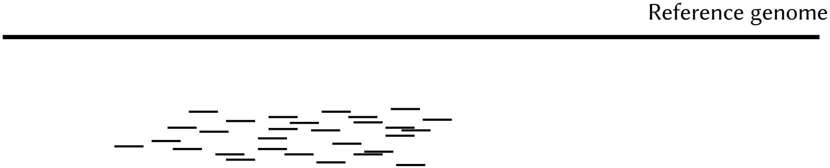
To map or not to map?

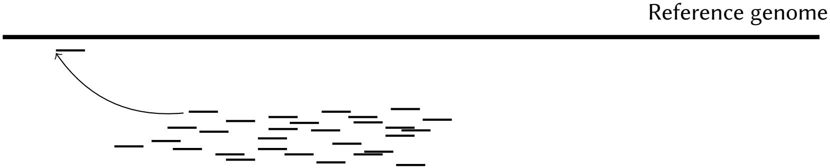
Formation RNA-Seq – Bilille

Mikaël Salson
mikael.salson@univ-lille.fr

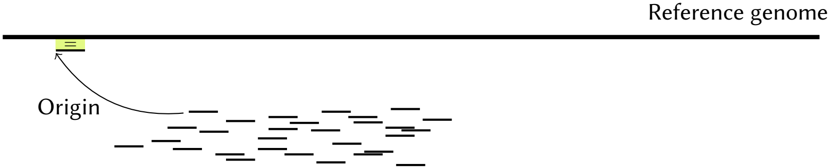
RNA-Seq read mapping



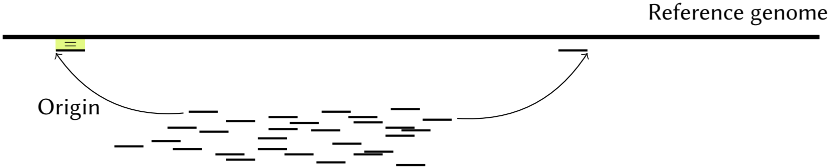
RNA-Seq read mapping



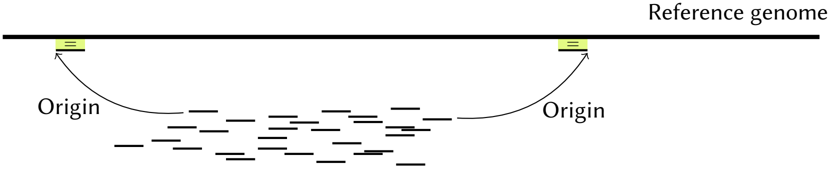
RNA-Seq read mapping



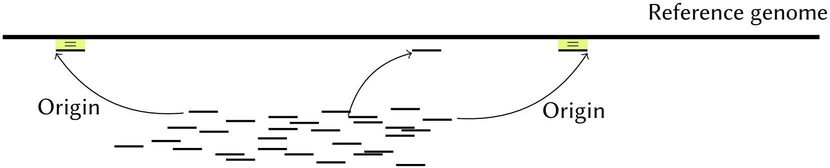
RNA-Seq read mapping



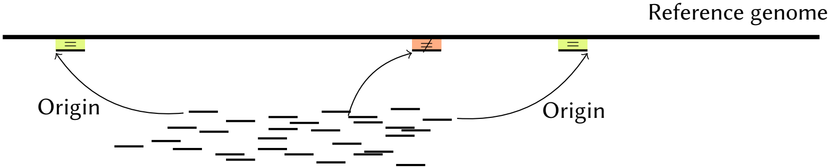
RNA-Seq read mapping



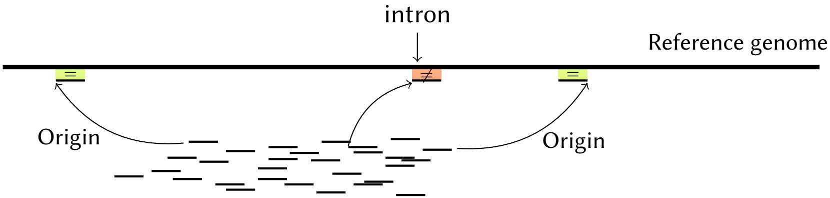
RNA-Seq read mapping



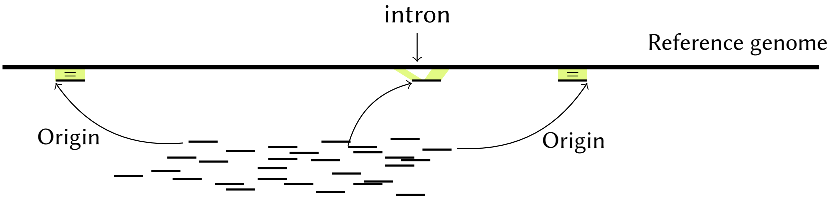
RNA-Seq read mapping



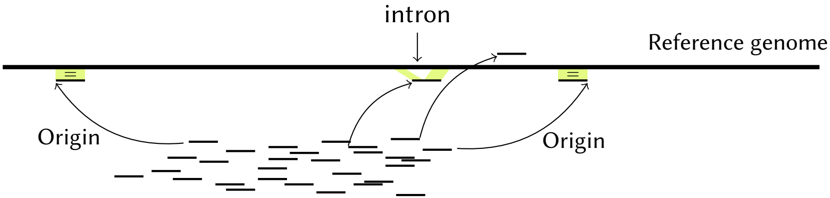
RNA-Seq read mapping



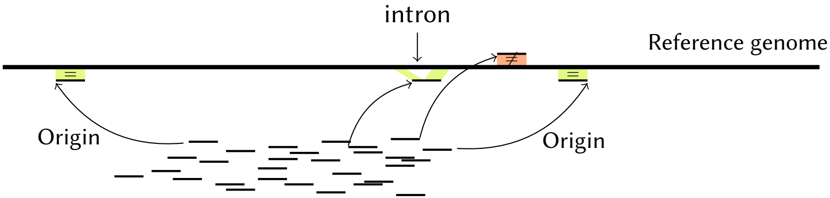
RNA-Seq read mapping



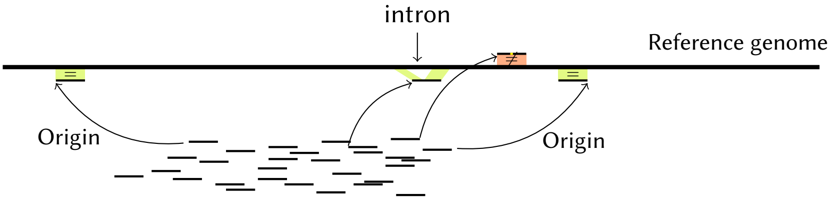
RNA-Seq read mapping



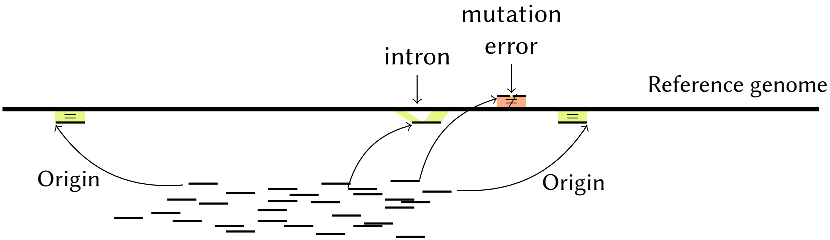
RNA-Seq read mapping



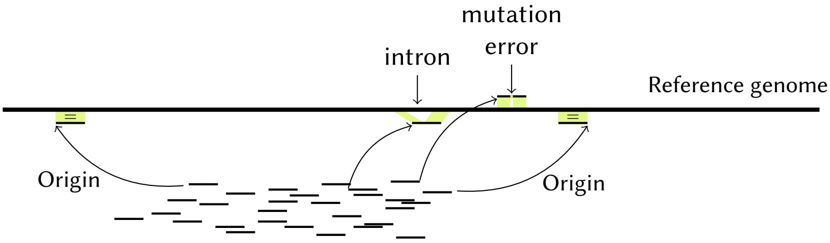
RNA-Seq read mapping



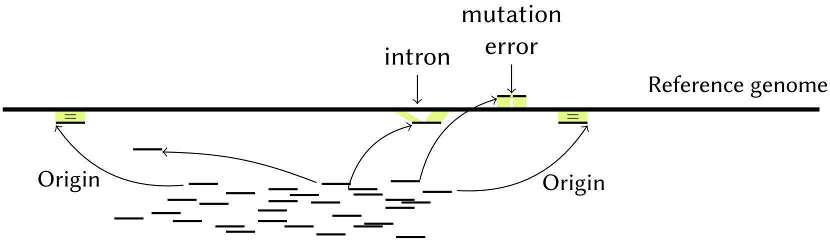
RNA-Seq read mapping



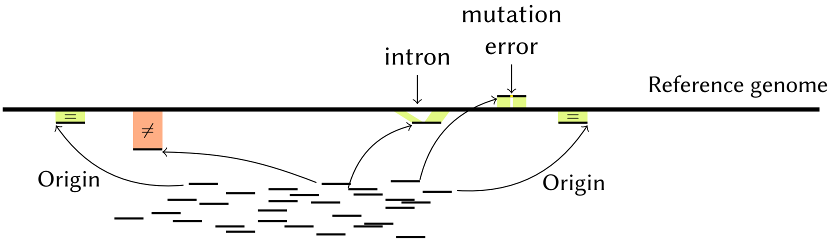
RNA-Seq read mapping



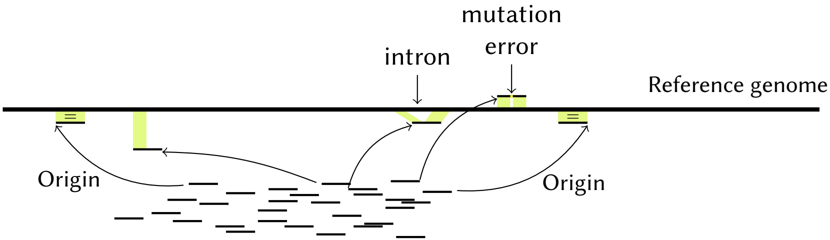
RNA-Seq read mapping



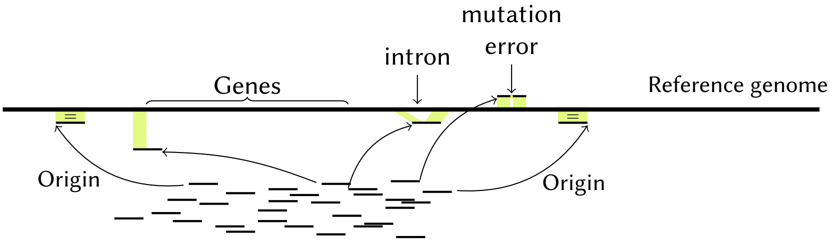
RNA-Seq read mapping



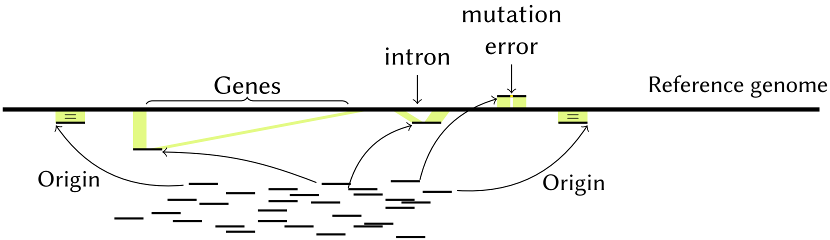
RNA-Seq read mapping



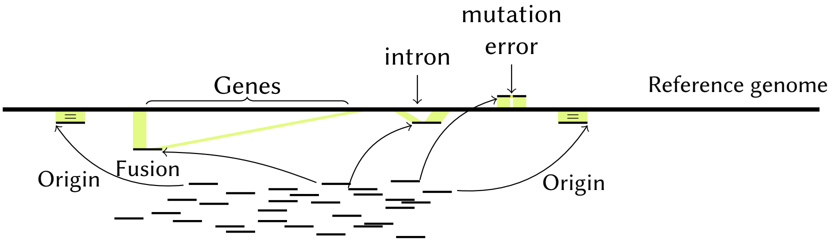
RNA-Seq read mapping



RNA-Seq read mapping



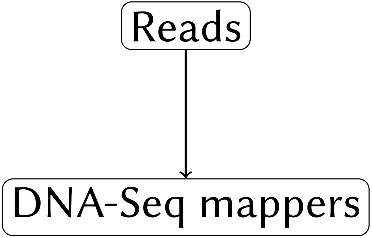
RNA-Seq read mapping



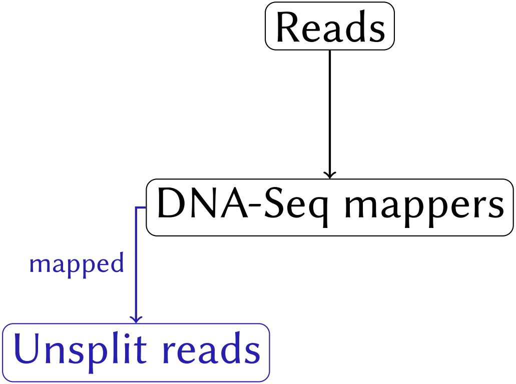
Split reads don’t align contiguously to the genome

Reads

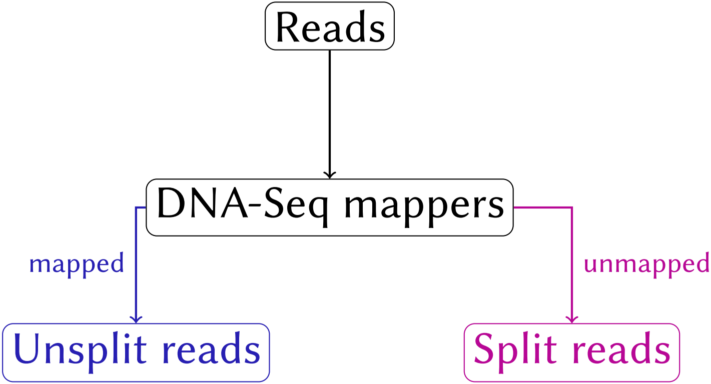
Split reads don't align contiguously to the genome



Split reads don’t align contiguously to the genome



Split reads don’t align contiguously to the genome



Reads

DNA-Seq mappers

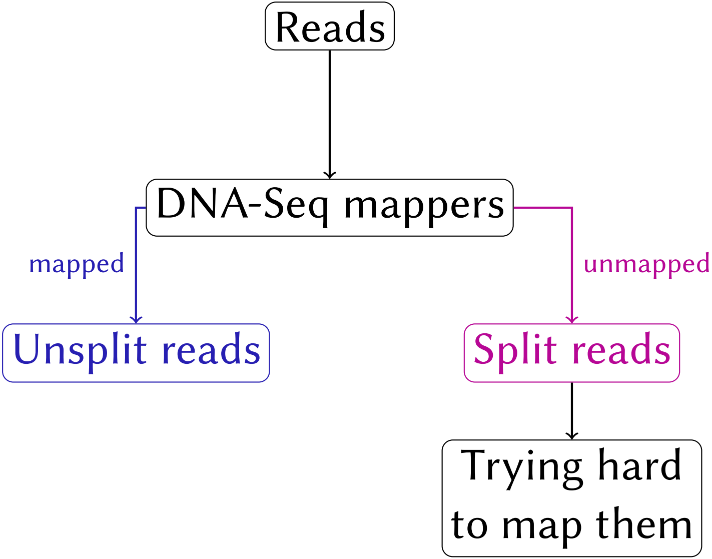
mapped

unmapped

Unsplit reads

Split reads

Split reads don’t align contiguously to the genome



Reads

DNA-Seq mappers

mapped

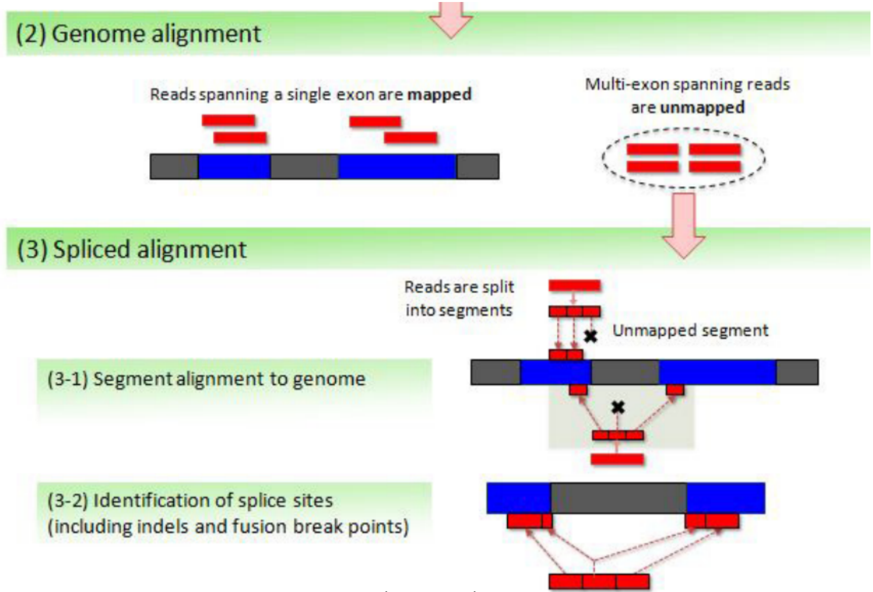
unmapped

Unsplit reads

Split reads

Trying hard
to map them

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

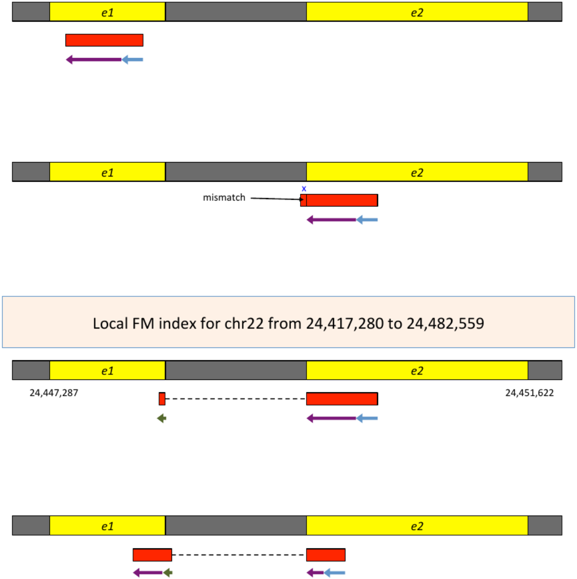
Unmapped segment

(3-1) Segment alignment to genome

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

© *Tophat2: Kim et al, 2013*

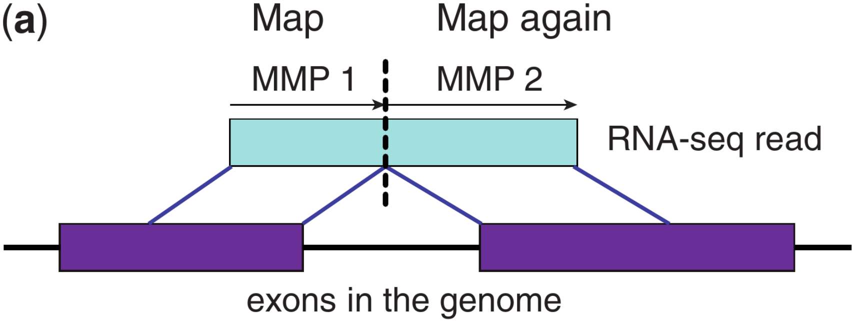
Mapping all reads by splitting them – HISAT2, STAR



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

© HISAT: Kim, Langmead, Salzberg, Nat. Methods, 2015

Mapping all reads by splitting them – HISAT2, STAR



© STAR: Dobin et al, Bioinformatics, 2013

Specificities of the approaches

Mapping methods

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

Specificities of the approaches

Mapping methods

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

Indexing methods

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

Indexing methods

$$T = \begin{array}{ccccccccc} \scriptstyle 0 & \scriptstyle 1 & \scriptstyle 2 & \scriptstyle 3 & \scriptstyle 4 & \scriptstyle 5 & \scriptstyle 6 & \scriptstyle 7 & \scriptstyle 8 \\ \mathrm{C} & \mathrm{T} & \mathrm{A} & \mathrm{G} & \mathrm{T} & \mathrm{T} & \mathrm{A} & \mathrm{G} & \$ \end{array}$$

Indexing methods

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

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

$$T = \begin{array}{ccccccccc} \scriptstyle 0 & \scriptstyle 1 & \scriptstyle 2 & \scriptstyle 3 & \scriptstyle 4 & \scriptstyle 5 & \scriptstyle 6 & \scriptstyle 7 & \scriptstyle 8 \\ \mathrm{C} & \mathrm{T} & \mathrm{A} & \mathrm{G} & \mathrm{T} & \mathrm{T} & \mathrm{A} & \mathrm{G} & \$ \end{array}$$

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

$$T = \begin{array}{ccccccccc} {\scriptstyle 0} & {\scriptstyle 1} & {\scriptstyle 2} & {\scriptstyle 3} & {\scriptstyle 4} & {\scriptstyle 5} & {\scriptstyle 6} & {\scriptstyle 7} & {\scriptstyle 8} \\ \mathrm{C} & \mathrm{T} & \mathrm{A} & \mathrm{G} & \mathrm{T} & \mathrm{T} & \mathrm{A} & \mathrm{G} & \$ \end{array}$$

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

k-**mer sets** - Burrows Wheeler transform?[1]

text

```
row_row_row_your_boat
row_row_row_your_boat
row_row_row_your_boat$
```

Burrows Wheeler transform (BWT)

```
trrrwwwwwwwwwooo___bbbyyyrrrrrrrruuutt$______aaaoooooooooooo___
```

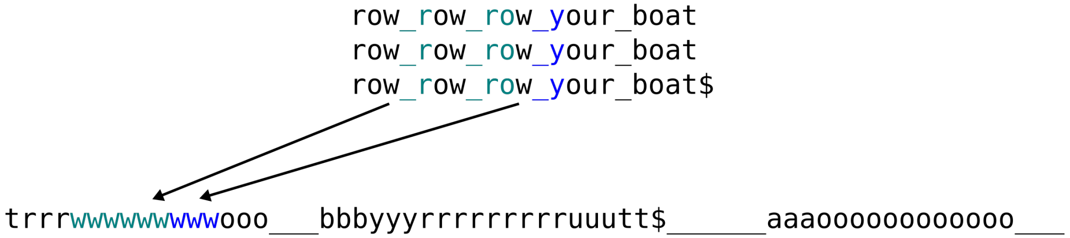
Compression through run length encoding

```
(t,1)(r,3)(w,9)(o,3) ... (_,3)
```

[1] Adapted from Ben Langmead's course

k-**mer sets** - Right contexts of w's

very similar right lexicographic contexts for all w's

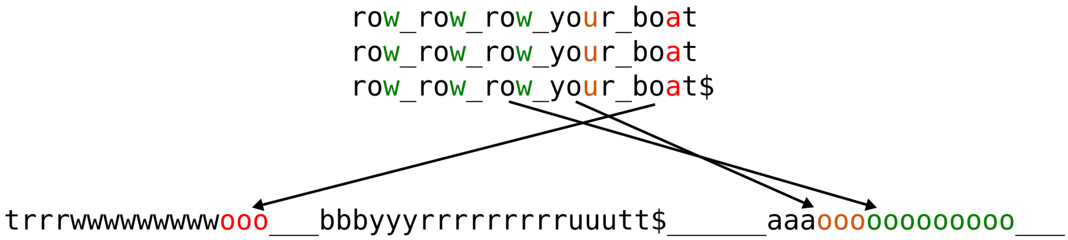


```
row_row_row_your_boat
row_row_row_your_boat
row_row_row_your_boat$
```

```
trrrwwwwwwwwwooo___bbbyyyrrrrrrrrruuutt$______aaaoooooooooooo___
```

k-**mer sets** - Right contexts of o's

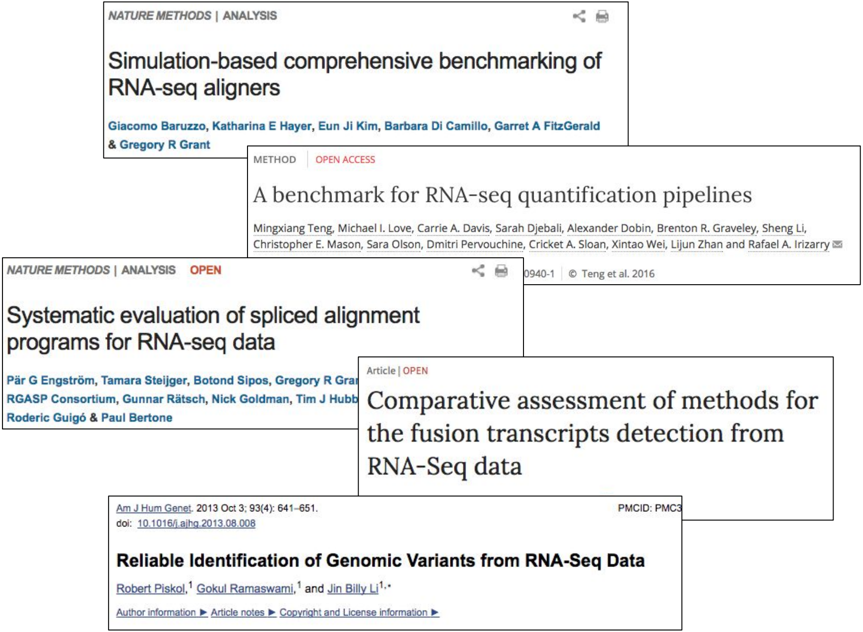
right lexicographic contexts for o's



row_row_row_your_boat
row_row_row_your_boat
row_row_row_your_boat$

trrrwwwwwwwwwooo___bbbyyyrrrrrrrrruuutt$______aaaooooooooooooo___

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 ✉

0940-1 | © Teng et al. 2016

NATURE METHODS | ANALYSIS OPEN

Systematic evaluation of spliced alignment programs for RNA-seq data

Pär G Engström, Tamara Steijger, Botond Sipos, Gregory R Gra
RGASP Consortium, Gunnar Rätsch, Nick Goldman, Tim J Hubb
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. PMCID: PMC3
doi: 10.1016/j.ajhg.2013.08.008

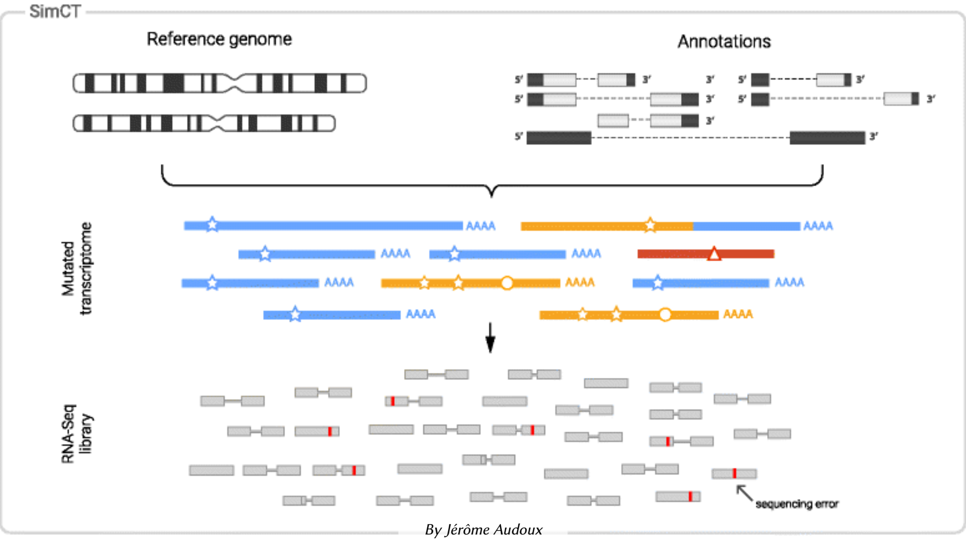
Reliable Identification of Genomic Variants from RNA-Seq Data

Robert Piskol,[1] Gokul Ramaswami,[1] and Jin Billy Li[1,*]

Author information ▶ Article notes ▶ Copyright and License information ▶

Benchmarking RNA-Seq aligners

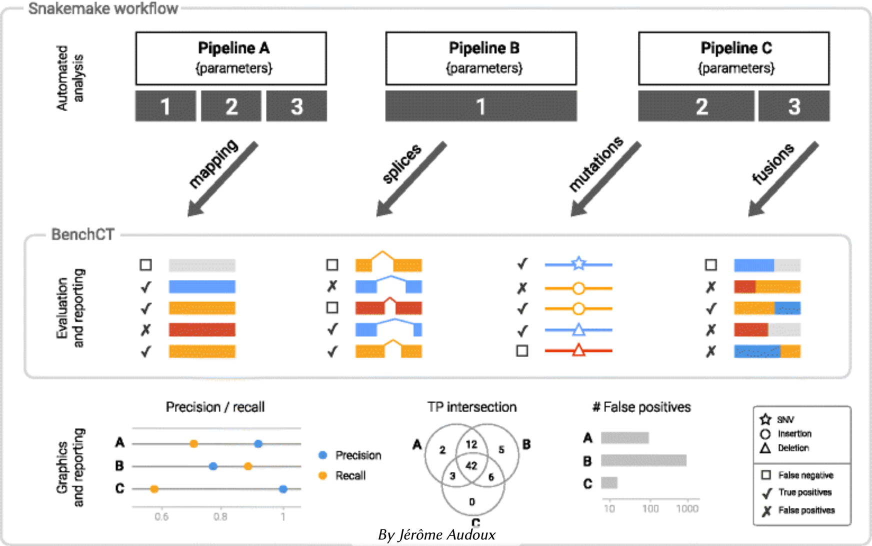
Audoux *et al*, BMC Bioinformatics, 2017



By Jérôme Audoux

Benchmarking RNA-Seq aligners

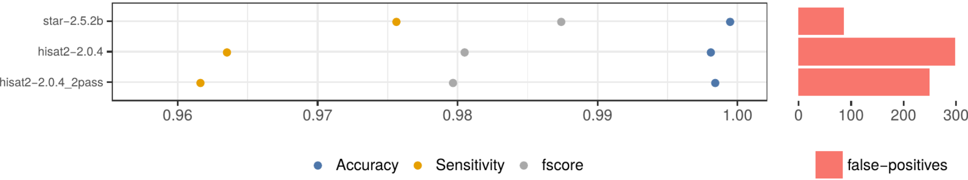
Audoux *et al*, BMC Bioinformatics, 2017



By Jérôme Audoux

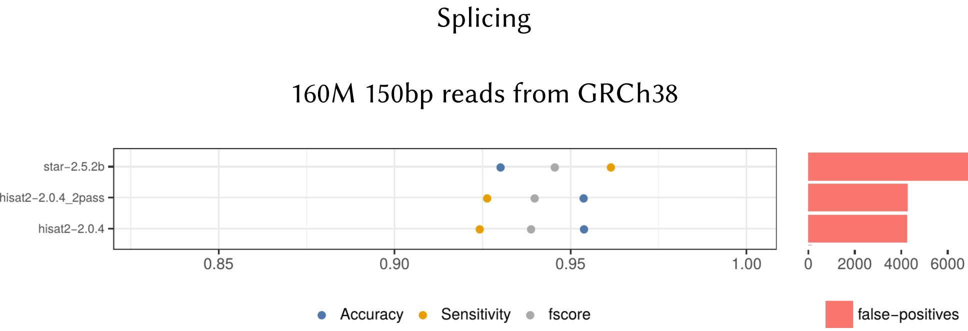
Sensitivity/accuracy of read mappers

160M 150bp reads from GRCh38



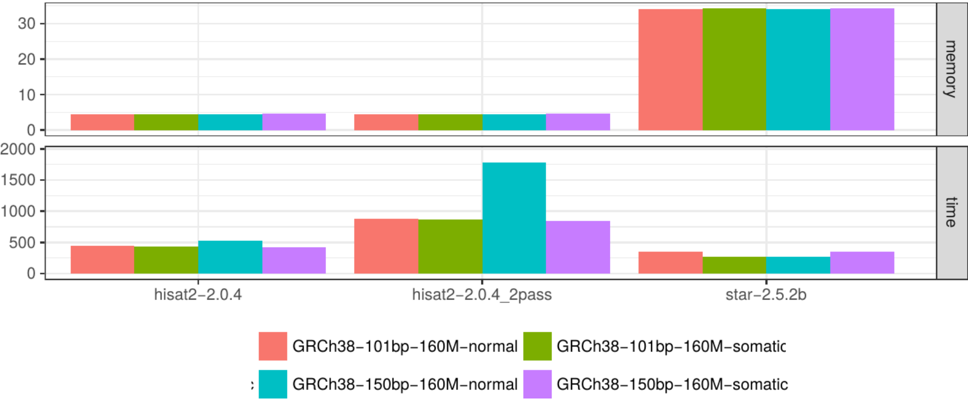
By Jérôme Audoux

STAR offers the best trade-off for splice detection



By Jérôme Audoux

Space/time for read mappers



By Jérôme Audoux

Many people uses TopHat2

($>$ 10,623 citations in Scholar, $> 1,000$ citations in 2021 only)

Many people uses TopHat2

($>$ 10,623 citations in Scholar, $> 1,000$ citations in 2021 only)

but don't

Many people uses TopHat2

($>$ 10,623 citations in Scholar, $> 1,000$ citations in 2021 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 may not require alignment

Kallisto

Bray et al, Nat. Biotechnology, 2016

Salmon

Patro et al, Nat. Methods, 2017

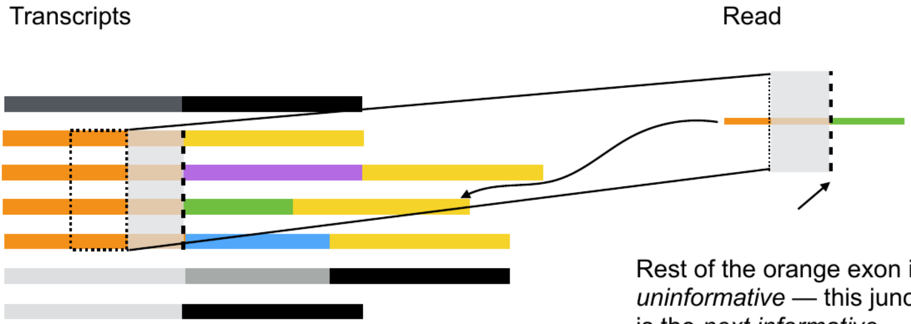
Alignment-free RNA-seq quantification

Quantifying transcripts may not require alignment

Kallisto	Salmon
Bray et al, Nat. Biotechnology, 2016	Patro et al, Nat. Methods, 2017

Two orders of magnitude faster than TopHat+Cufflinks

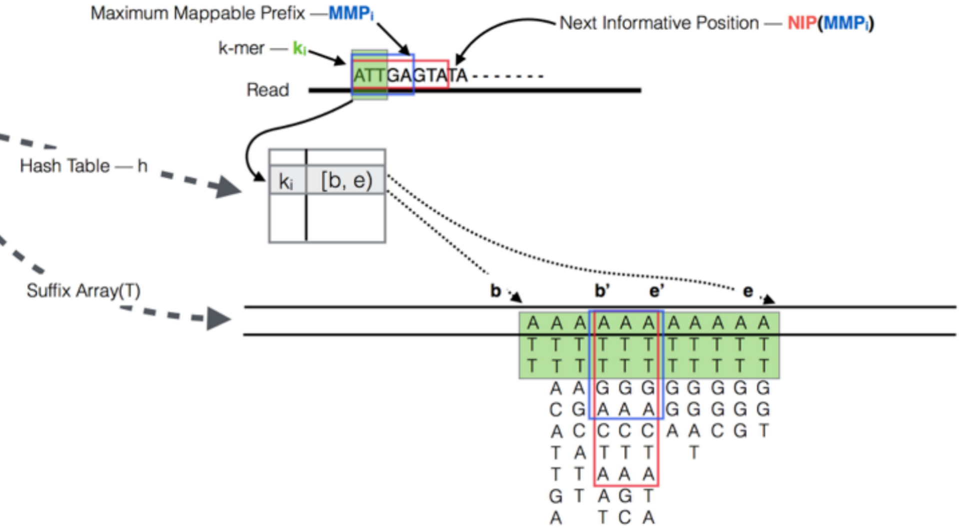
How to quantify without aligning?



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

© Rob Patro (Salmon)

How to quantify without aligning?



© Rob Patro (Salmon)

Ultra fast methods with good results…

Ultra fast methods with good results…

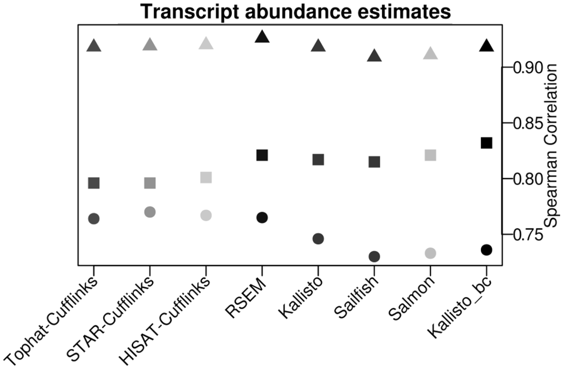
> « *With the exception of the underperforming Flux Capacitor and eXpress, we found that the other algorithms performed similarly.* »

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…

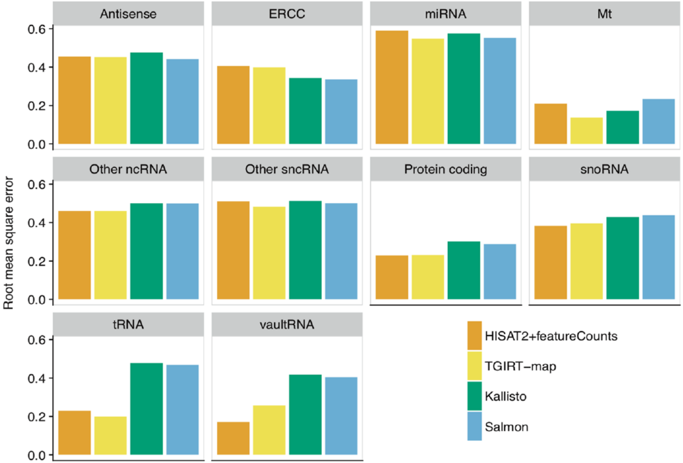
« *With the exception of the underperforming Flux Capacitor and eXpress, we found that the other algorithms performed similarly.* »

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.* »

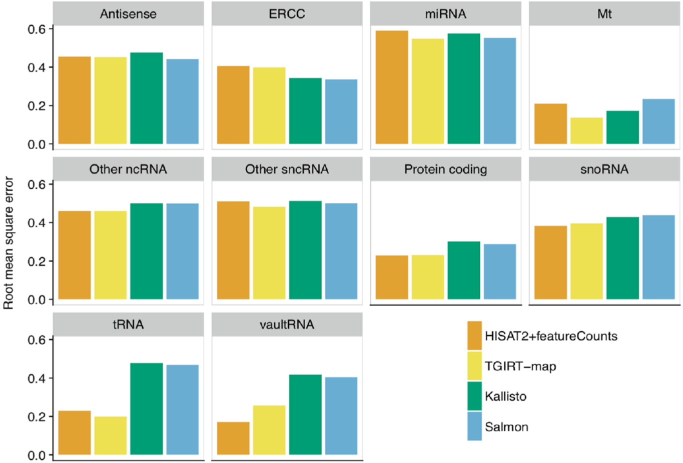
Germain et al, Nucleic Acid Research, 2016

Good performances may not hold true for all the data



CC BY *Wu et al, 2018*

Good performances may not hold true for all the data



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

Where the differences come from?

Where the differences come from?

1. Alignement vs pseudo-alignment

Where the differences come from?

1. Alignement vs pseudo-alignment

2. Genome reference vs transcriptome reference
   see Srivastava *et al*, 2020 ↗

Where the differences come from?

1. Alignement vs pseudo-alignment

2. Genome reference vs transcriptome reference
   see Srivastava *et al*, 2020

3. Quantification method

How to quantify multi-mapped reads?

When a read maps at multiple loci,
what transcript/gene should be counted?

See Deschamps-Francoeur *et al*, 2020 (thanks Pierre!)

How to quantify multi-mapped reads?

When a read maps at multiple loci,
what transcript/gene should be counted?

See Deschamps-Francoeur *et al*, 2020 ↗ (thanks Pierre!)

- None
  (*eg*. HTSeq-count, STAR genecount, featureCounts)

How to quantify multi-mapped reads?

When a read maps at multiple loci,
what transcript/gene should be counted?

See Deschamps-Francoeur *et al*, 2020 (thanks Pierre!)

- None
  (*eg.* HTSeq-count, STAR genecount, featureCounts)

- Split counts evenly
  (*eg.* Cufflinks, featureCounts (with an option))

How to quantify multi-mapped reads?

When a read maps at multiple loci,
what transcript/gene should be counted?

See Deschamps-Francoeur *et al*, 2020 ↗ (thanks Pierre!)

- None
  (*eg.* HTSeq-count, STAR genecount, featureCounts)

- Split counts evenly
  (*eg.* Cufflinks, featureCounts (with an option))

- Rescue based on single mapping reads
  (*eg.* Cufflinks (with an option))

How to quantify multi-mapped reads?

When a read maps at multiple loci,
what transcript/gene should be counted?

See Deschamps-Francoeur *et al*, 2020 ↗ (thanks Pierre!)

- None
  (*eg.* HTSeq-count, STAR genecount, featureCounts)
- Split counts evenly
  (*eg.* Cufflinks, featureCounts (with an option))
- Rescue based on single mapping reads
  (*eg.* Cufflinks (with an option))
- Expectation maximization
  (*eg.* RSEM, Salmon, Kallisto)

Up-to-date RNA-Seq analyses

High number of citations $\neq$ Best software

Up-to-date RNA-Seq analyses

High number of citations $\neq$ Best software

Alignment isn't an end in itself

Up-to-date RNA-Seq analyses

High number of citations $\neq$ Best software

Alignment isn’t an end in itself

Alignment-free methods may be suitable for you