

To map or not to map?

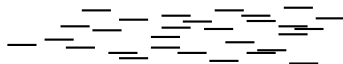
Formation RNA-Seq – Lille

Mikaël Salson

`mikael.salson@univ-lille.fr`

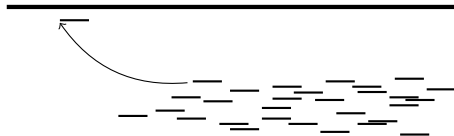
RNA-Seq read mapping

Reference genome

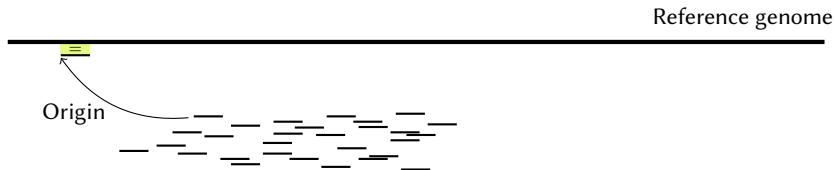


RNA-Seq read mapping

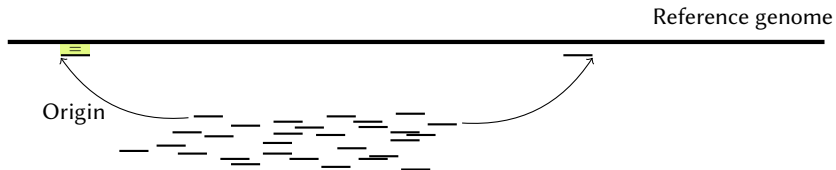
Reference genome



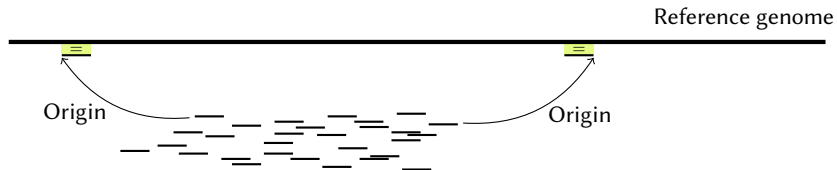
RNA-Seq read mapping



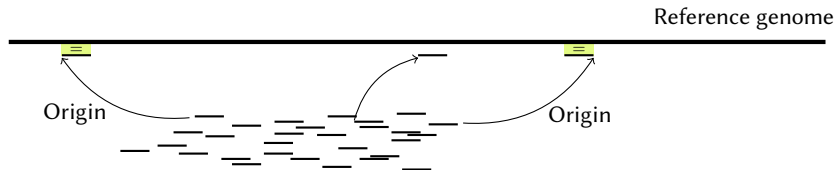
RNA-Seq read mapping



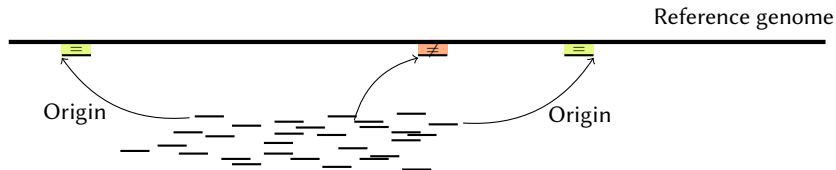
RNA-Seq read mapping



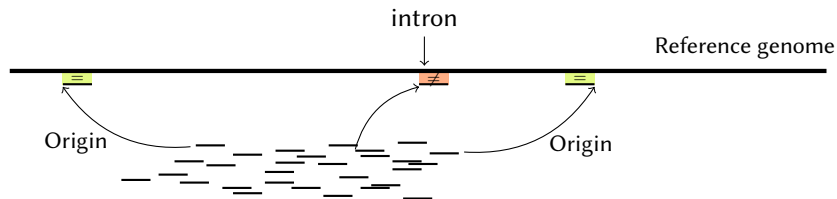
RNA-Seq read mapping



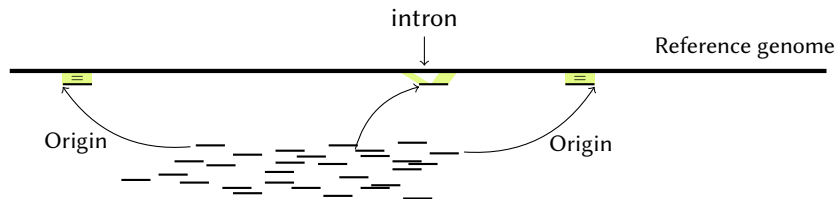
RNA-Seq read mapping



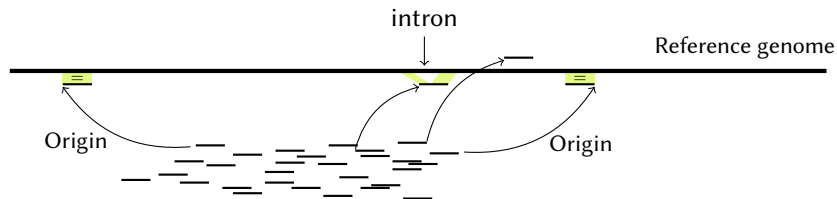
RNA-Seq read mapping



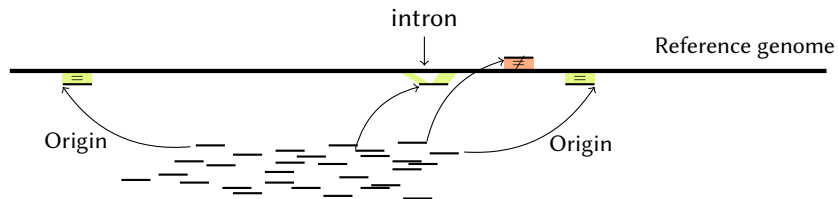
RNA-Seq read mapping



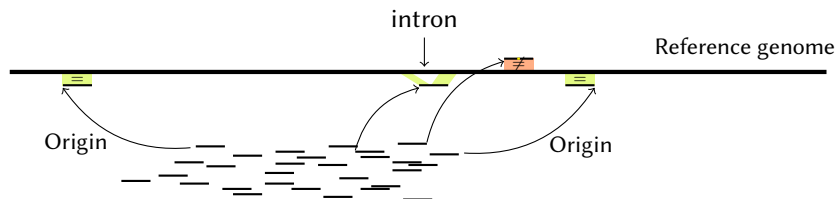
RNA-Seq read mapping



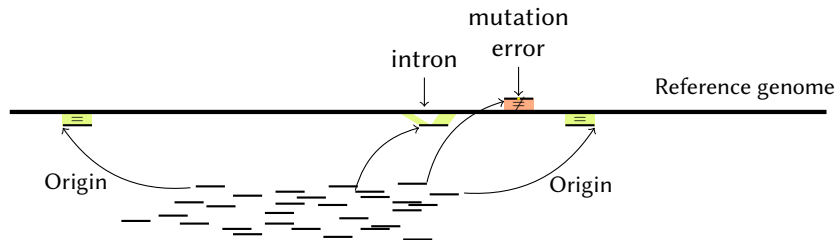
RNA-Seq read mapping



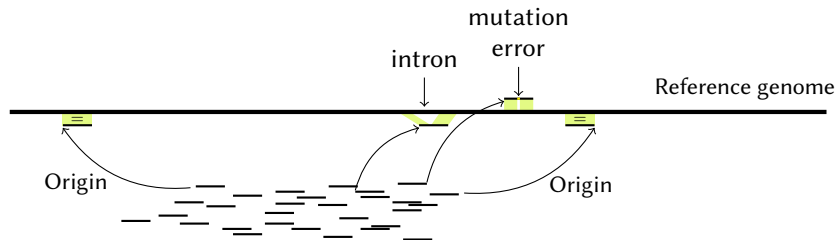
RNA-Seq read mapping



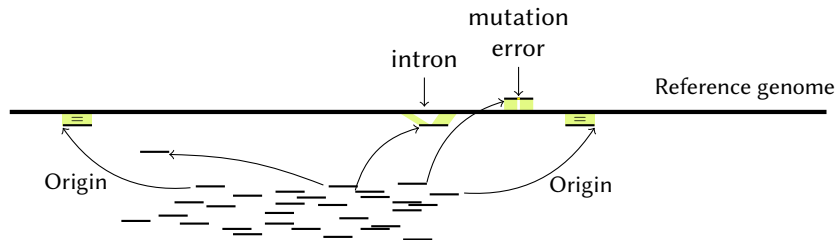
RNA-Seq read mapping



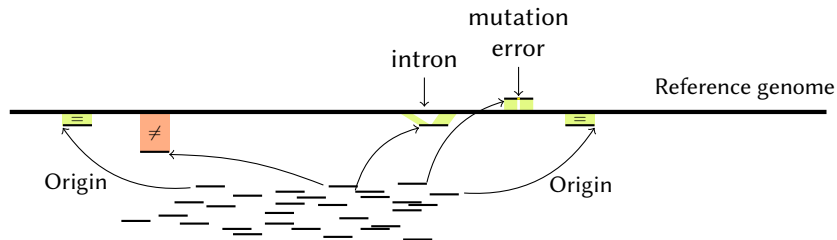
RNA-Seq read mapping



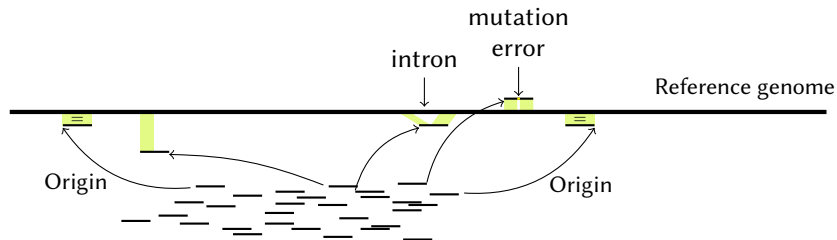
RNA-Seq read mapping



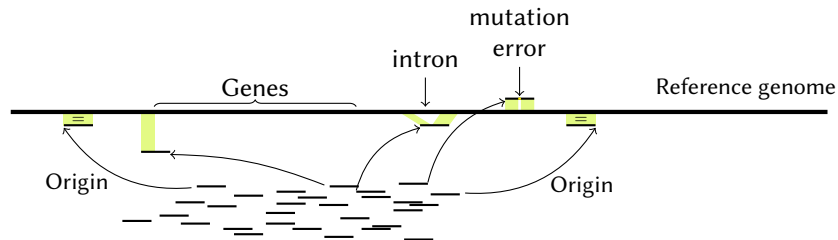
RNA-Seq read mapping



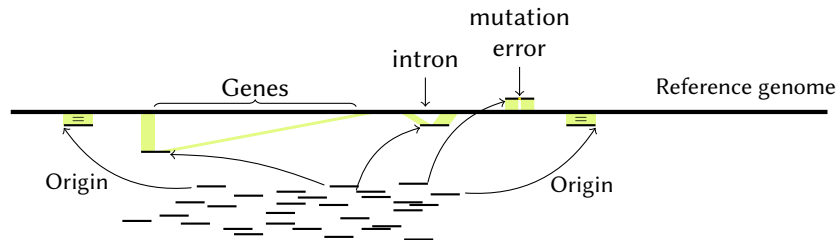
RNA-Seq read mapping



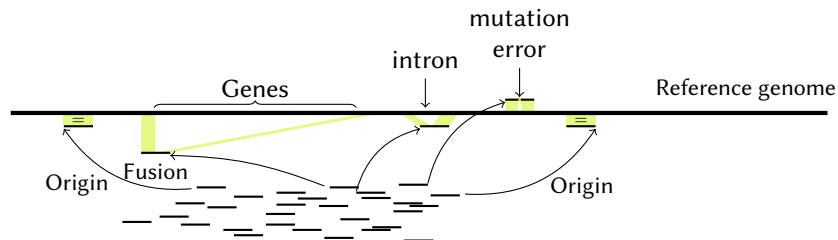
RNA-Seq read mapping



RNA-Seq read mapping



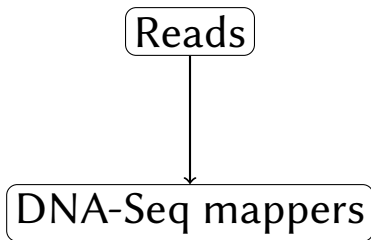
RNA-Seq read mapping



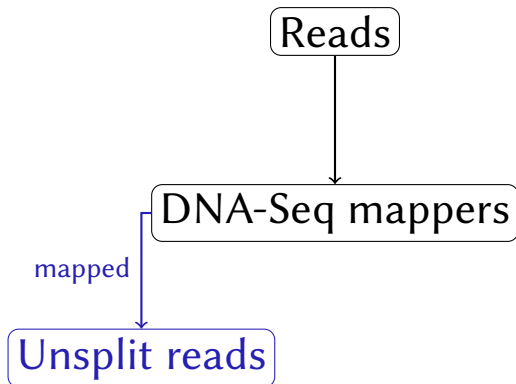
Split reads don't align contiguously to the genome

Reads

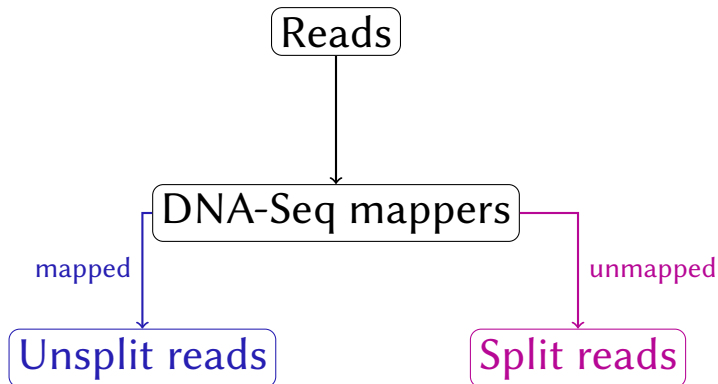
Split reads don't align contiguously to the genome



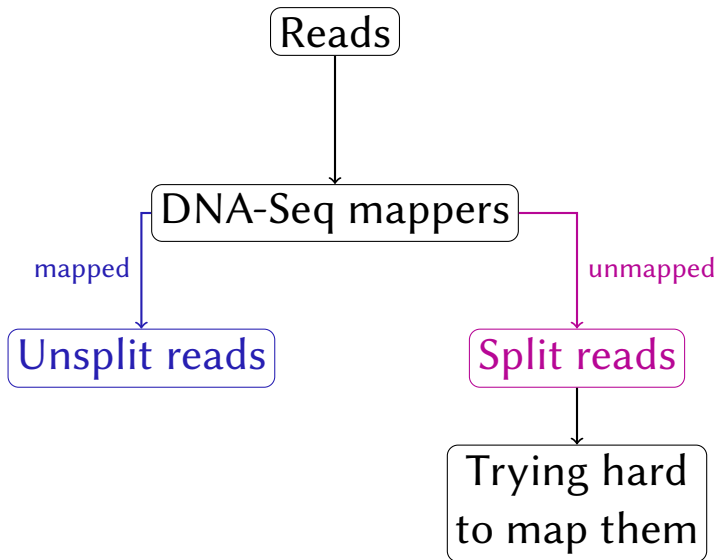
Split reads don't align contiguously to the genome



Split reads don't align contiguously to the genome



Split reads don't align contiguously to the genome



Mapping split reads by... splitting them – TopHat2

(2) Genome alignment

Reads spanning a single exon are **mapped**



Multi-exon spanning reads are **unmapped**

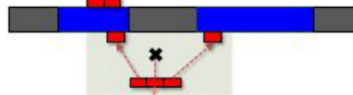


(3) Spliced alignment

Reads are split into segments

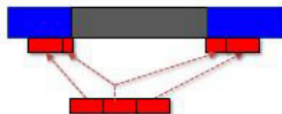


Unmapped segment

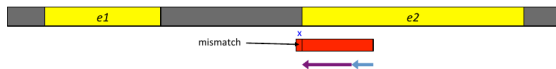


(3-1) Segment alignment to genome

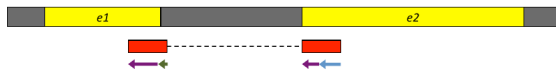
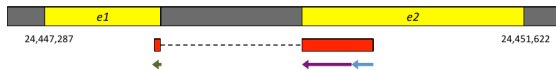
(3-2) Identification of splice sites
(including indels and fusion break points)



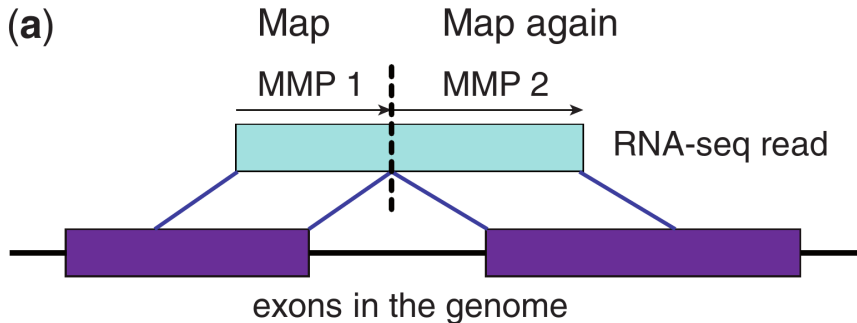
Mapping all reads by splitting them – HISAT2, STAR



Local FM index for chr22 from 24,417,280 to 24,482,559



Mapping all reads by splitting them – HISAT2, STAR



Mapping methods

TopHat2	Exact contiguous fixed-length
HISAT	Maximal mappable suffix
STAR	Maximal mappable prefix

Mapping methods

TopHat2	Exact contiguous fixed-length
HISAT	Maximal mappable suffix
STAR	Maximal mappable prefix

Indexing methods

TopHat2	FM-index
HISAT	Multiple FM-indices
STAR	Suffix Array

Indexing methods

$T = \overset{0}{C} \overset{1}{T} \overset{2}{A} \overset{3}{G} \overset{4}{T} \overset{5}{T} \overset{6}{A} \overset{7}{G} \overset{8}{\$}$

Indexing methods

0 1 2 3 4 5 6 7 8
 $T = \text{CTAGTTAG\$}$

TS	8	6	2	0	7	3	5	1	4
	\$	A	A	C	G	G	T	T	T
		G	G	T	\$	T	A	A	T
		\$	T	A		T	G	G	A
			T	G		A	\$	T	G
			A	T		G		T	\$
			G	T		\$		A	
			\$	A				G	
				G				\$	

Indexing methods

0 1 2 3 4 5 6 7 8
 $T = \text{CTAGTTAG\$}$

TS	8	6	2	0	7	3	5	1	4
	\$	A	A	C	G	G	T	T	T
	C	G	G	T	\$	T	A	A	T
	T	\$	T	A	C	T	G	G	A
	A	C	T	G	T	A	\$	T	G
	G	T	A	T	A	G	C	T	\$
	T	A	G	T	G	\$	T	A	C
	T	G	\$	A	T	C	A	G	T
	A	T	C	G	T	T	G	\$	A
	G	T	T	\$	A	A	T	C	G

Indexing methods

0 1 2 3 4 5 6 7 8
 $T = \text{CTAGTTAG\$}$

TS	8	6	2	0	7	3	5	1	4
	\$	A	A	C	G	G	T	T	T
	C	G	G	T	\$	T	A	A	T
	T	\$	T	A	C	T	G	G	A
	A	C	T	G	T	A	\$	T	G
	G	T	A	T	A	G	C	T	\$
	T	A	G	T	G	\$	T	A	C
	T	G	\$	A	T	C	A	G	T
	A	T	C	G	T	T	G	\$	A
	G	T	T	\$	A	A	T	C	G

Burrows-Wheeler Transform

What approach is the best? (slide courtesy of J. Audoux)

NATURE METHODS | ANALYSIS

Simulation-based comprehensive benchmarking of RNA-seq aligners

Giacomo Baruzzo, Katharina E Hayer, Eun Ji Kim, Barbara Di Camillo, Garret A FitzGerald & Gregory R Grant

METHOD | OPEN ACCESS

A benchmark for RNA-seq quantification pipelines

Mingxiang Teng, Michael I. Love, Carrie A. Davis, Sarah Djebali, Alexander Dobin, Brenton R. Graveley, Sheng Li, Christopher E. Mason, Sara Olson, Dmitri Pervouchine, Cricket A. Sloan, Xintao Wei, Lijun Zhan and Rafael A. Irizarry

NATURE METHODS | ANALYSIS OPEN

Systematic evaluation of spliced alignment programs for RNA-seq data

Pär G Engström, Tamara Steijger, Botond Sipos, Gregory R Grant, RGASP Consortium, Gunnar Rättsch, Nick Goldman, Tim J Hubbard, Roderic Guigó & Paul Bertone

Article | OPEN

Comparative assessment of methods for the fusion transcripts detection from RNA-Seq data

Am J Hum Genet. 2013 Oct 3; 93(4): 641-651.
doi: [10.1016/j.ajhg.2013.08.008](https://doi.org/10.1016/j.ajhg.2013.08.008)

PMCID: PMC3

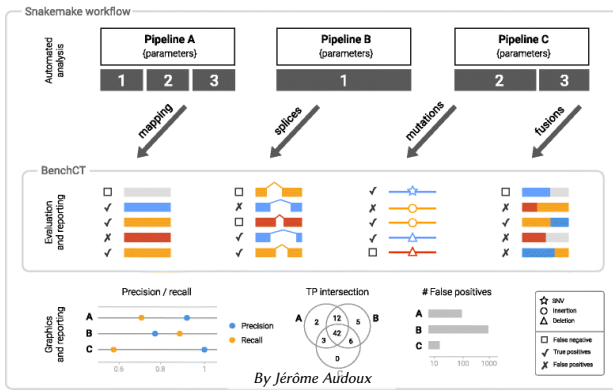
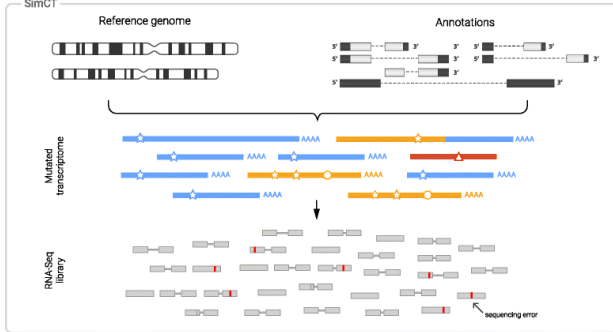
Reliable Identification of Genomic Variants from RNA-Seq Data

Robert Piskol,¹ Gokul Ramaswami,¹ and Jin Billy Li^{1,*}

[Author information](#) ▶ [Article notes](#) ▶ [Copyright and License information](#) ▶

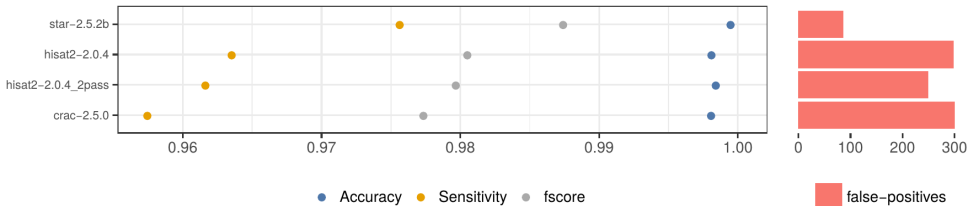
Benchmarking RNA-Seq data

Audoux *et al*, BMC Bioinformatics, 2017



Sensitivity/accuracy of read mappers

160M 150bp reads from GRCh38

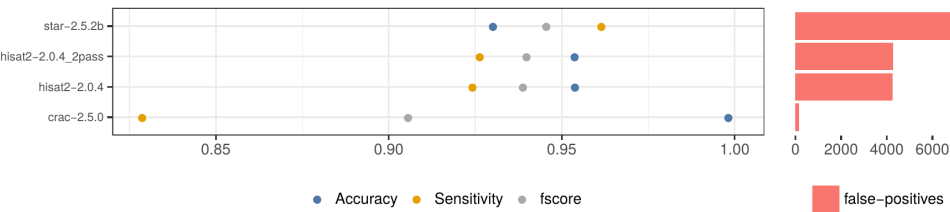


By Jérôme Audoux

STAR offers the best trade-off for splice detection

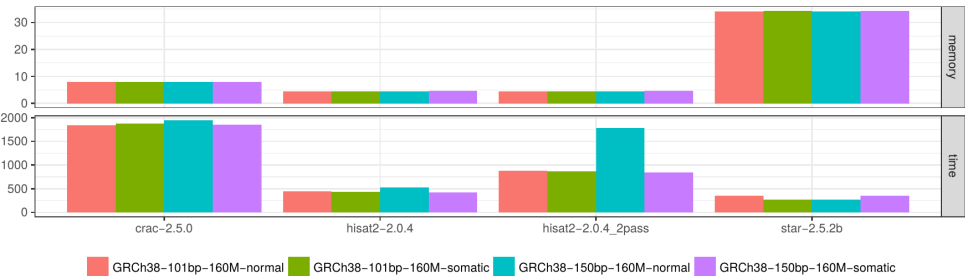
Splicing

160M 150bp reads from GRCh38



By Jérôme Audoux

Space/time for read mappers



By Jérôme Audoux

Many people uses TopHat2

(> 6,500 citations in Scholar, > 800 citations in 2019 only)

Many people uses TopHat2

(> 6,500 citations in Scholar, > 800 citations in 2019 only)

but don't

Many people uses TopHat2

(> 6,500 citations in Scholar, > 800 citations in 2019 only)

but don't

On TopHat2 website (since Feb 2016) [↗](#)

TopHat2 « *is now largely superseded by HISAT2 which provides the same core functionality (i.e. spliced alignment of RNA-Seq reads), in a more accurate and **much more efficient** way* » .

Do you really need to map reads?

Does it matter to have a base pair precision location for hundreds of millions of reads?

Alignment-free RNA-seq quantification

Quantifying transcripts does not require alignment

Kallisto

Bray et al, Nat. Biotechnology, 2016 [↗](#)

Salmon

Patro et al, Nat. Methods, 2017 [↗](#)

Alignment-free RNA-seq quantification

Quantifying transcripts does not require alignment

Kallisto

Bray et al, Nat. Biotechnology, 2016 [↗](#)

Salmon

Patro et al, Nat. Methods, 2017 [↗](#)

Two orders of magnitude faster than TopHat+Cufflinks

How to quantify without aligning?

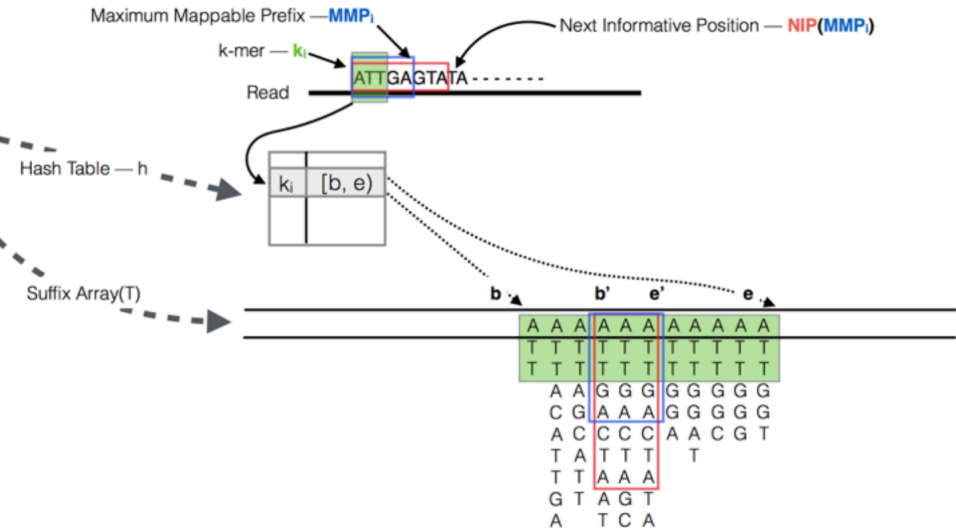
Transcripts

Read



Rest of the orange exon is *uninformative* — this junction is the *next informative position*.

How to quantify without aligning?



Ultra fast methods with good results...

Ultra fast methods with good results...

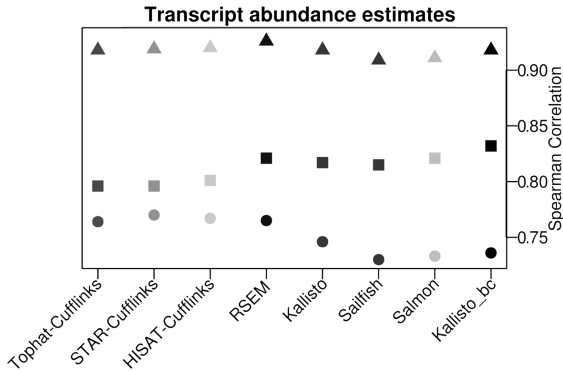
« With the exception of the underperforming Flux Capacitor and eXpress, we found that the other algorithms performed similarly. »

Teng et al, Genome Biology, 2016 [↗](#)

Ultra fast methods with good results...

« With the exception of the underperforming Flux Capacitor and eXpress, we found that the other algorithms performed similarly. »

Teng et al, Genome Biology, 2016 [↗](#)



Germain et al, Nucleic Acid Research, 2016 [↗](#)

Ultra fast methods with good results...

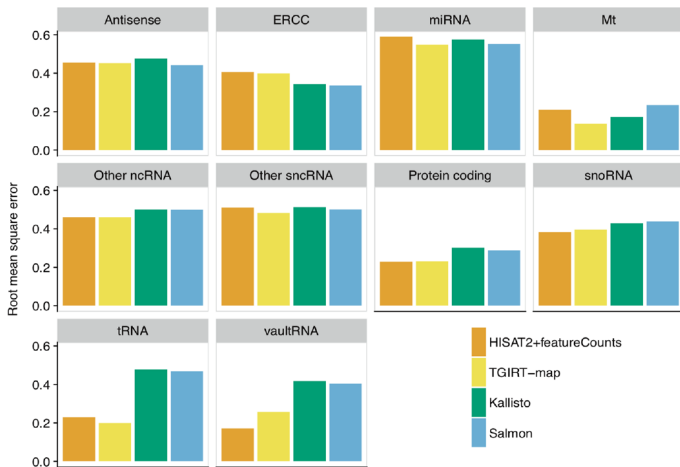
« With the exception of the underperforming Flux Capacitor and eXpress, we found that the other algorithms performed similarly. »

Teng et al, Genome Biology, 2016 [↗](#)

« It is particularly noteworthy that Salmon, which (like Sailfish and Kallisto) bypasses traditional alignment and thereby quantifies a single sample in a matter of minutes, had a comparable performance to Cufflinks and RSEM. Importantly, we confirmed these results using a variety of assays on both empirical and simulated data. »

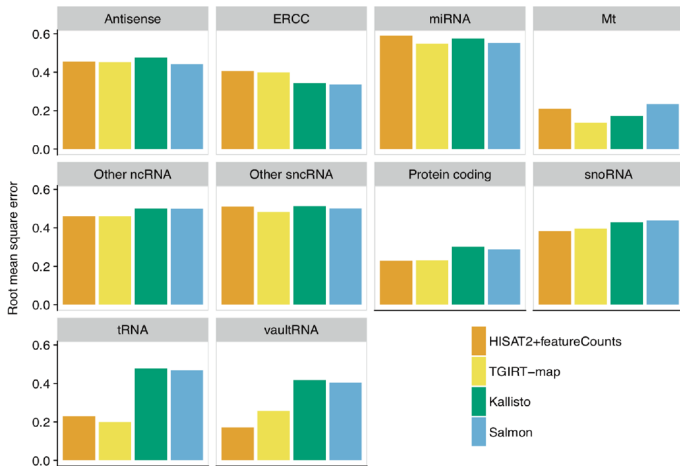
Germain et al, Nucleic Acid Research, 2016 [↗](#)

Good results that depend on the type of RNAs



CC BY Wu et al, 2018

Good results that depend on the type of RNAs



CC BY Wu et al, 2018

« We have found that alignment-based tools were more accurate in quantifying lowly-expressed or small genes. »

Wu et al, BMC Genomics, 2018

Up-to-date RNA-Seq analyses

High number of citations \neq Best software

High number of citations \neq Best software

Alignment isn't an end in itself

High number of citations \neq Best software

Alignment isn't an end in itself

Alignment-free methods may be suitable for you