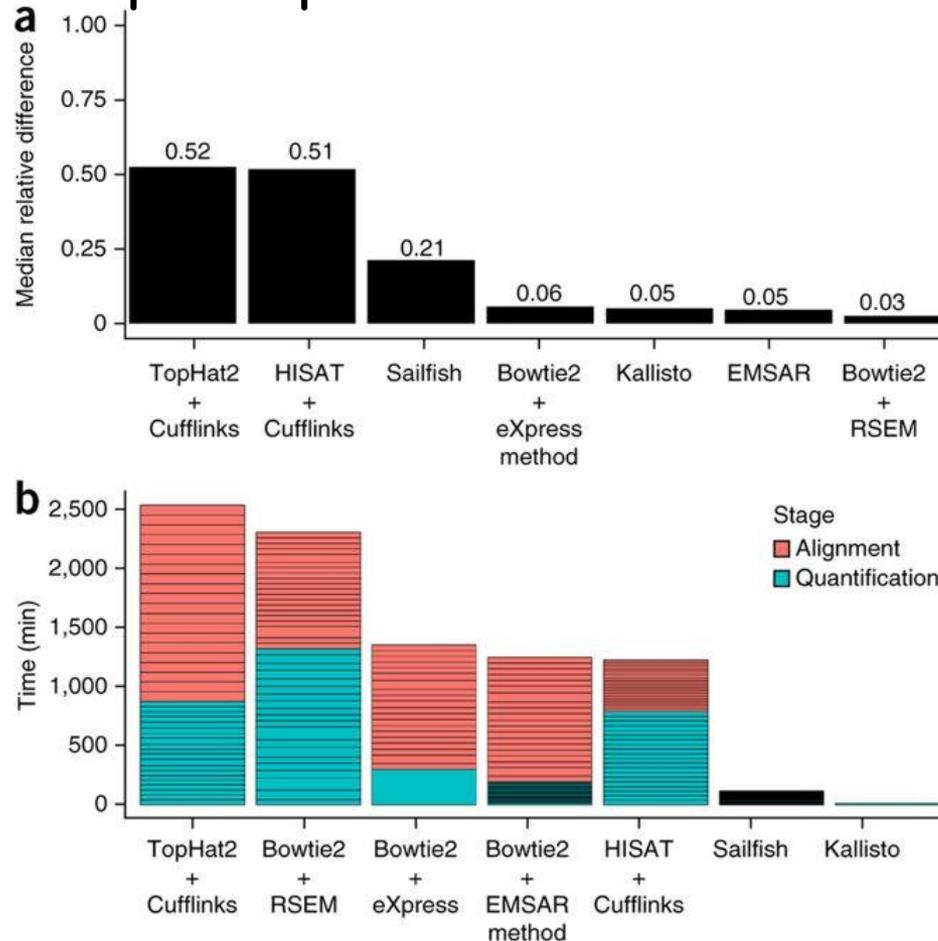
Kallisto

- Kallisto pseudoaligns reads to a reference transcript database, producing a list of transcripts that are compatible with each read while avoiding alignment of individual bases
- EM algorithm deconvolutes pseudoalignments to obtain transcript abundance
- Advantage : fast

Performance of Kallisto and other methods

Kallisto can quantify 30 million human reads in less than 3 minutes on a Mac desktop computer

Relative difference is defined as the absolute difference between the estimated abundance and the ground truth divided by the average of the two.

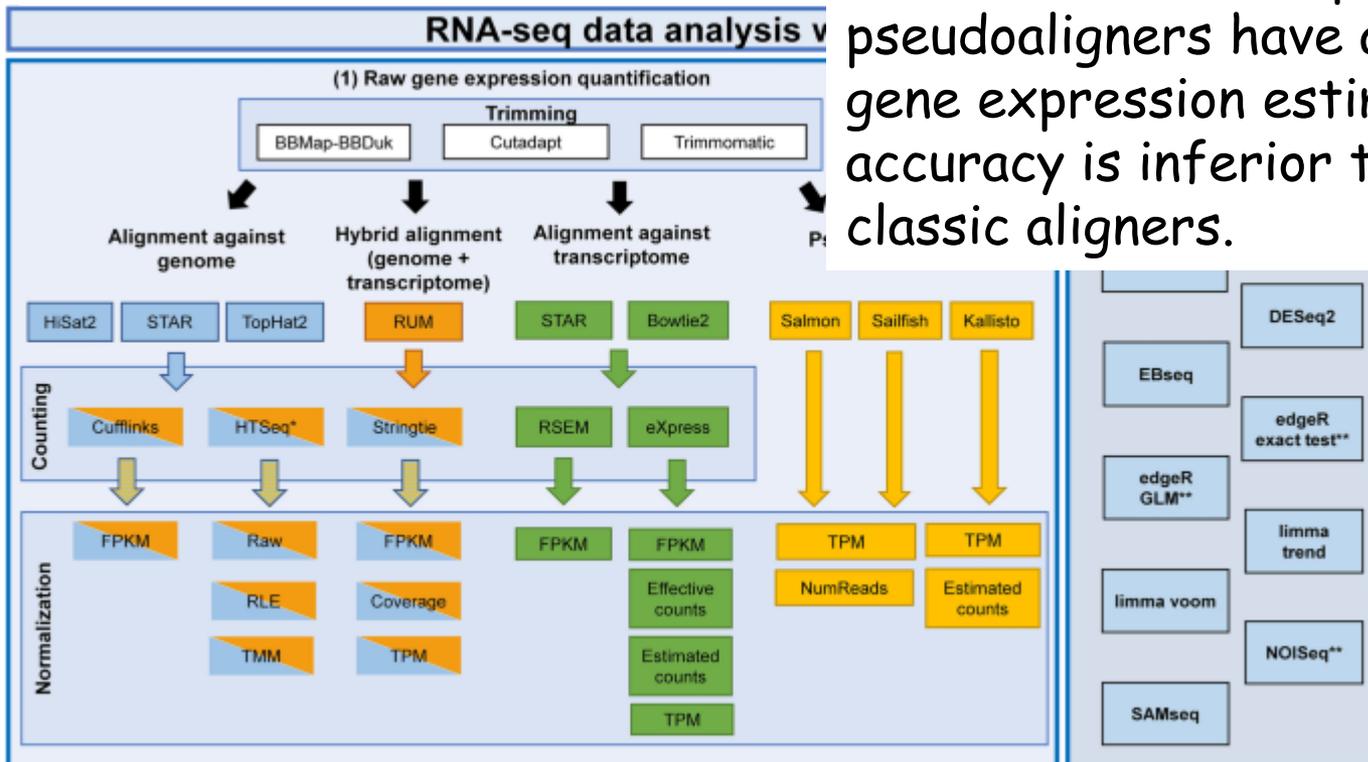


Systematic comparison and assessment of RNA-seq procedures for gene expression quantitative analysis

Luis A. Corchete , Elizabeth A. Rojas, Diego Alonso-López, Javier De Las Rivas, Norma C. Gutiérrez & Francisco J. Burguillo

Scientific Reports 10, Article number: 19737 (2020) | [Cite this article](#)

We have seen that pipelines based on pseudoaligners have a good precision in gene expression estimation, but their accuracy is inferior to that of the classic aligners.



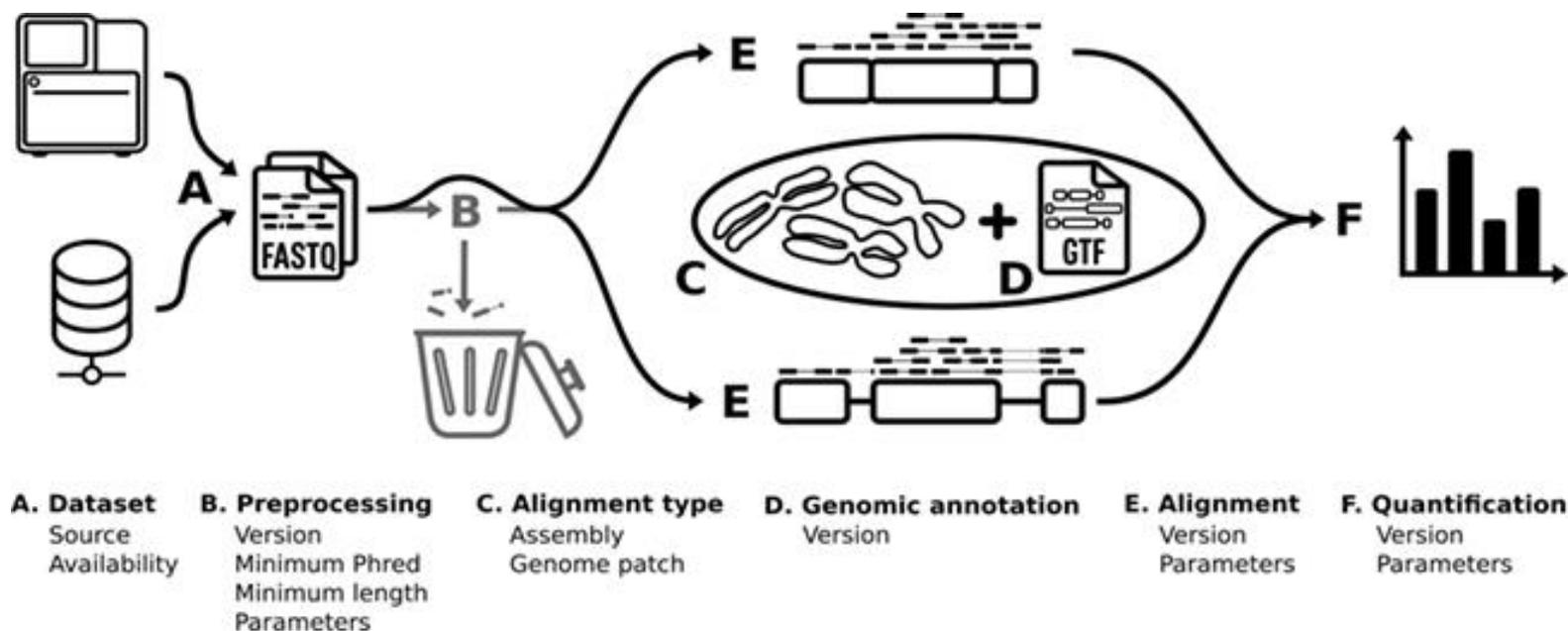
Current RNA-seq methodology reporting limits reproducibility

Joël Simoneau, Simon Dumontier, Ryan Gosselin, Michelle S Scott ✉

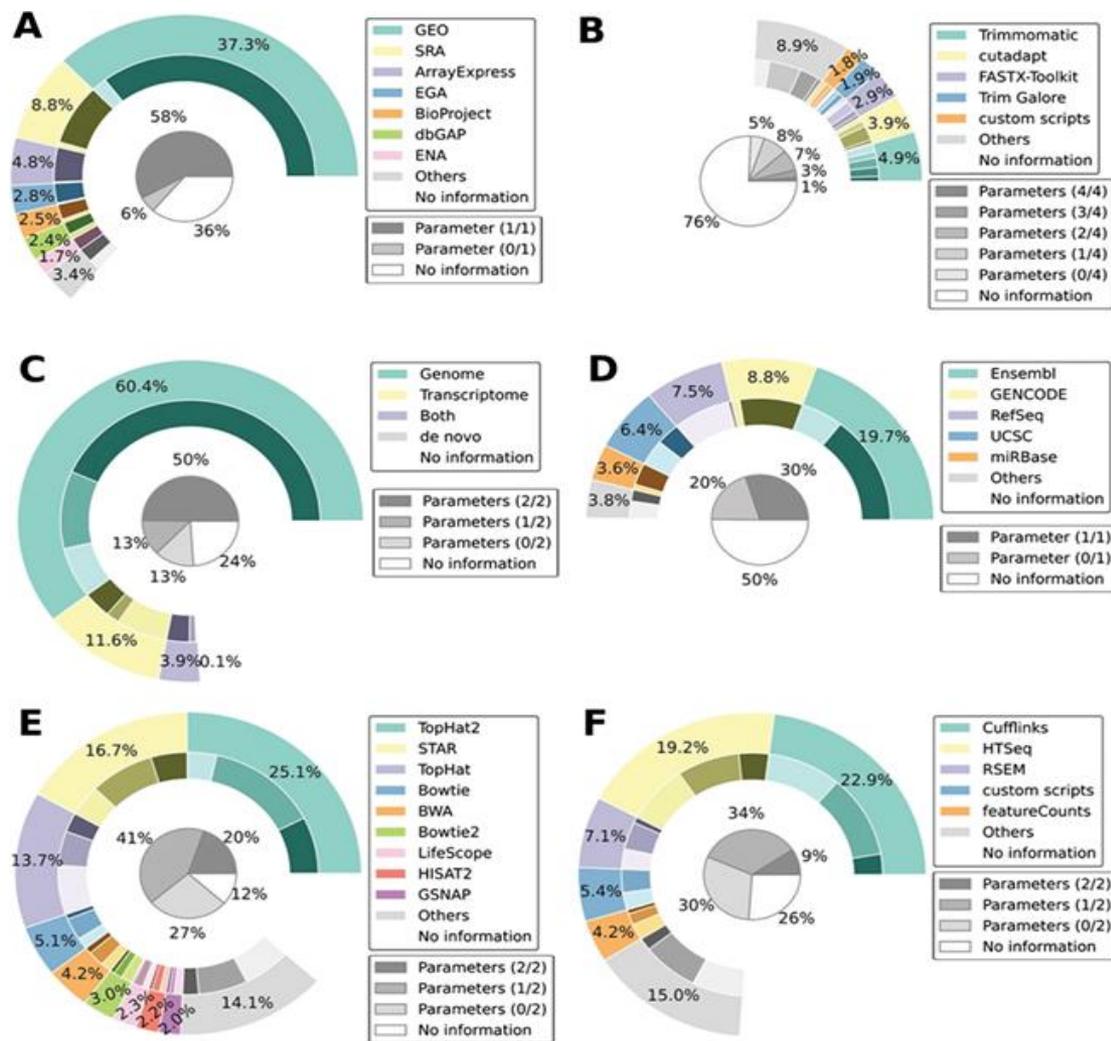
Briefings in Bioinformatics, bbz124, <https://doi.org/10.1093/bib/bbz124>

Published: 10 December 2019 Article history ▾

Figure 1 RNA-seq bioinformatics pipeline. Schematic of the RNA-seq bioinformatics methodology.



RNA-seq reported methodology is incomplete. Distribution of software and reference usage for the six methodological steps of an RNA-seq experiment (A. dataset, B. preprocessing tool, C. alignment type, D. genomic annotation, E. alignment tool and F. quantification tool).



THANKS

Questions?