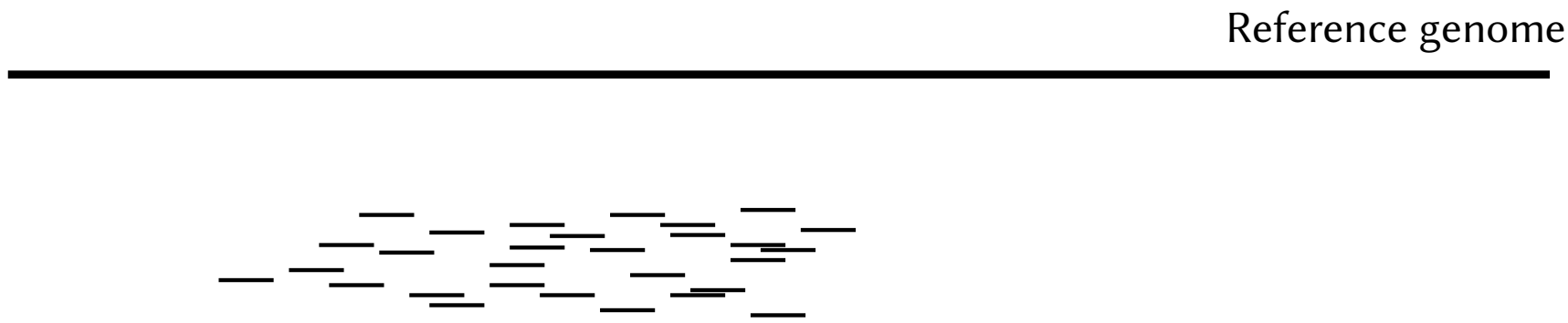
# To map or not to map?

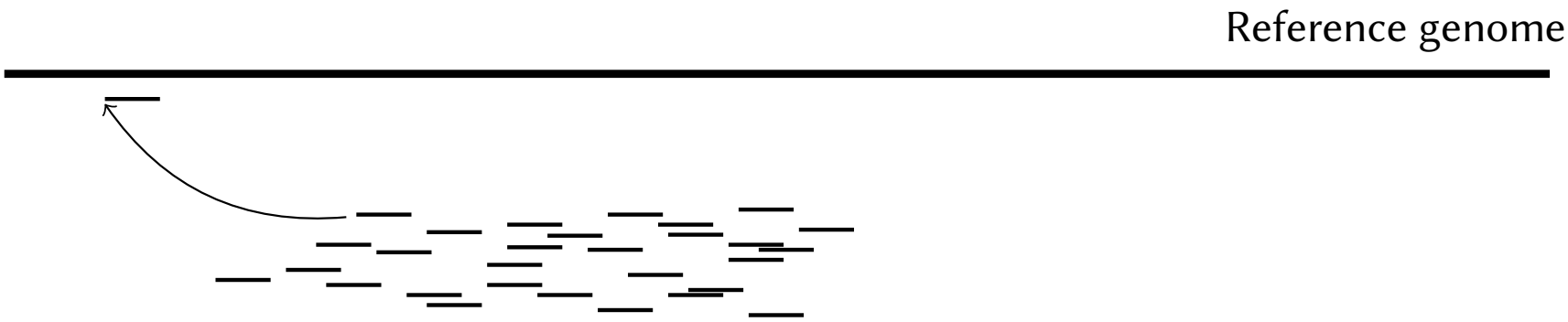
Formation RNA-Seq – Bilille

Mikaël Salson
mikael.salson@univ-lille.fr

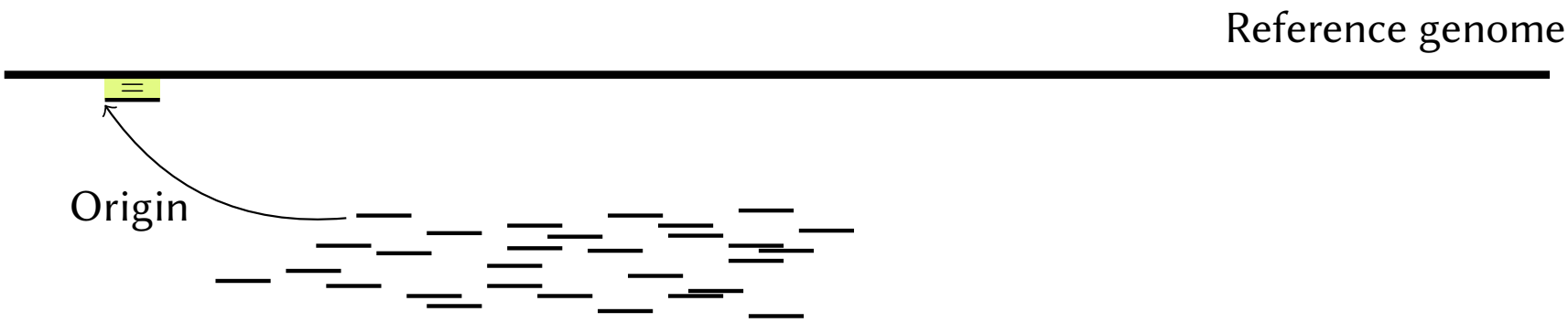
# RNA-Seq read mapping



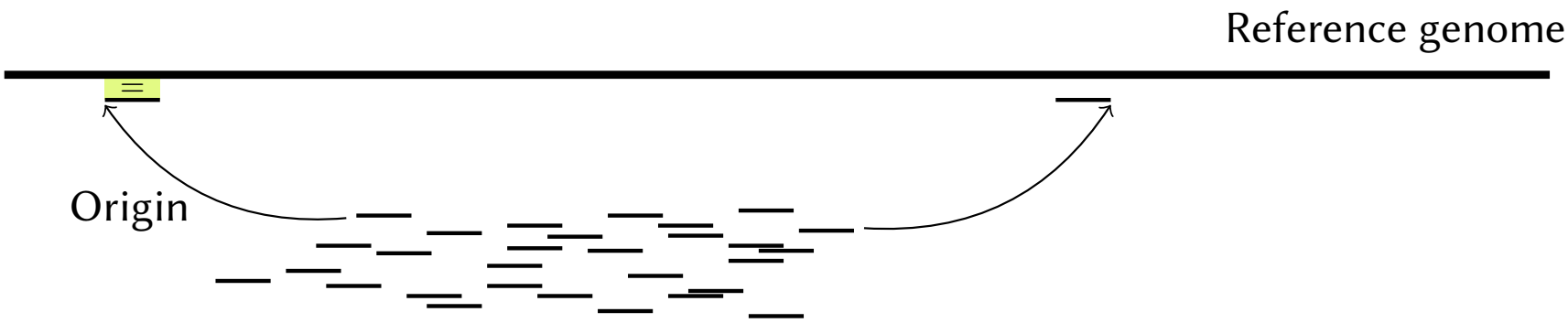
# RNA-Seq read mapping



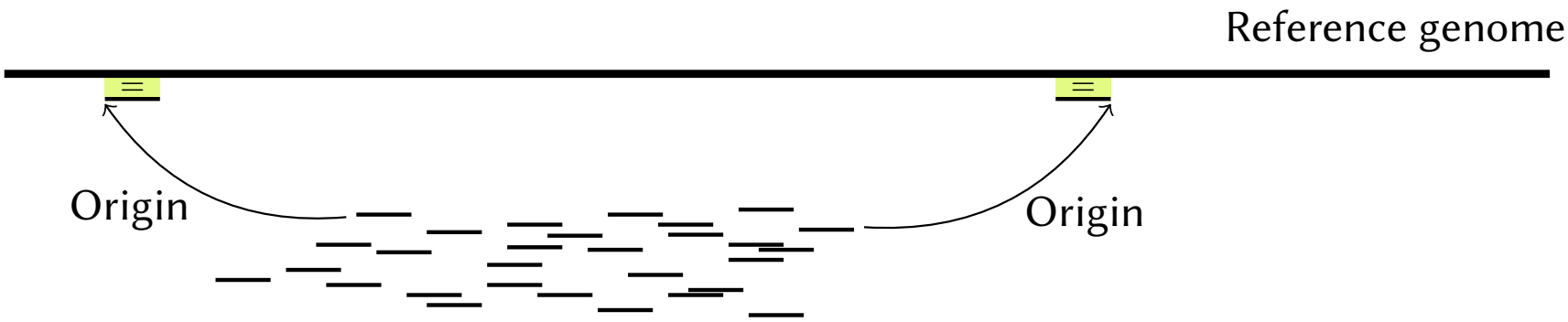
# RNA-Seq read mapping



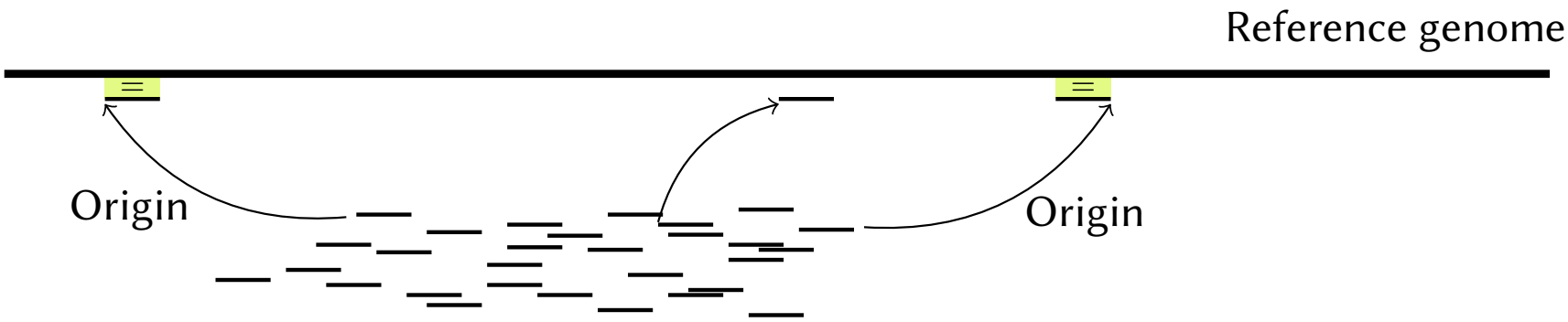
# RNA-Seq read mapping



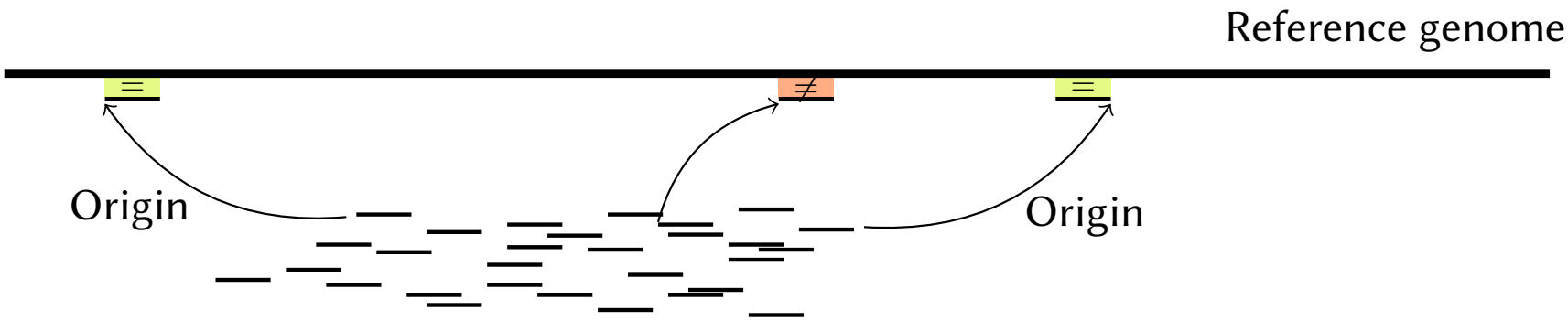
# RNA-Seq read mapping



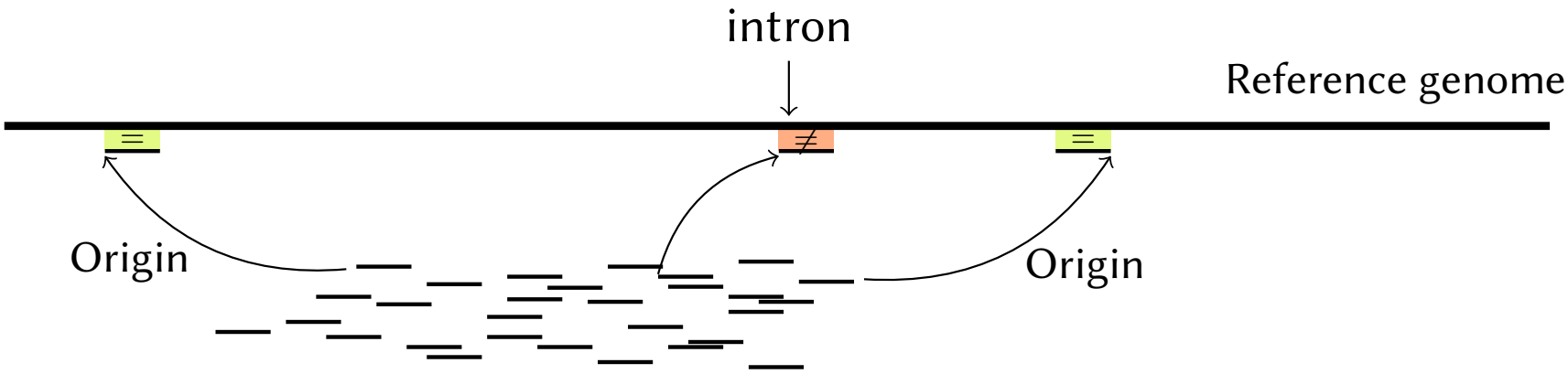
# RNA-Seq read mapping



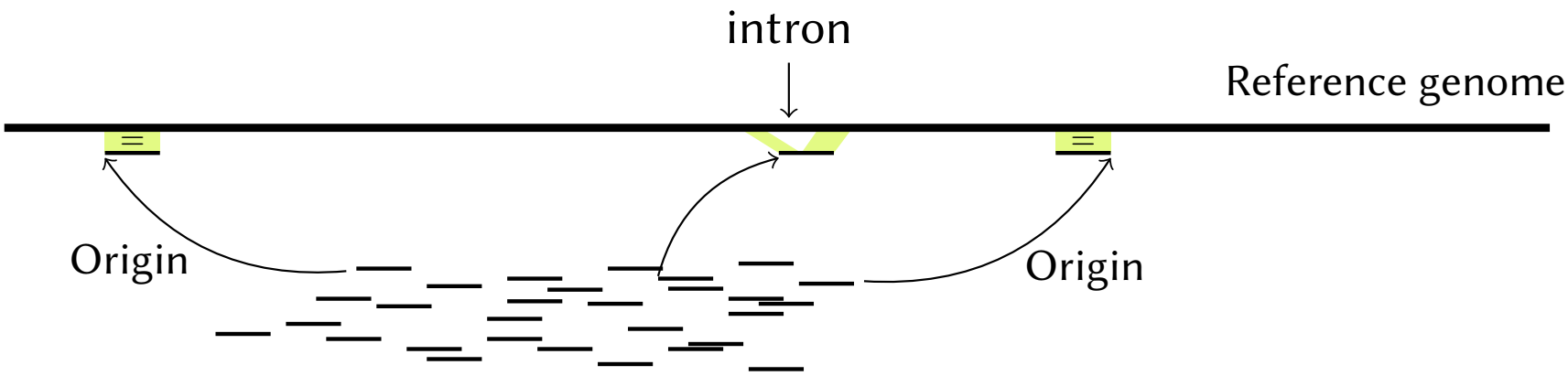
# RNA-Seq read mapping



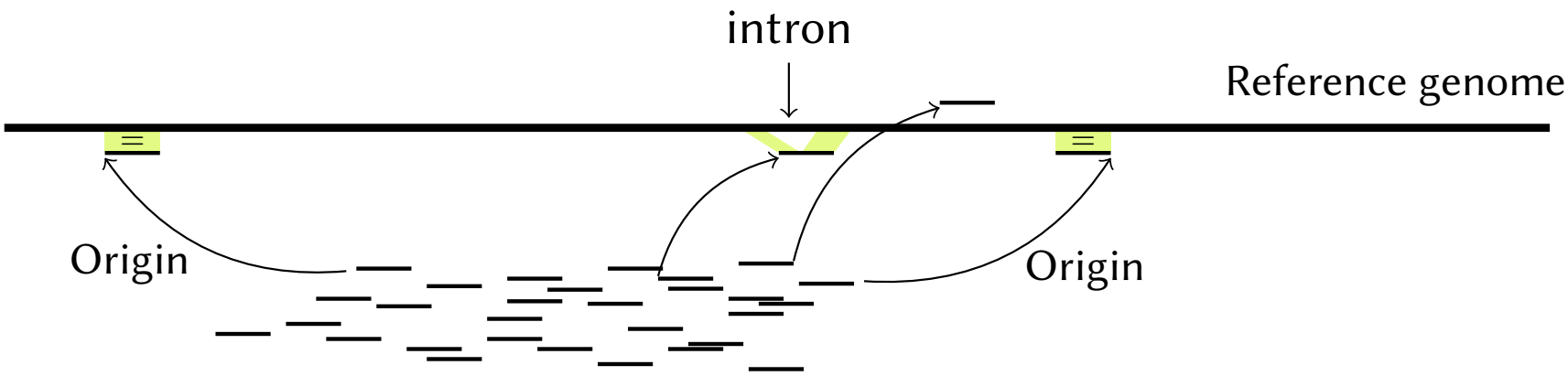
# RNA-Seq read mapping



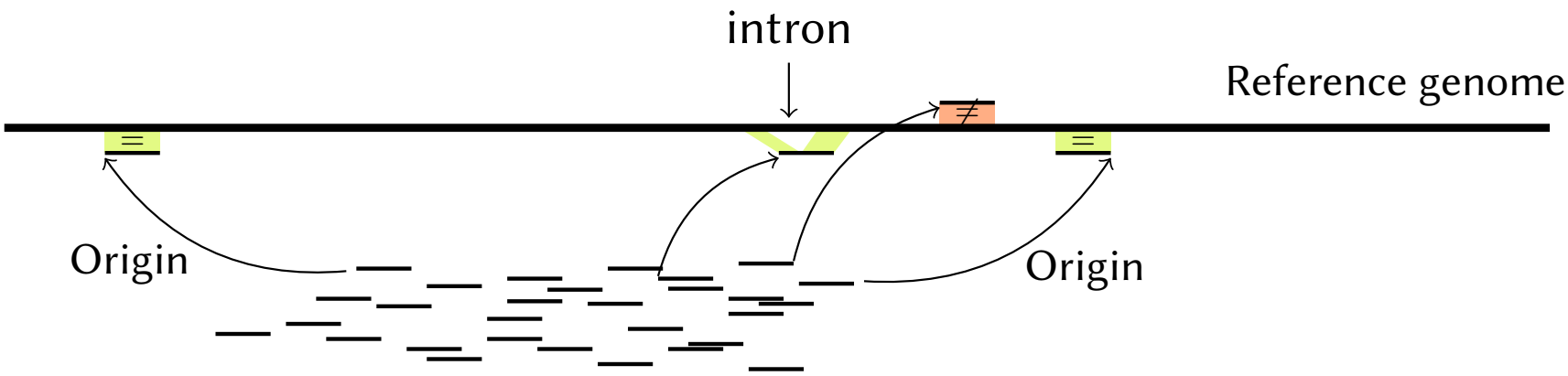
# RNA-Seq read mapping



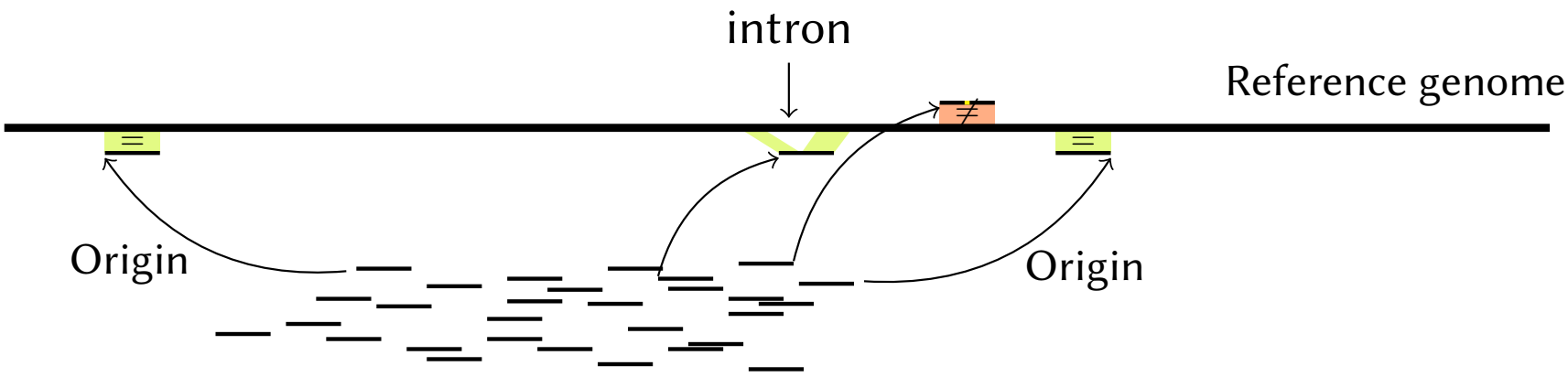
# RNA-Seq read mapping



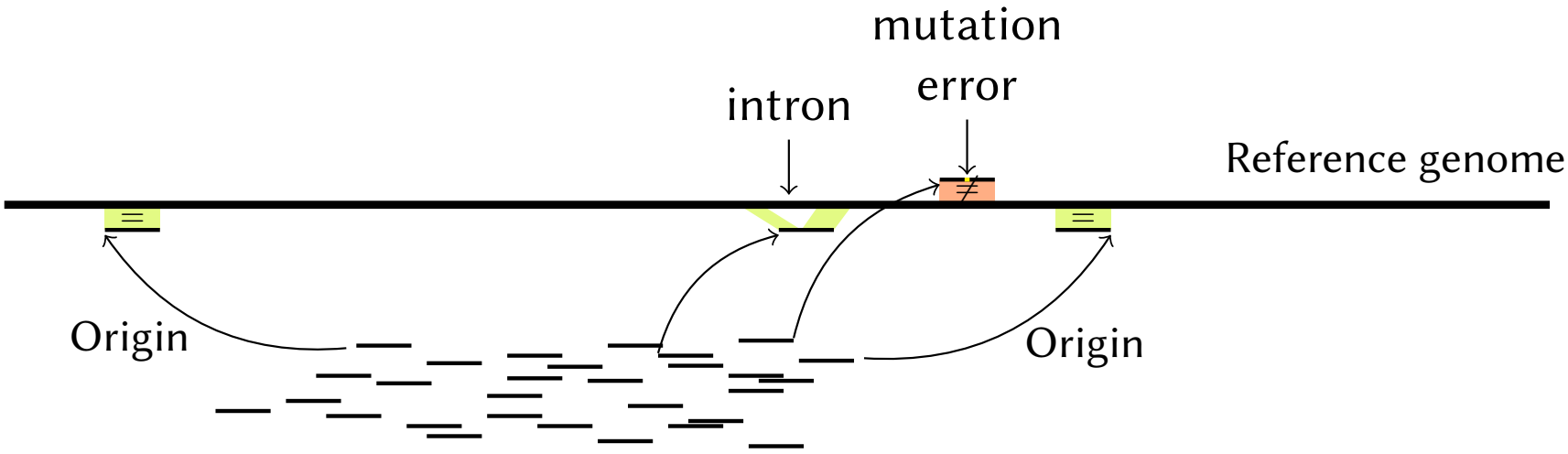
# RNA-Seq read mapping



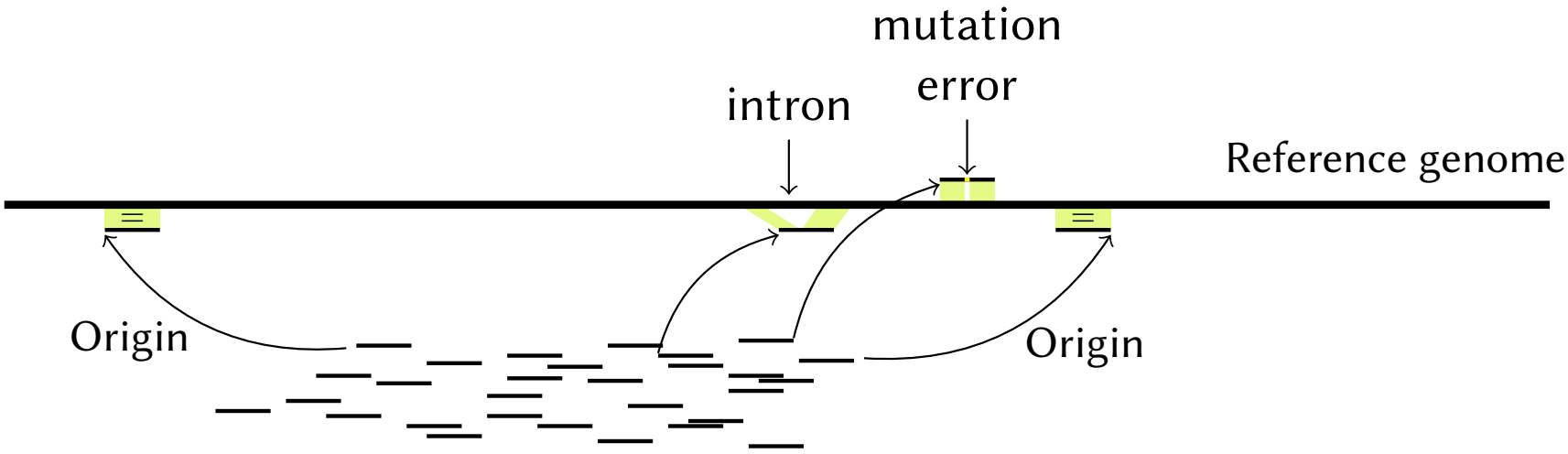
# RNA-Seq read mapping



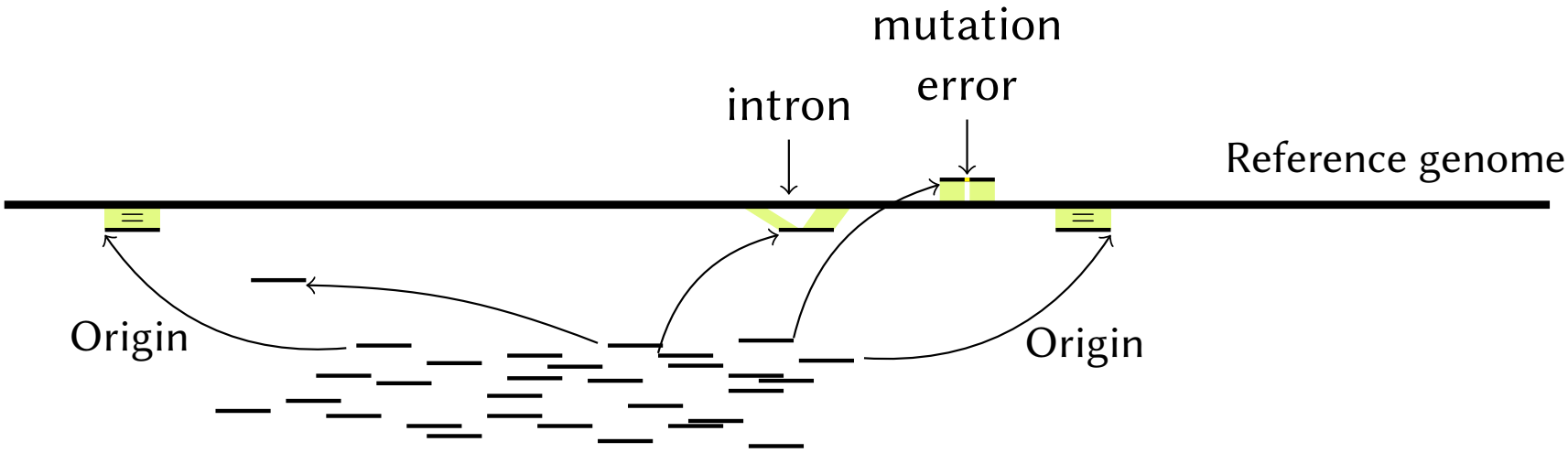
# RNA-Seq read mapping



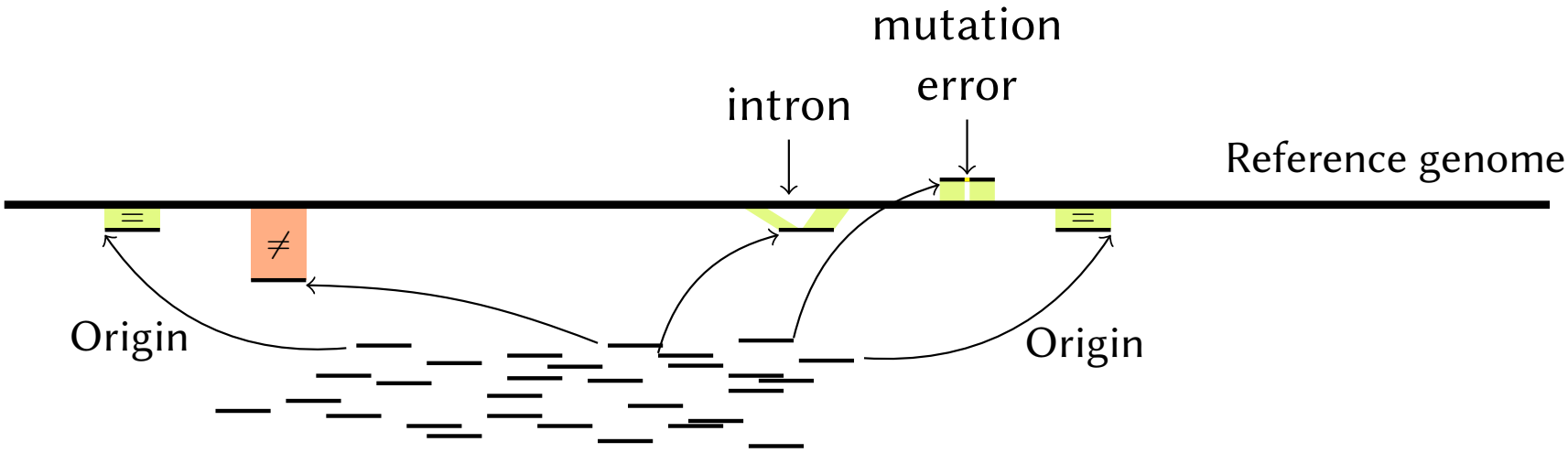
# RNA-Seq read mapping



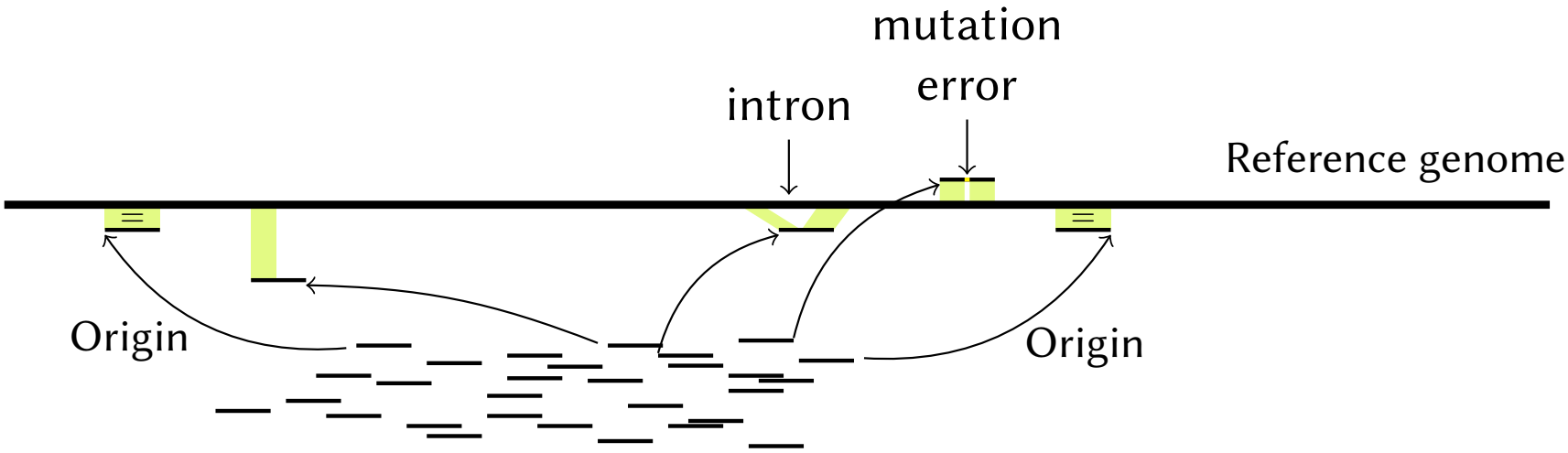
# RNA-Seq read mapping



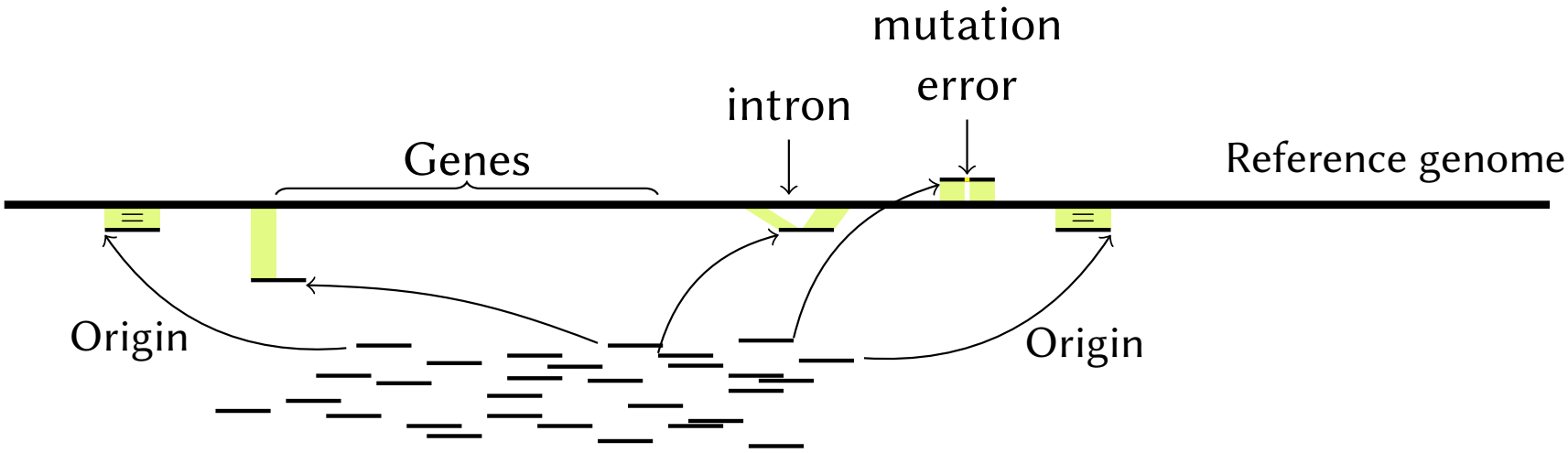
# RNA-Seq read mapping



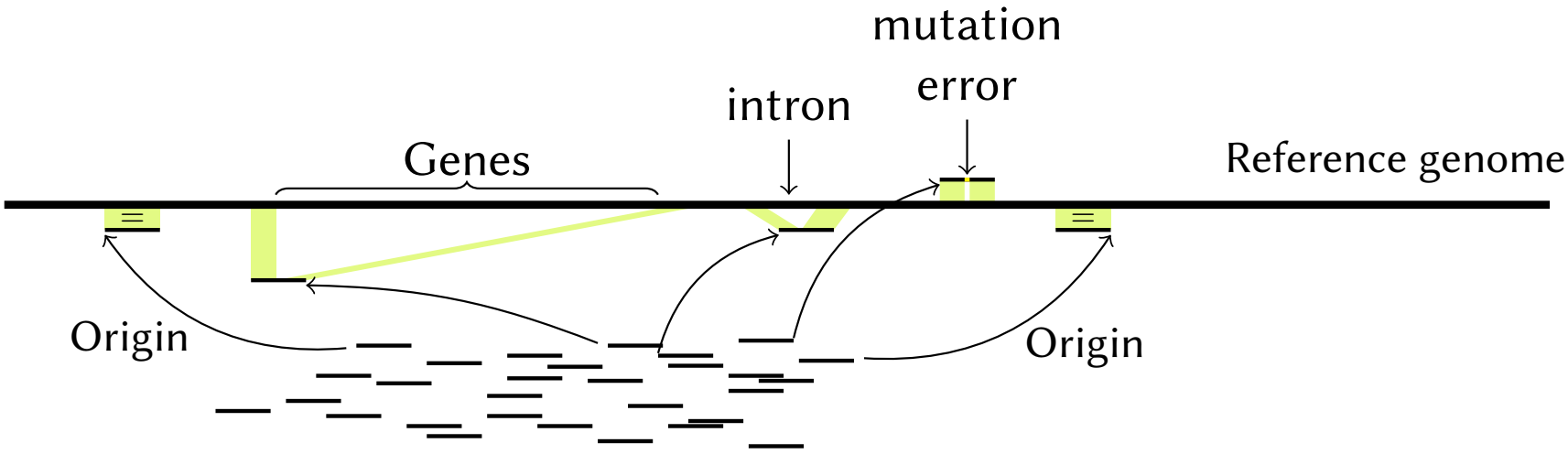
# RNA-Seq read mapping



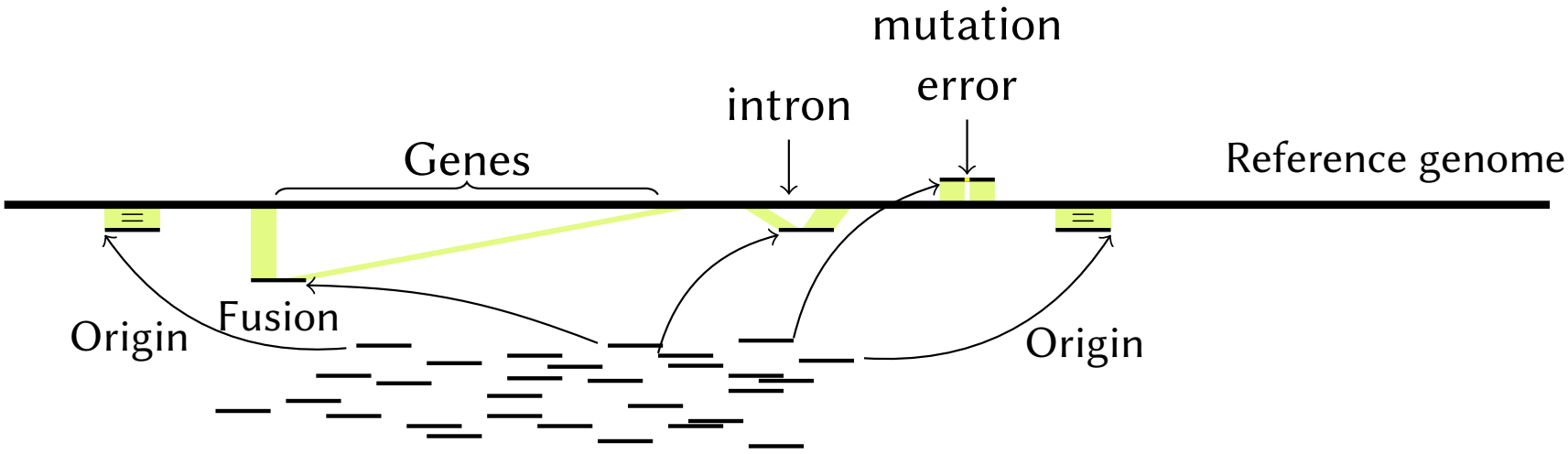
# RNA-Seq read mapping



# RNA-Seq read mapping

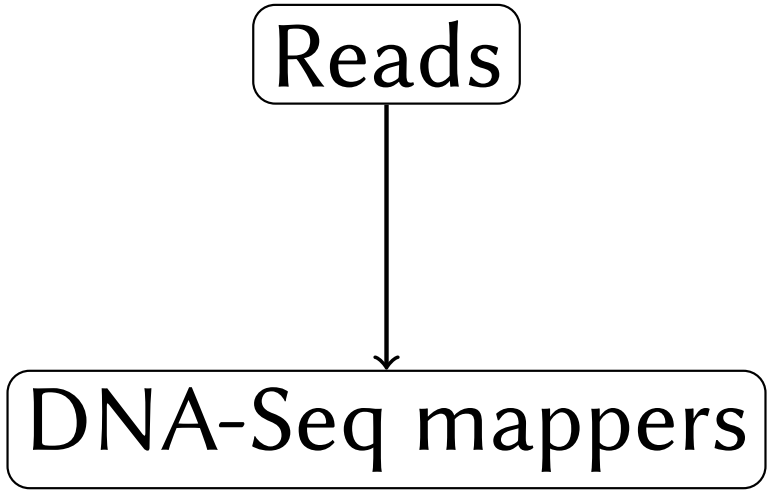


# RNA-Seq read mapping

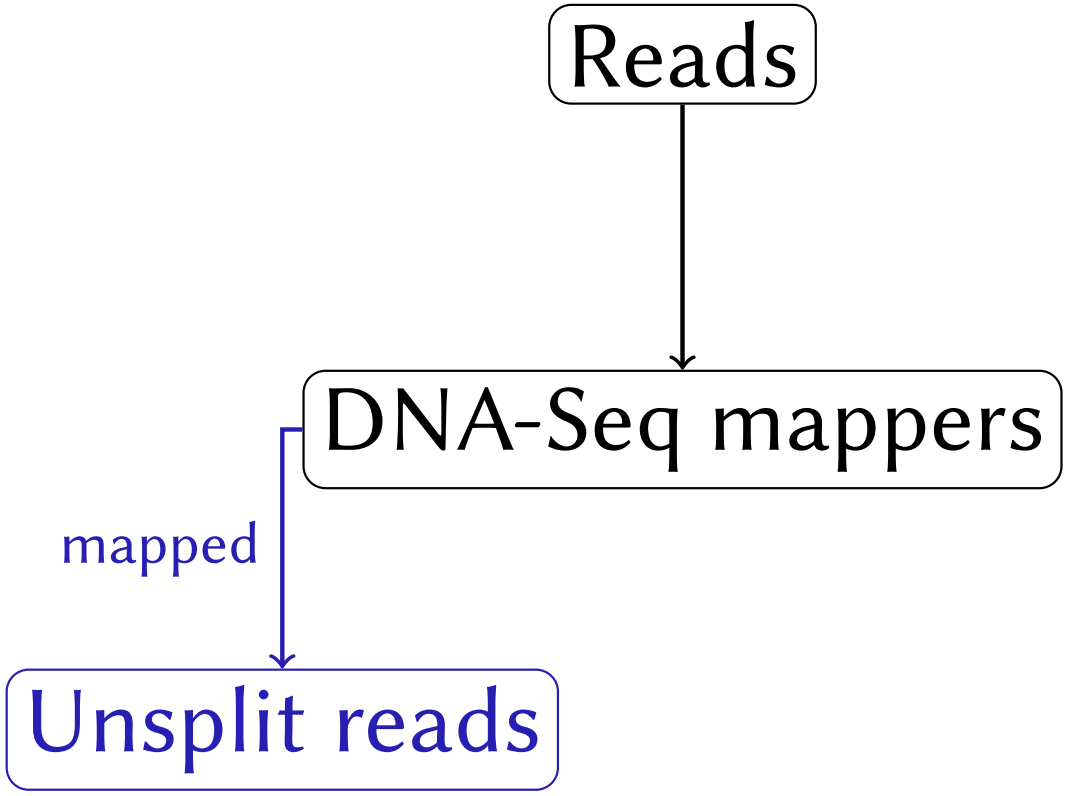


# Split reads don't align contiguously to the genome

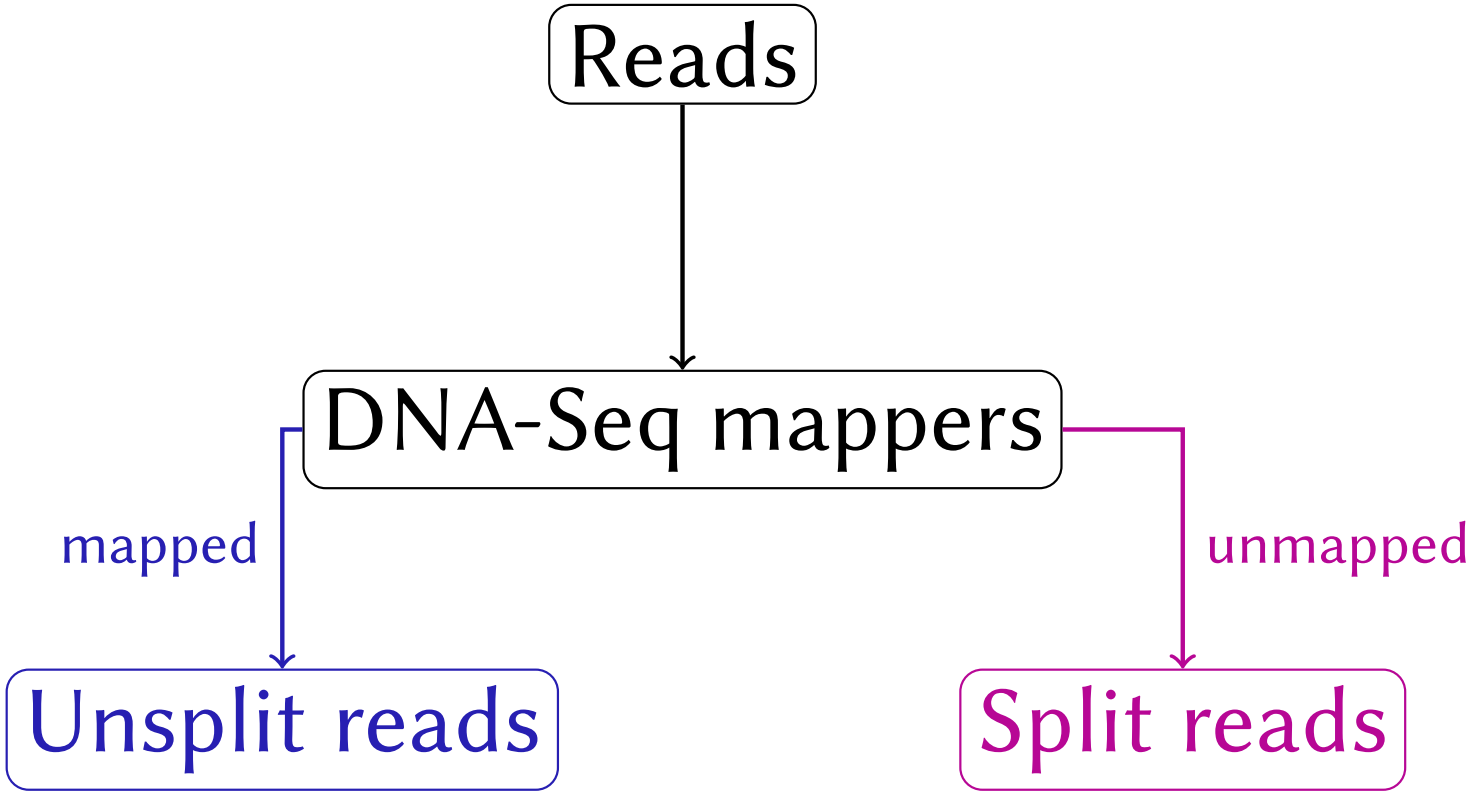
# Split reads don't align contiguously to the genome



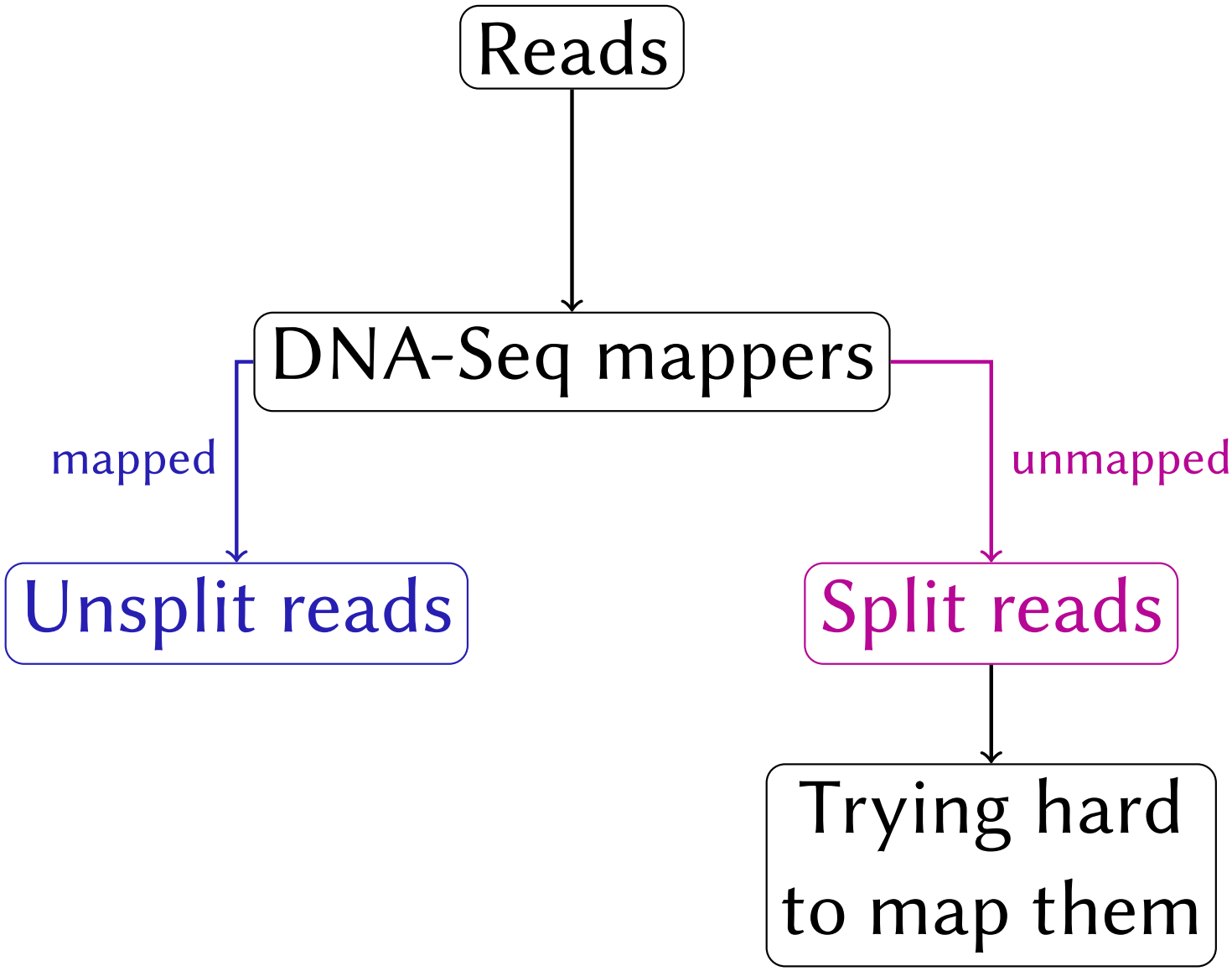
# Split reads don't align contiguously to the genome



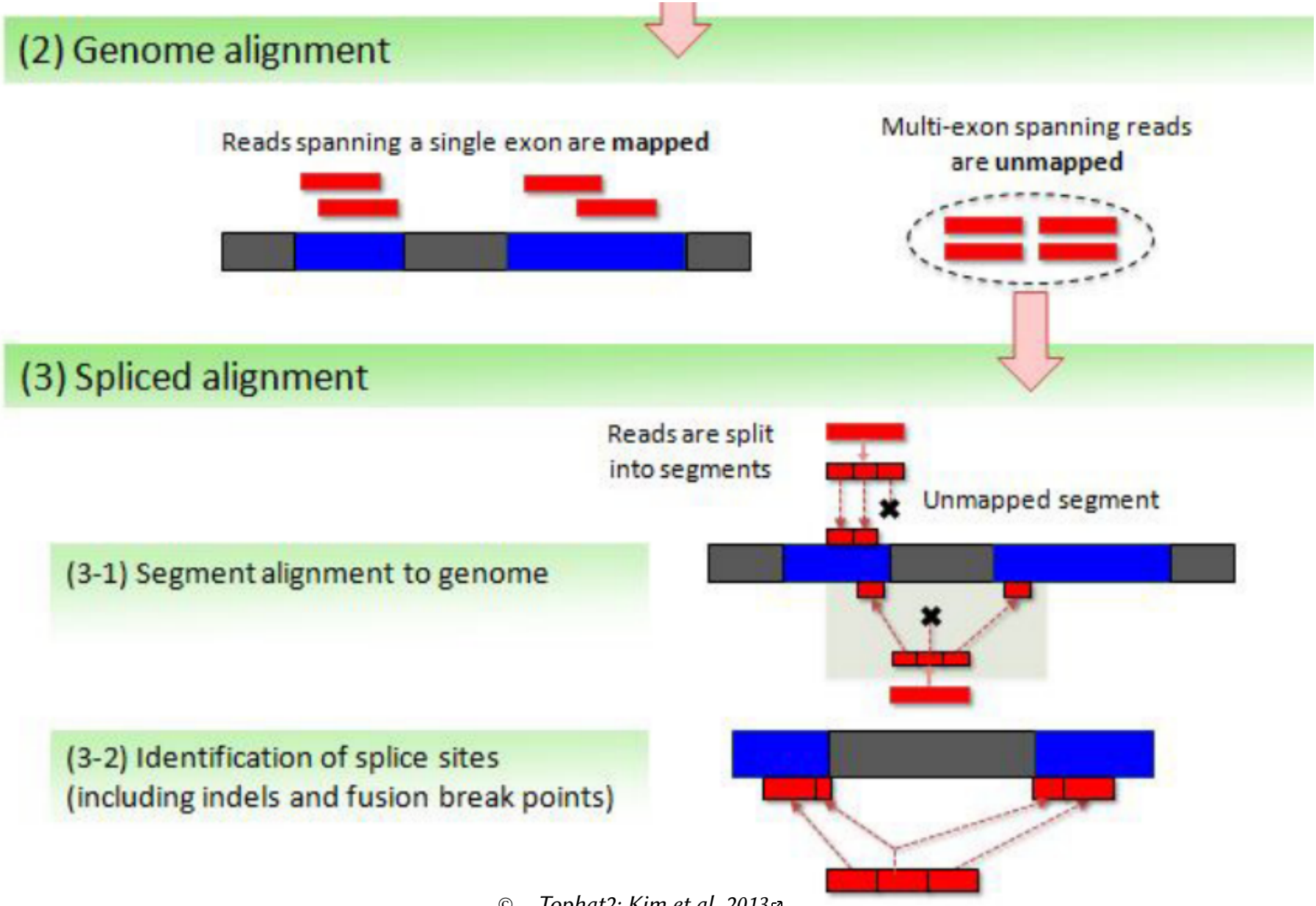
# Split reads don't align contiguously to the genome



# Split reads don't align contiguously to the genome

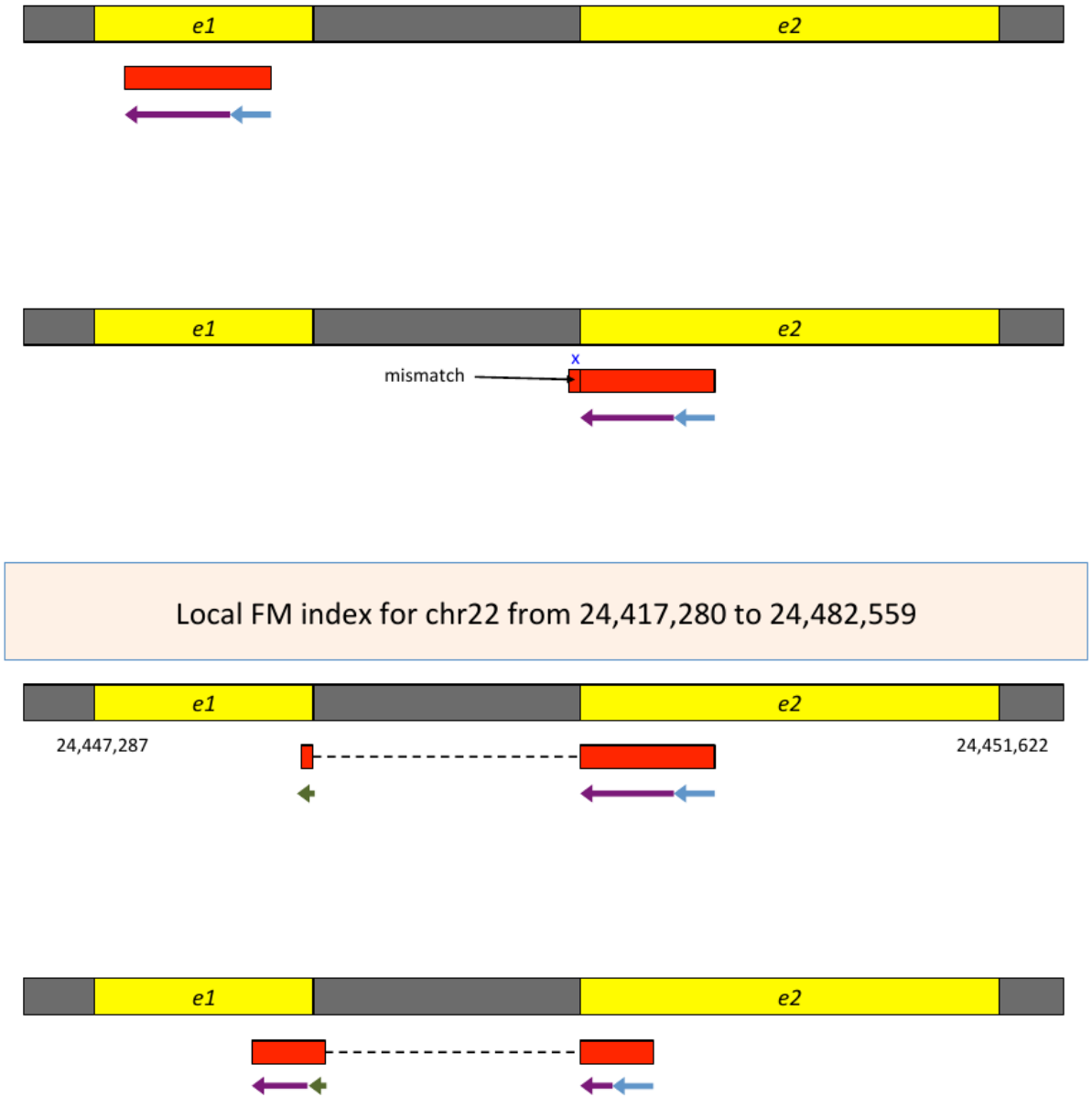


# Mapping split reads by… splitting them – TopHat2



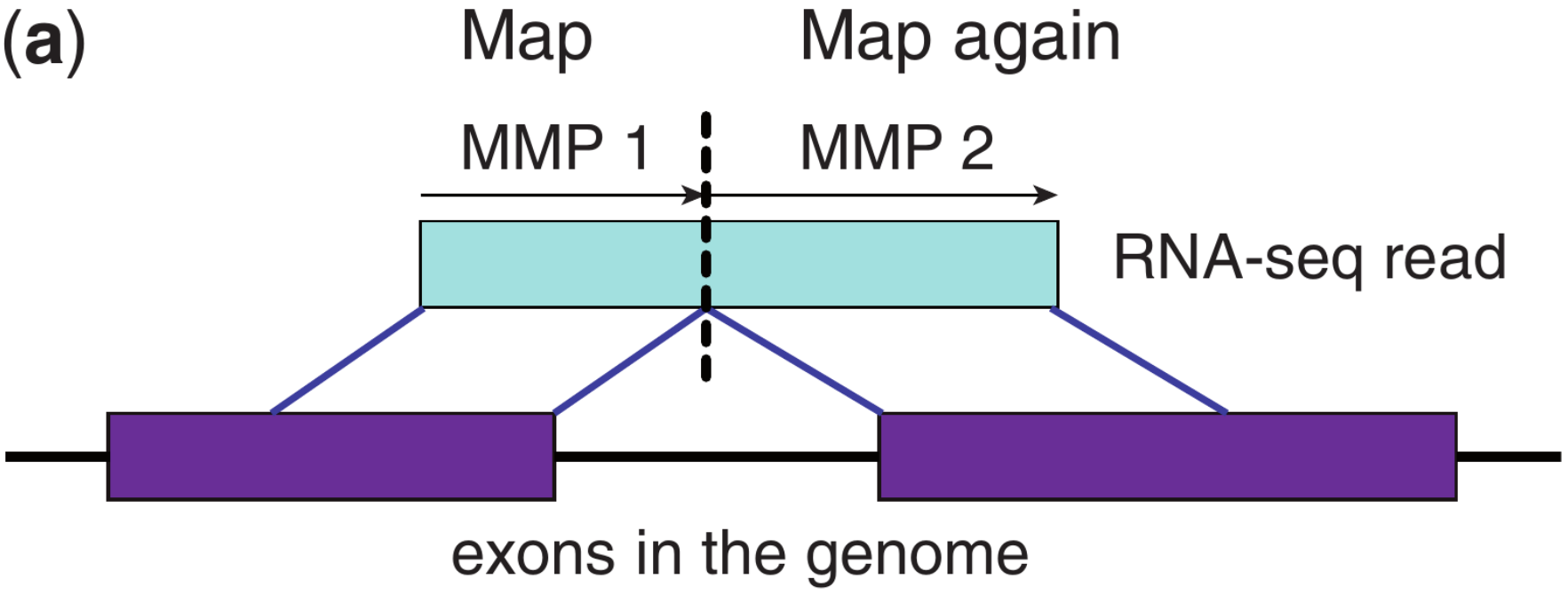

© *Tophat2: Kim et al, 2013*

# Mapping all reads by splitting them – HISAT2, STAR




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

# Mapping all reads by splitting them – HISAT2, STAR




© *STAR: Dobin et al, Bioinformatics, 2013*

# 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

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

# Indexing methods

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

| 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

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

| 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 |

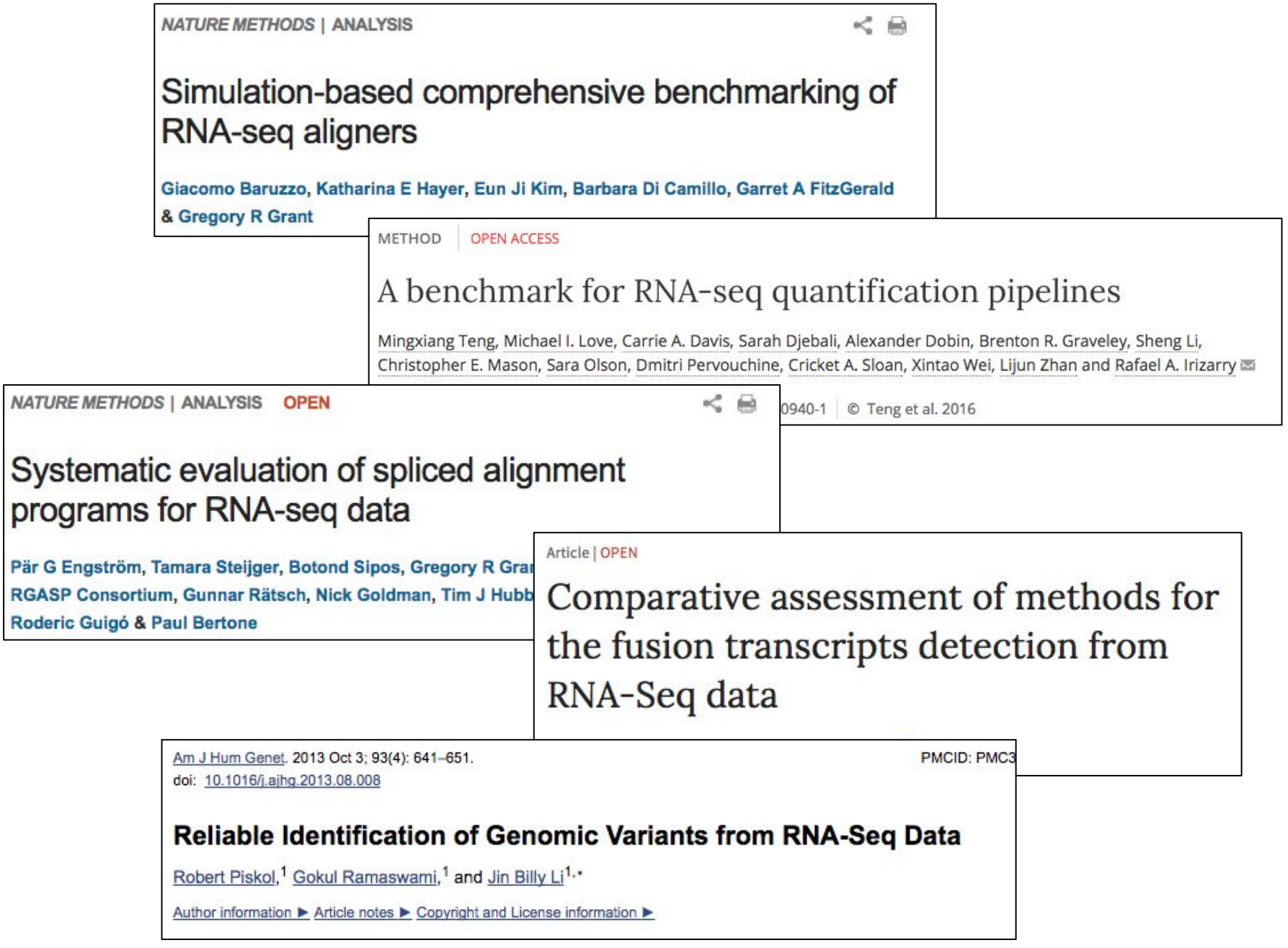
# Indexing methods

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

| 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** |

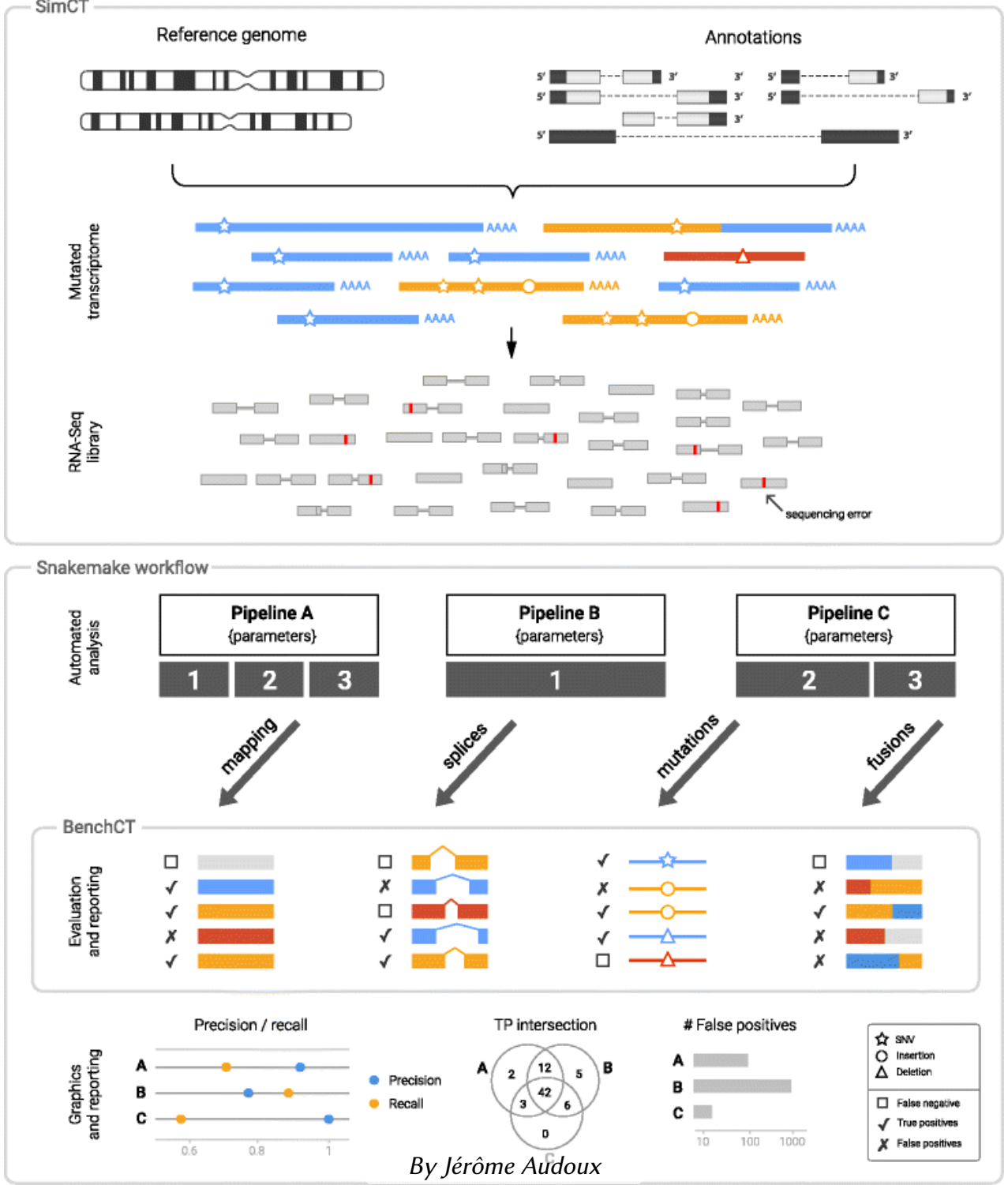
Burrows-Wheeler Transform

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



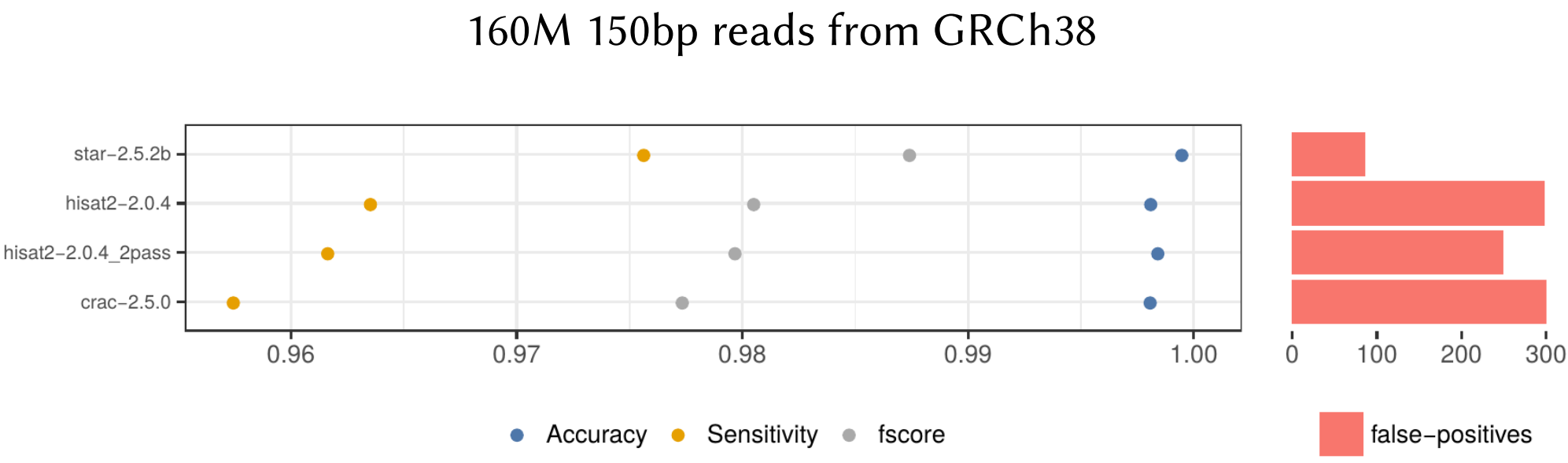
# Benchmarking RNA-Seq data

Audoux *et al*, BMC Bioinformatics, 2017



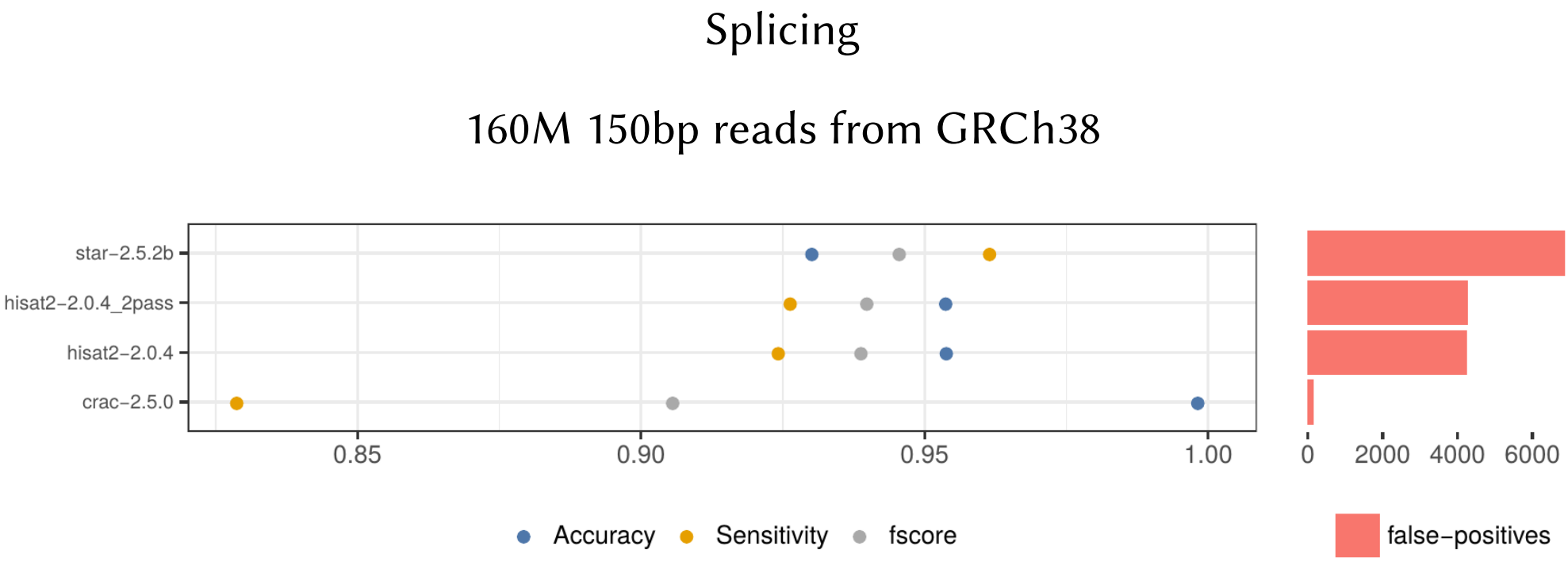

*By Jérôme Audoux*

# Sensitivity/accuracy of read mappers



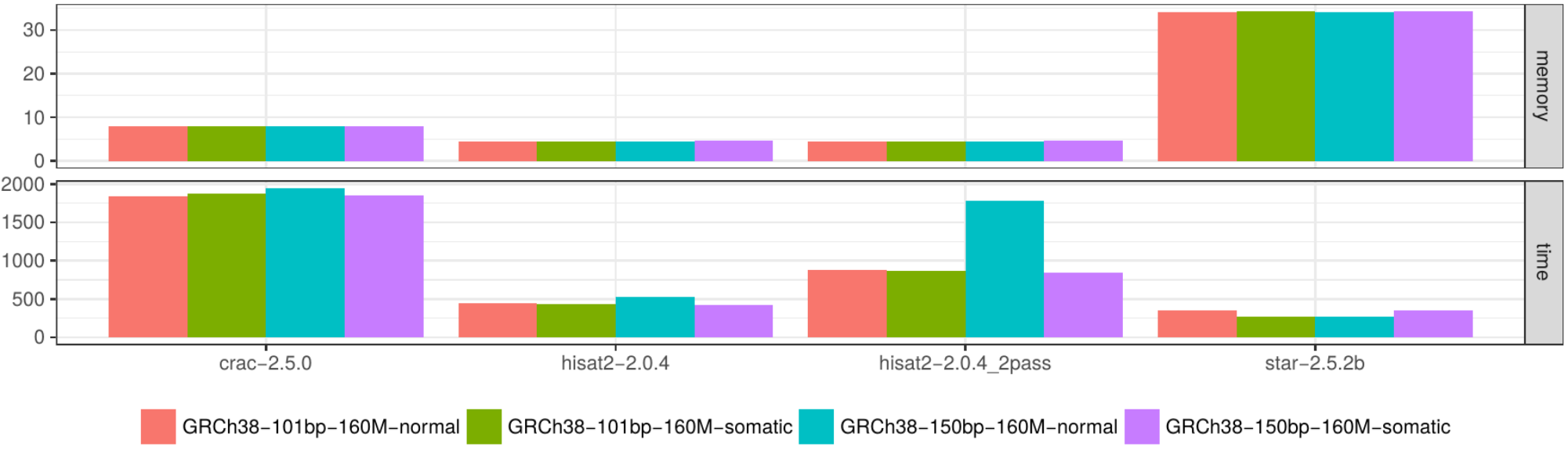

*By Jérôme Audoux*

# STAR offers the best trade-off for splice detection




*By Jérôme Audoux*

# Space/time for read mappers




*By Jérôme Audoux*

# Many people uses TopHat2

($>$ 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

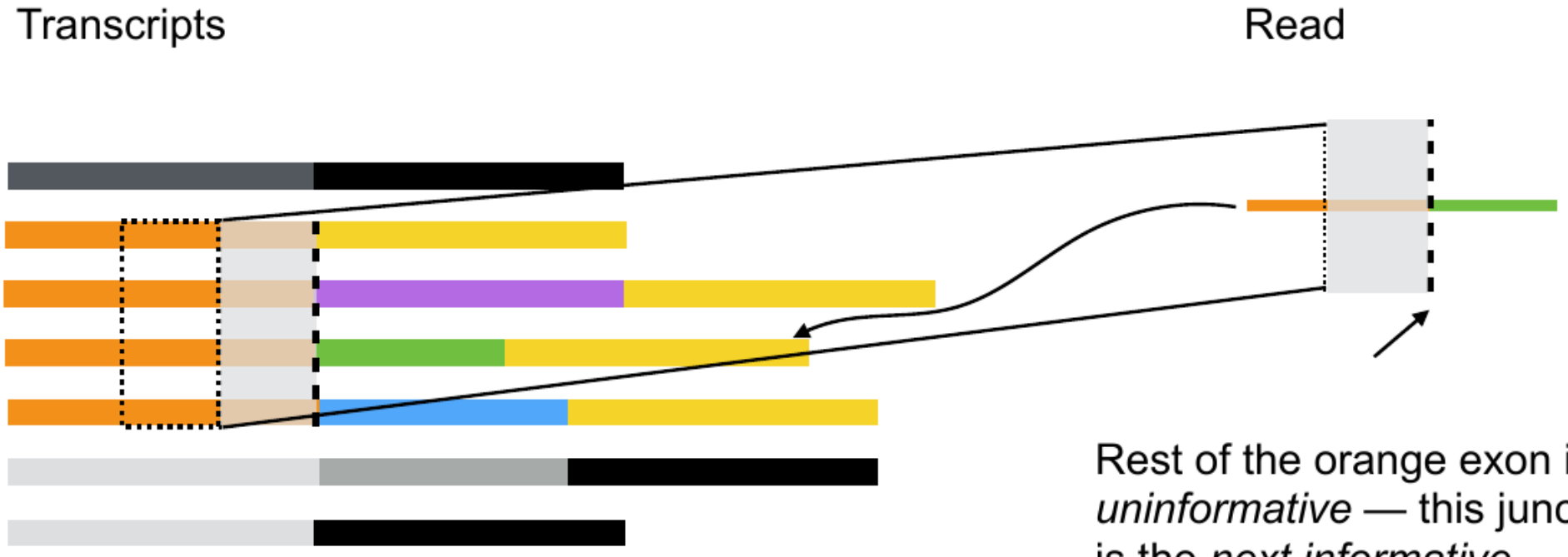
Bray et al, Nat. Biotechnology, 2016

Salmon

Patro et al, Nat. Methods, 2017

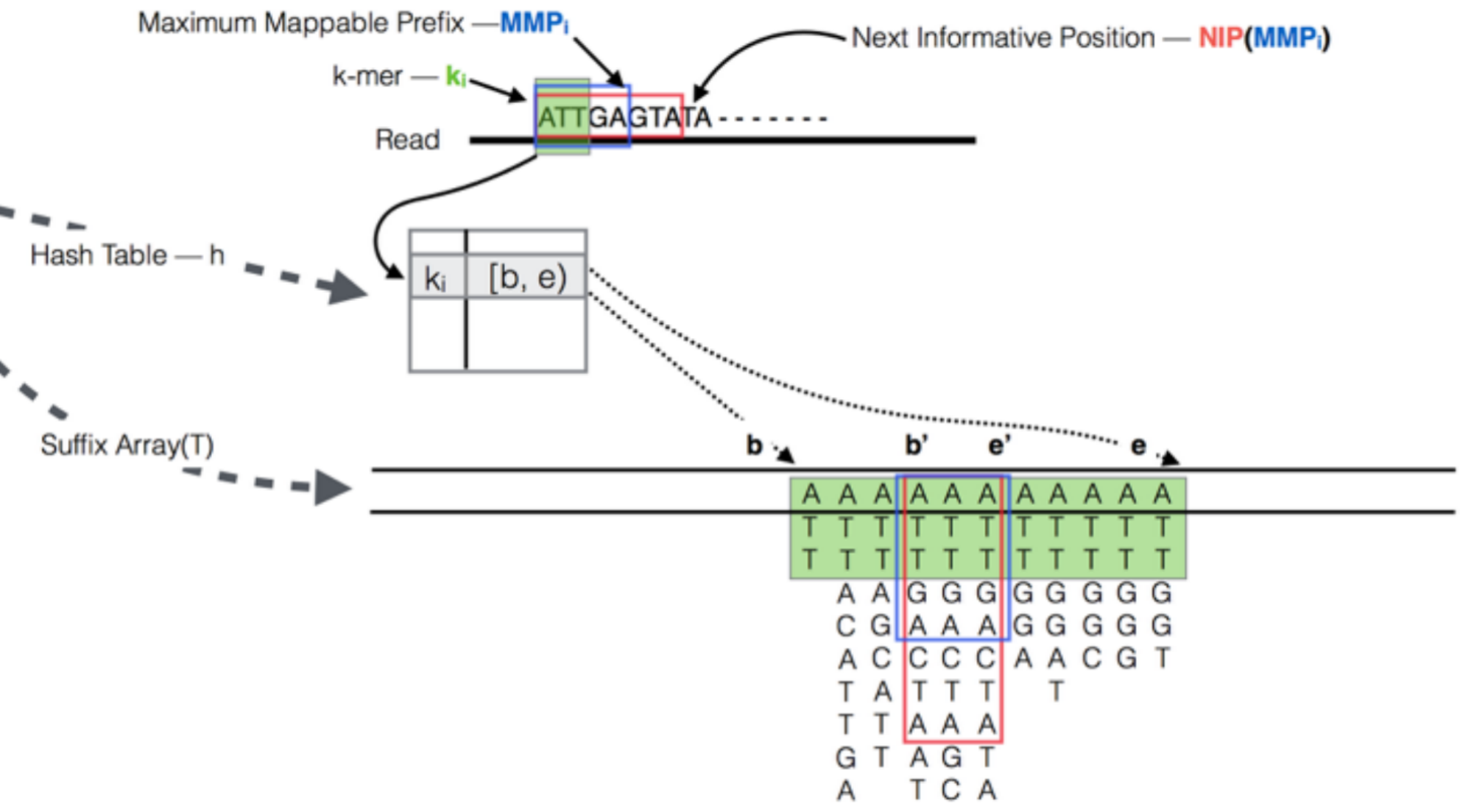
Two orders of magnitude faster than TopHat+Cufflinks

# How to quantify without aligning?




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…

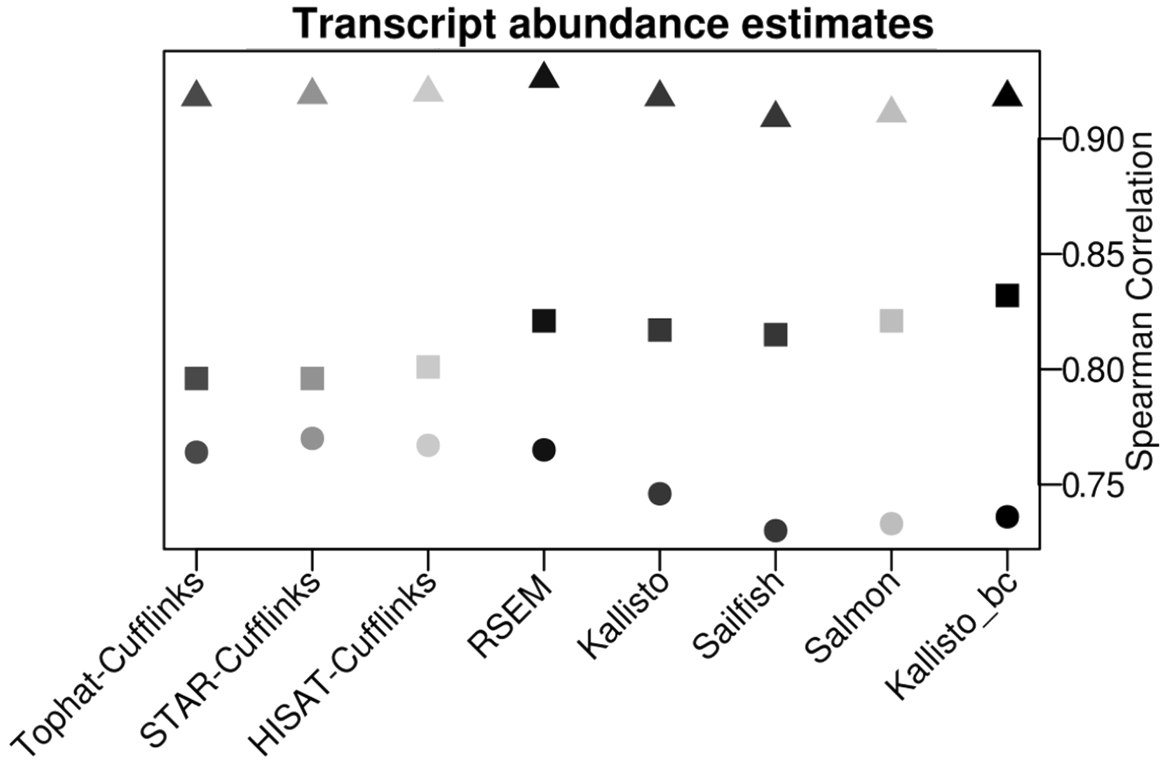
« *With the exception of the underperforming Flux Capacitor and eXpress, we found that the other algorithms performed similarly.* »

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…

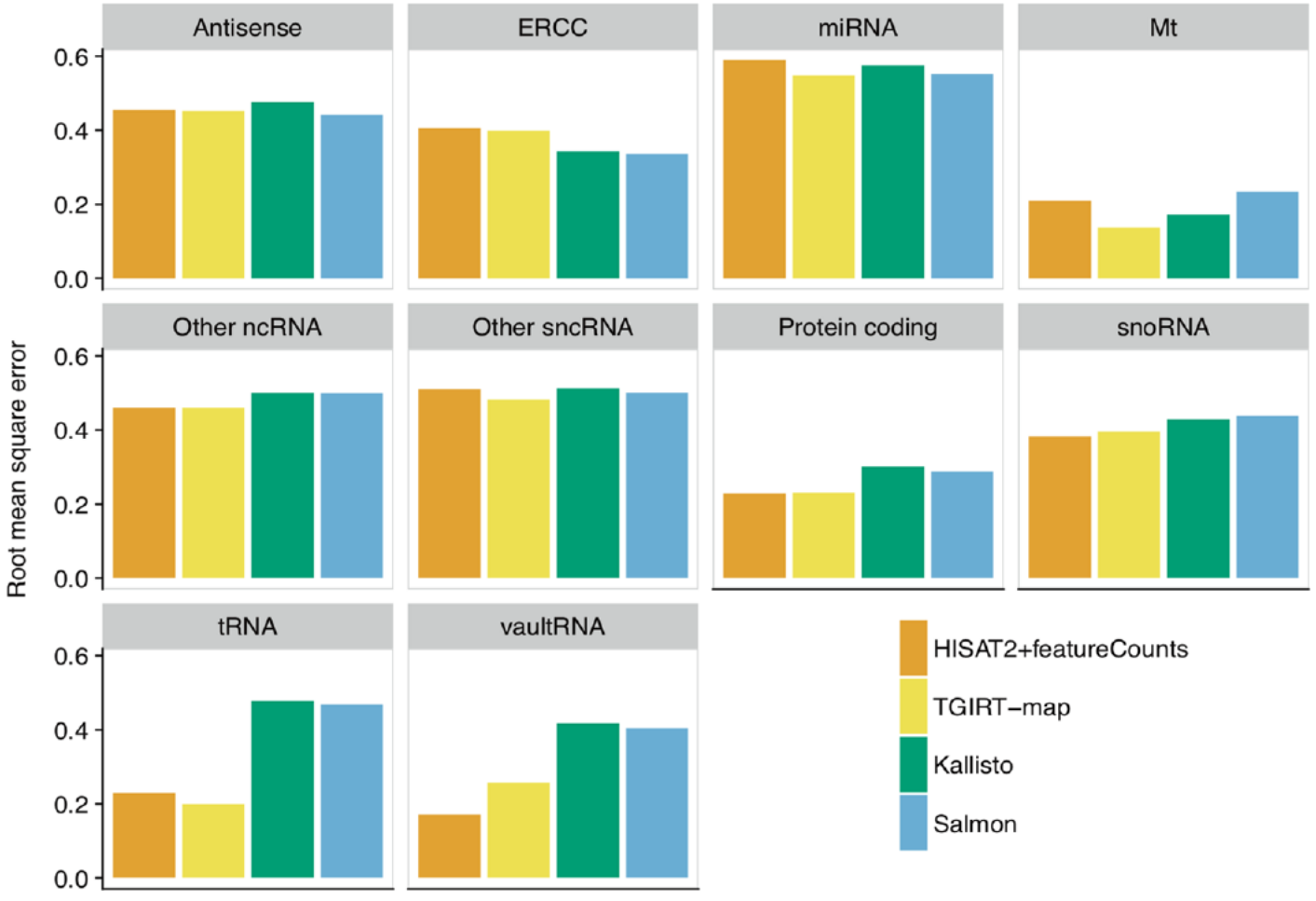
« *With the exception of the underperforming Flux Capacitor and eXpress, we found that the other algorithms performed similarly.* »

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.* »

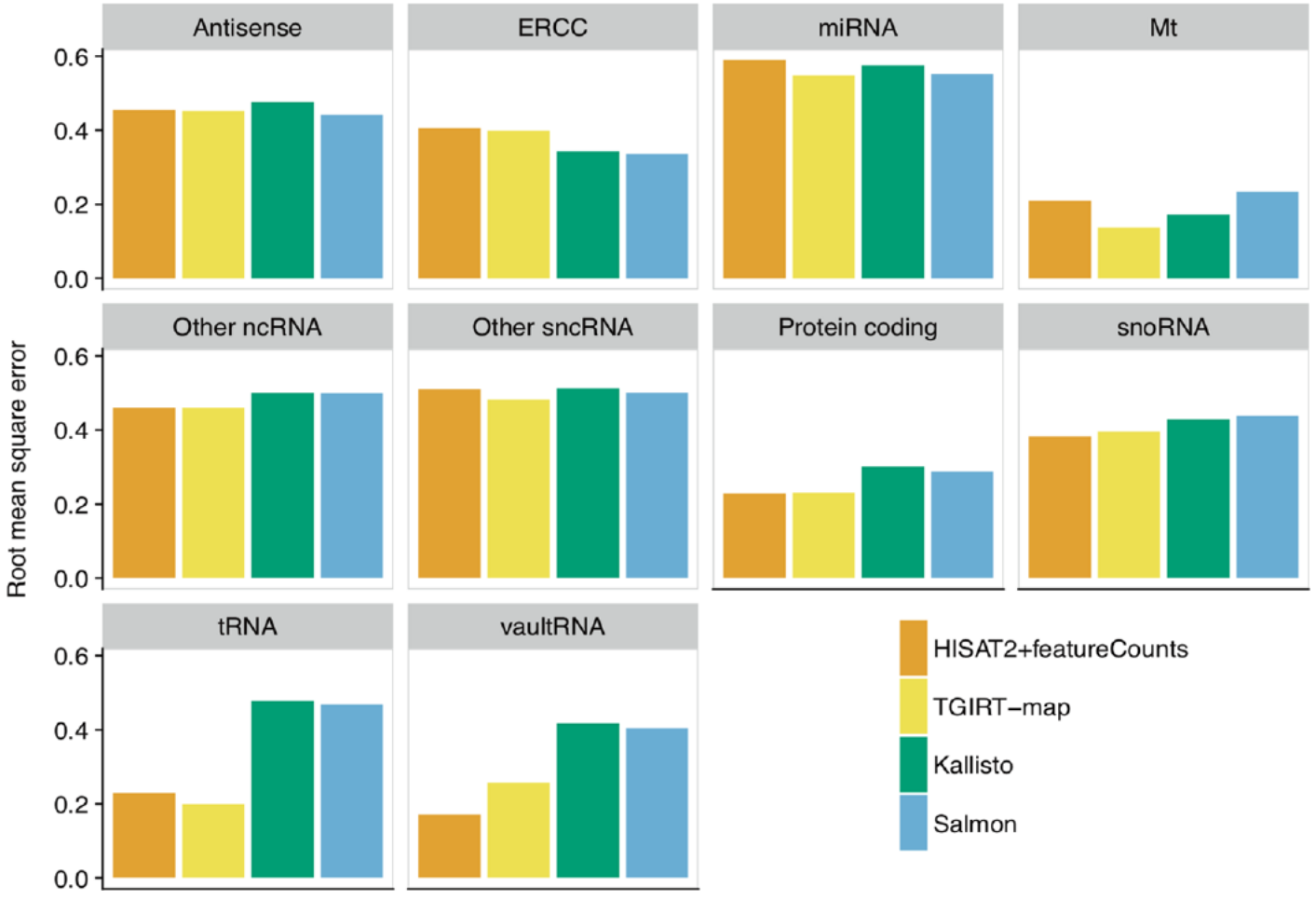
Germain et al, Nucleic Acid Research, 2016

# Good results that depend on the type of RNAs




 *Wu et al, 2018*

# Good results that depend on the type of RNAs




 *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

**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**