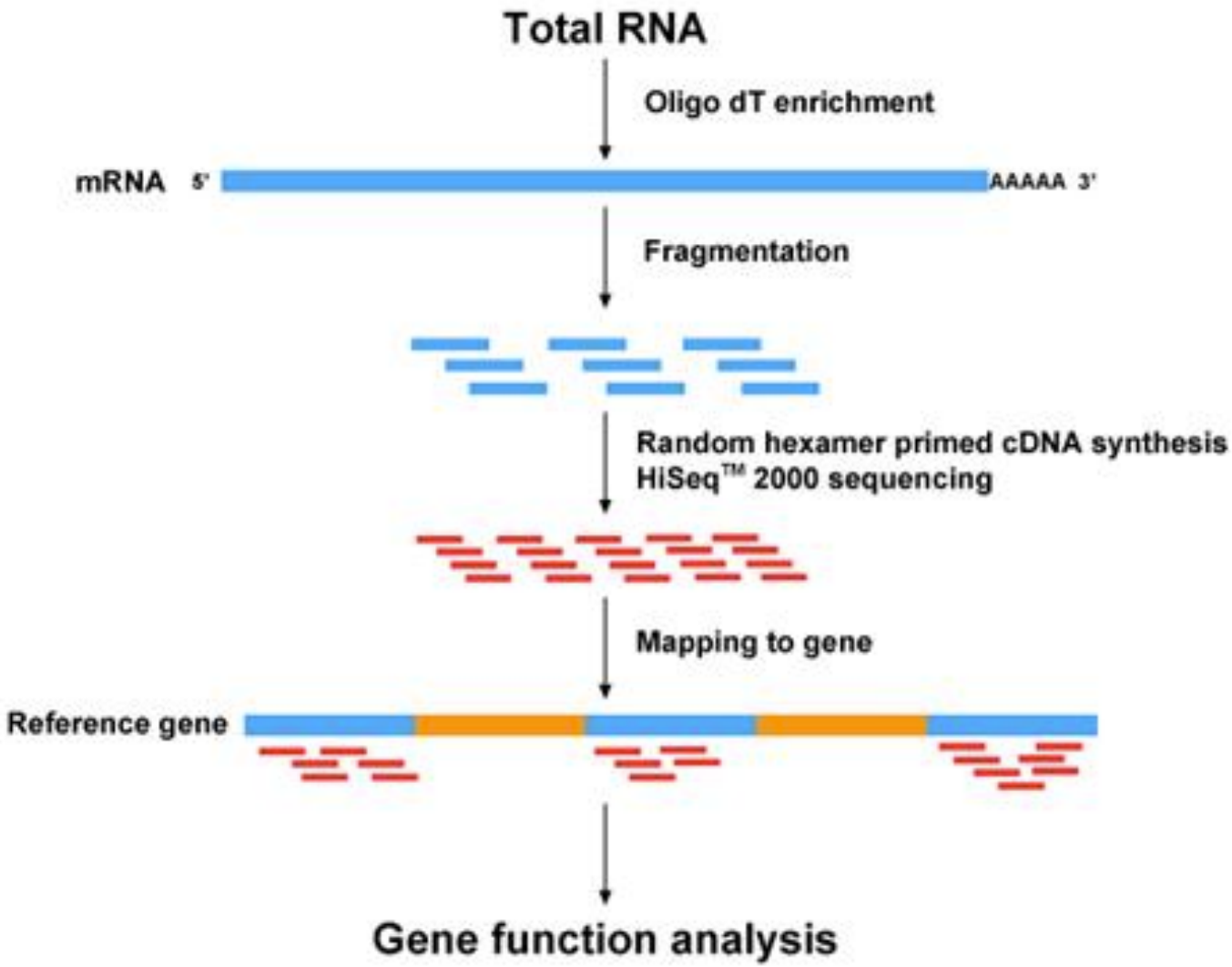
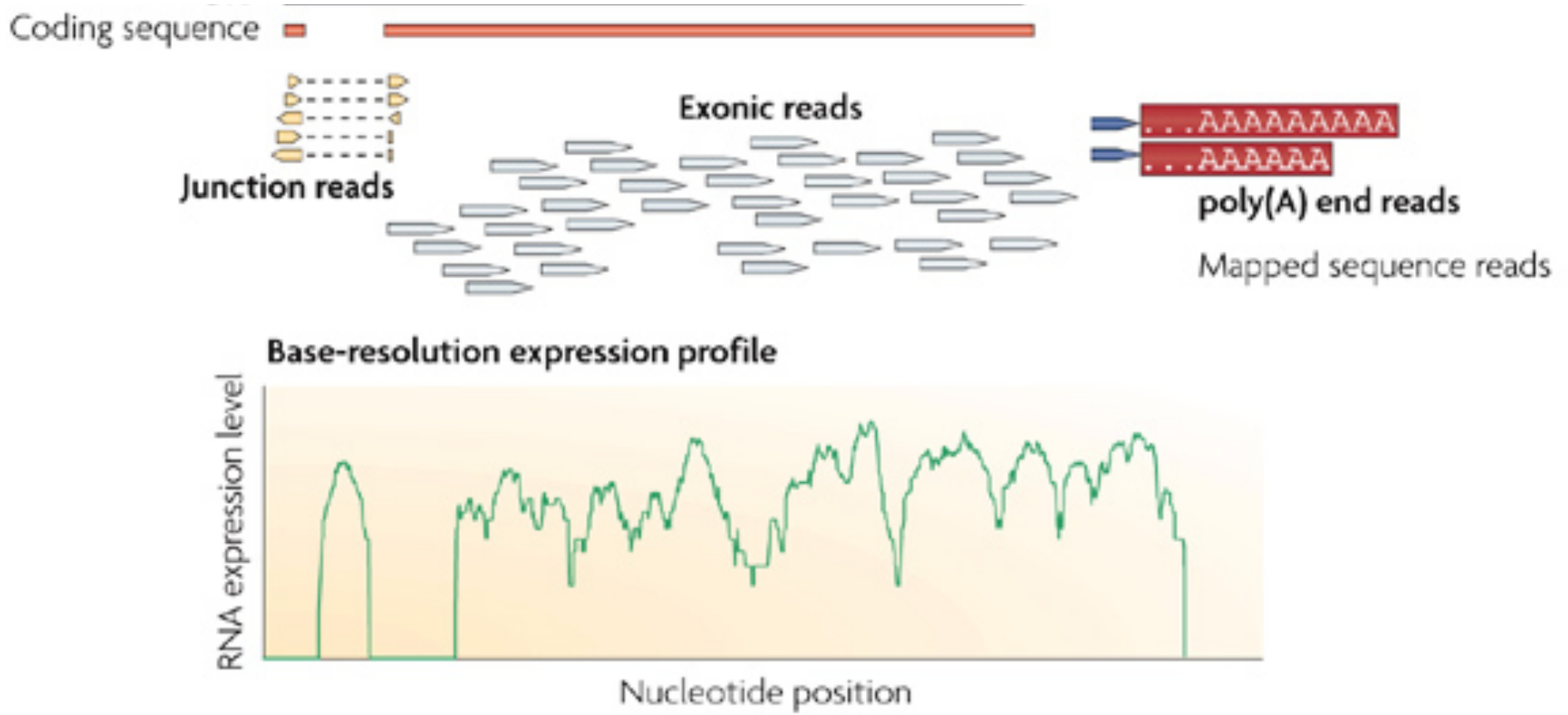
RNA-SEQ DATA ANALYSIS PIPELINE: READ MAPPING

TRANSCRIPTOME RECONSTRUCTION

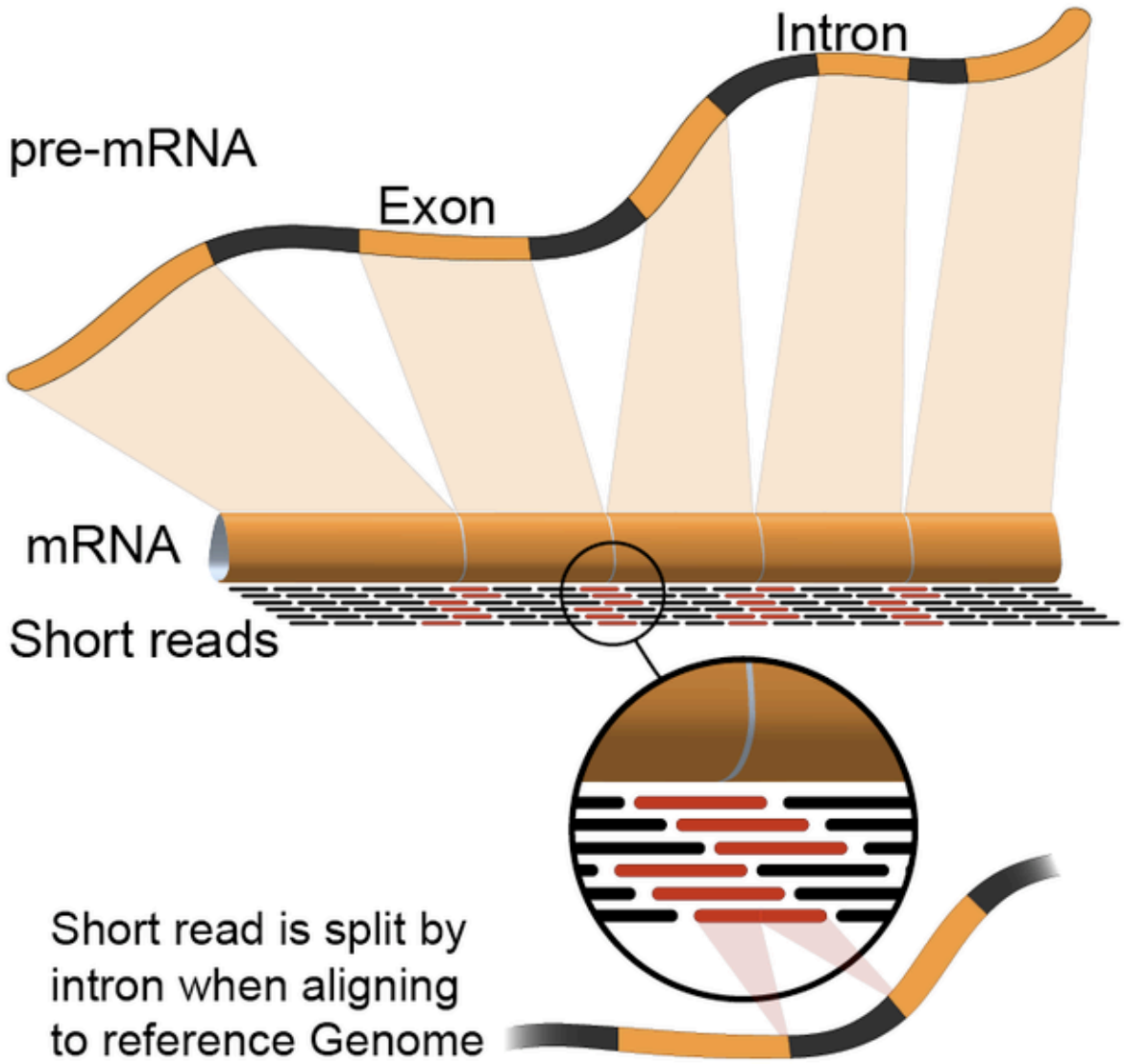


TRANSCRIPTOME RECONSTRUCTION



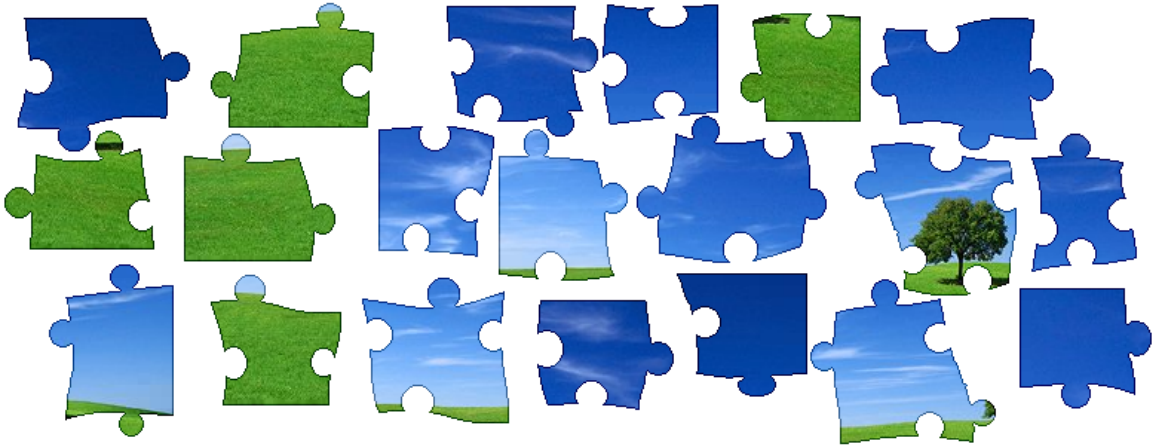
Nature Reviews | Genetics

TRANSCRIPTOME RECONSTRUCTION



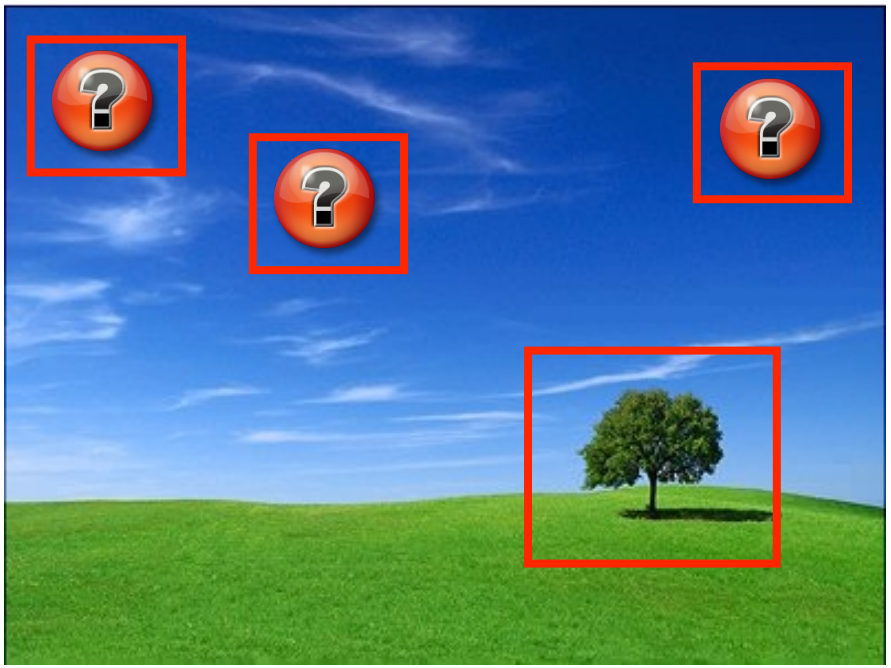
READ MAPPING

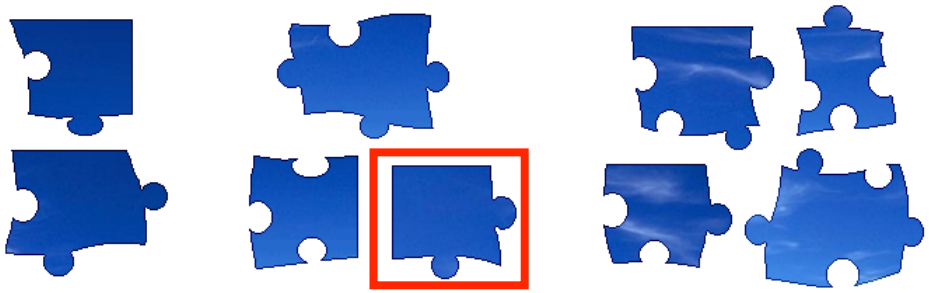
- Mapping reads onto a reference genome is analogous to complete a puzzle for which we know the image on the box cover



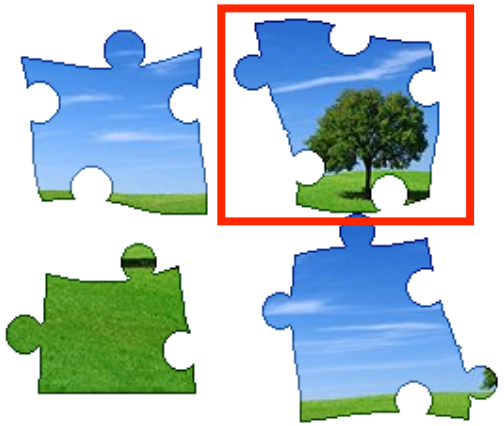
READ MAPPING

- Bot not all puzzle pieces are equally easy to place correctly



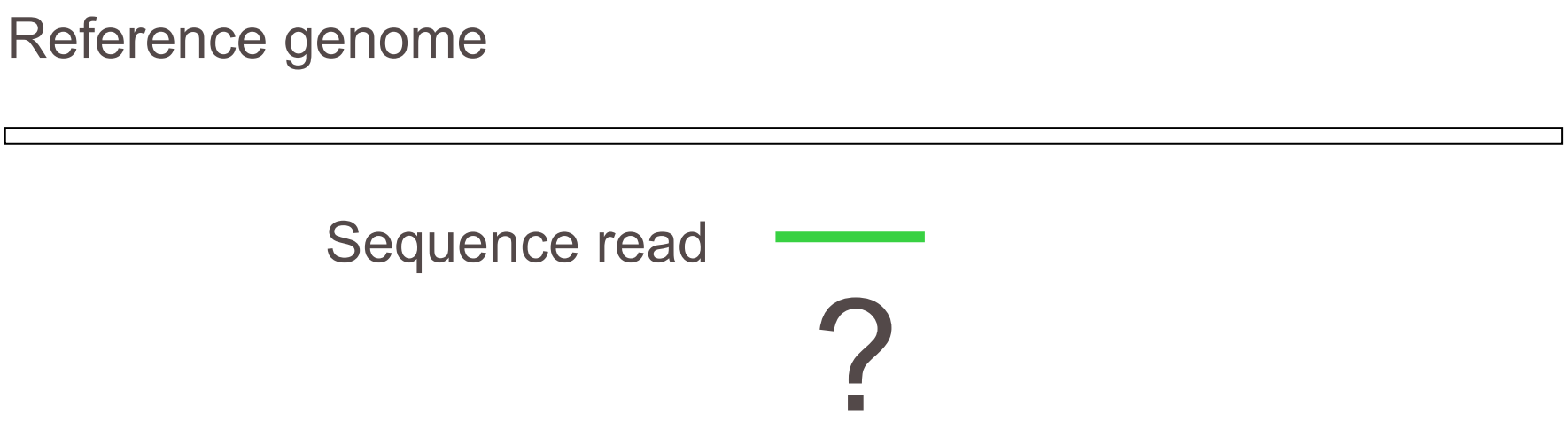


Some pieces are very similar to each other….



…while others contain features that facilitate their placement

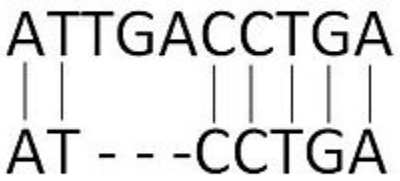
READ MAPPING



- For each read, one must map it to a genomic locus, from which transcription the read most probably originated
- In practice, this would mean comparing the read sequence to the sequence of the reference genome

READ MAPPING

- Read mapping is an alignment of each read sequence to the sequence of the reference genome of the species from which the samples come from



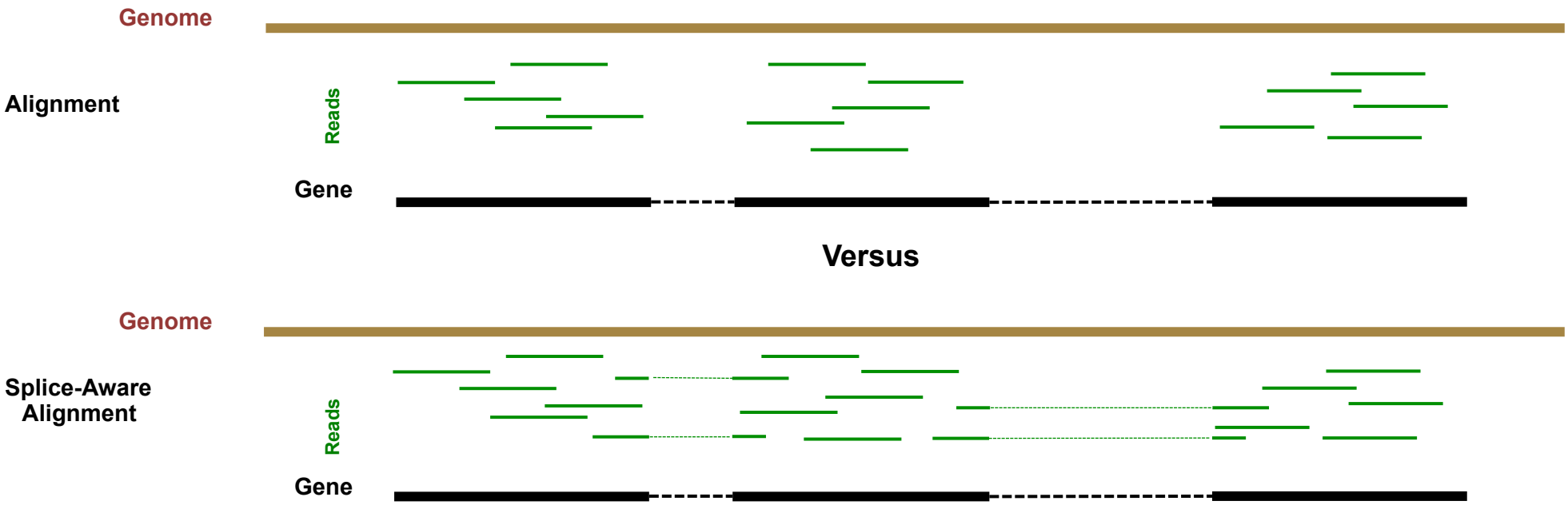
- Since the number of reads is in the order of millions, tens of millions or hundreds of millions, and the size of the genome can be in the order of billions of nucleotides, computing all these alignments using standard alignment algorithms is unfeasible.

READ MAPPING

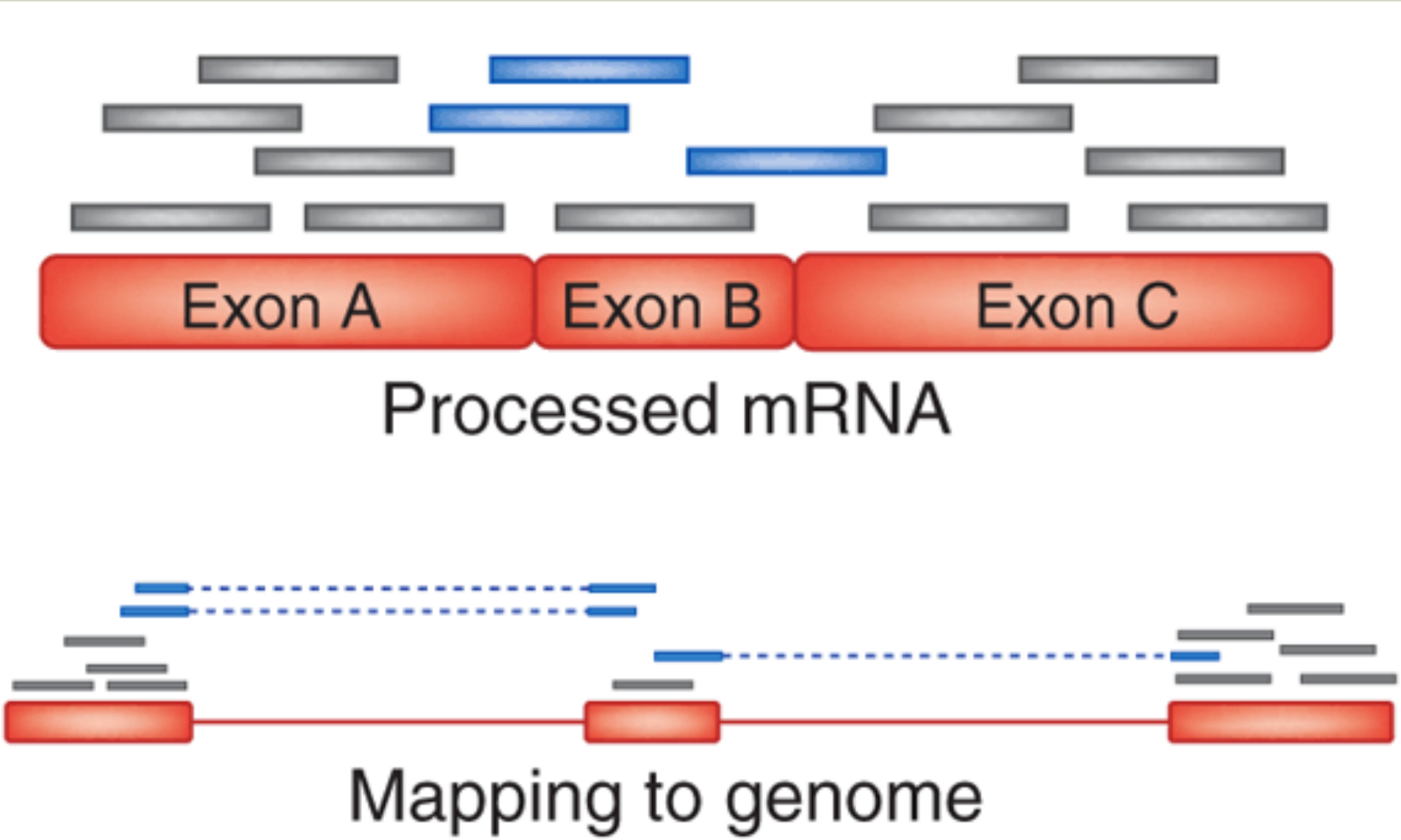
- When aligning reads to a reference, the first choice is between mapping reads onto a genomic sequence or onto the sequence of all known transcripts for the species under analysis
- This is an important issue that affects the analysis outcomes and the employed pipeline
- If you align reads to the genome, all reads spanning exon-exon junctions in the mature mRNA cannot be mapping from the beginning to the end on the genome. Hence, you would need a splicing-aware aligner
- When aligning against the transcriptome, this problem does not arise, but all reads coming from exons that can be included in several isoforms of the same genes will be mapped onto multiple transcripts

READ MAPPING

READ MAPPING



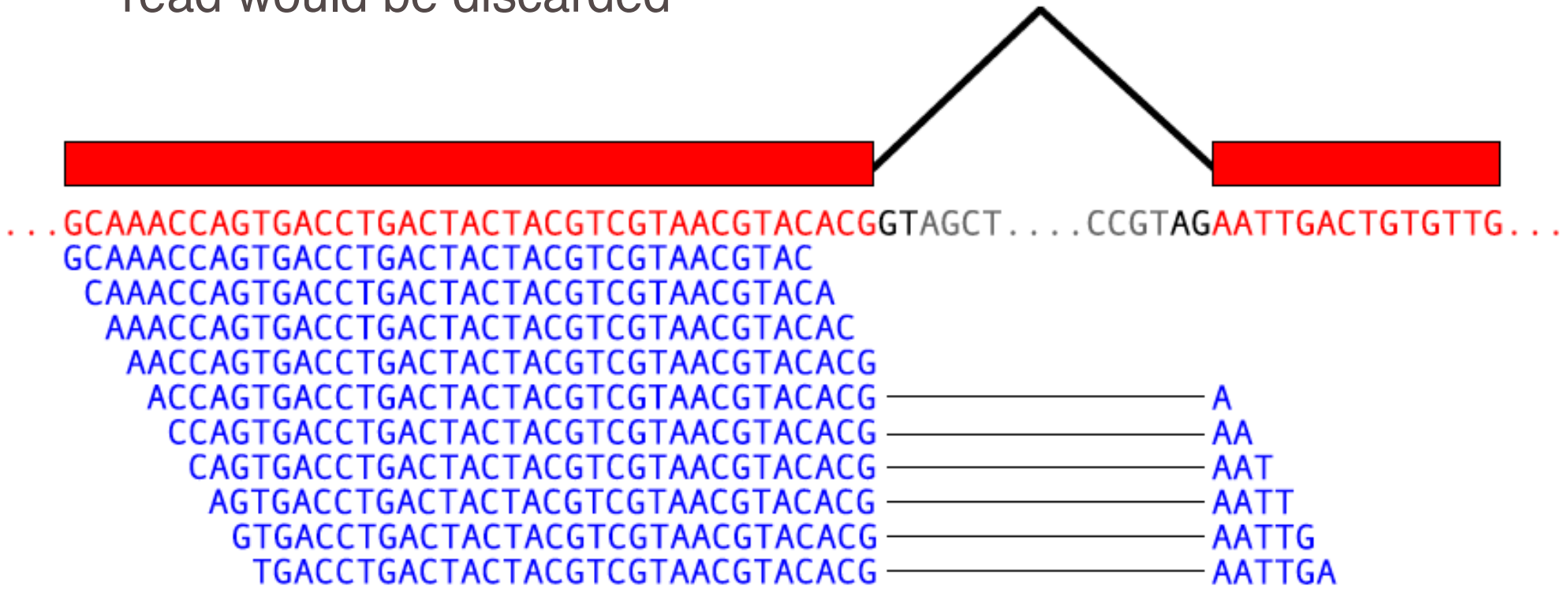
READ MAPPING



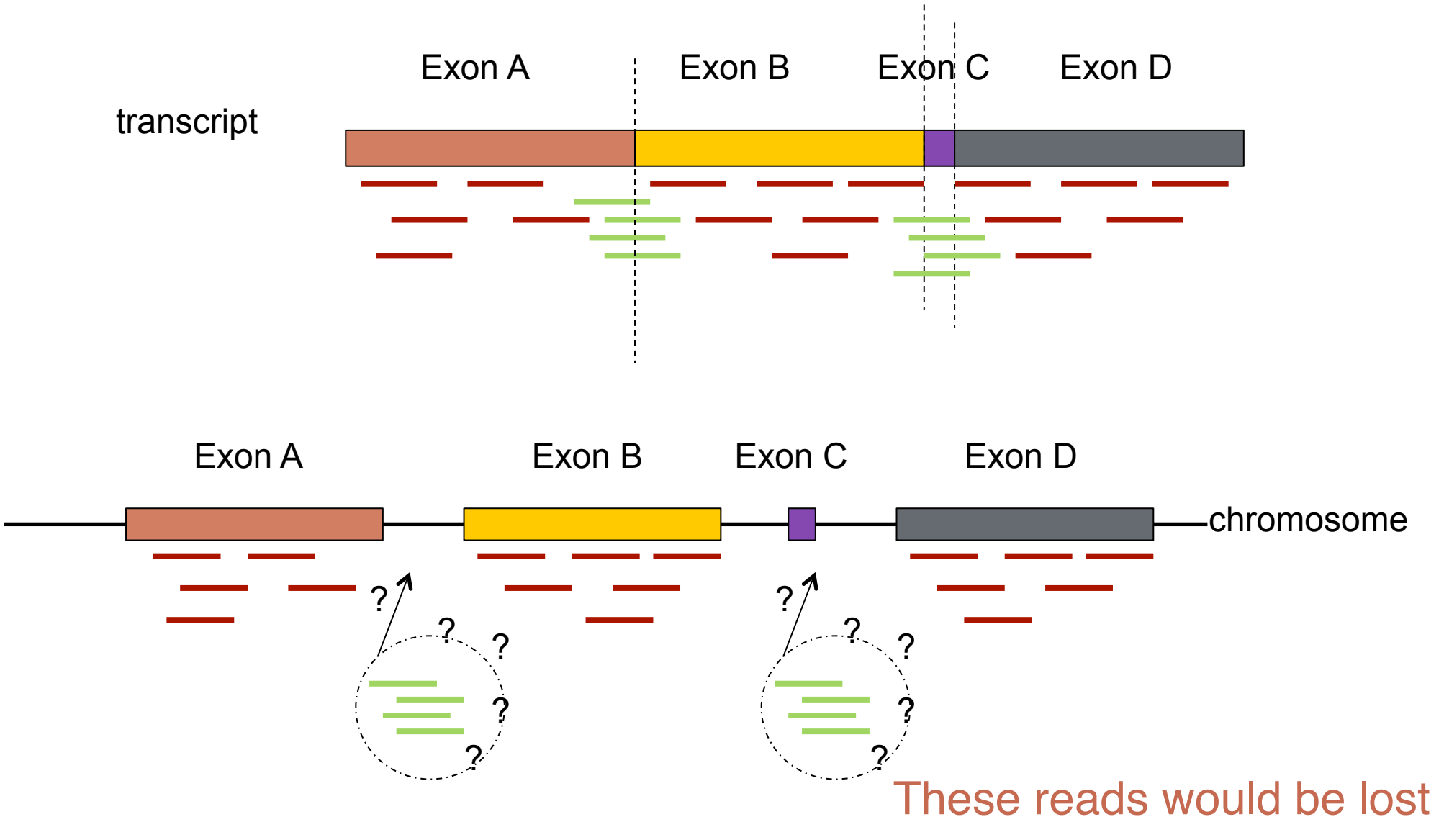
Trapnell & Salzberg, Nature Biotechnology 2009

READ MAPPING

- Reads at the exon-exon junctions are broken during the mapping step. If the two (or more) fragments are too short, the alignment could be judged as of insufficient extension and the read would be discarded



READ MAPPING



These reads would be lost

READ MAPPING

TopHat
A spliced read mapper for RNA-Seq

McKUSICK-NATHANS Institute of Genetic Medicine

CBCB

TopHat is a fast splice junction mapper for RNA-Seq reads. It aligns RNA-Seq reads to mammalian-sized genomes using the ultra high-throughput short read aligner Bowtie, and then analyzes the mapping results to identify splice junctions between exons.

TopHat is a collaborative effort between the Institute of Genetic Medicine at Johns Hopkins University, the Departments of Mathematics and Molecular and Cell Biology at the University of California, Berkeley and the Department of Stem Cell and Regenerative Biology at Harvard University.

OSI certified

» **TopHat 2.0.11 release 3/4/2014**

Version 2.0.11 is a maintenance release with the following simple fix:

- This version is compatible with Bowtie2 v2.2.1, although it does not support a 64-bit Bowtie2 index yet.

» **TopHat 2.0.10 release 11/13/2013**

Version 2.0.10 is a maintenance release with the following fixes and changes:

- Improved support for adding unpaired reads to PE reads in the same TopHat2 run (please see the manual entry for this usage). This includes reporting separate counts for the additional unpaired reads and making sure that the SAM flags in the output files reflect the paired or unpaired origin of the reads.
- Added the possibility to run TopHat just for the purpose of preparing the transcriptome index files (please see the manual entry for this special usage).
- The input read files can have different file formats, as TopHat now autodetects the FASTA/FASTQ format of each input file.
- Fixed a bug that could sometimes incorrectly rename the reads in the output alignments.
- The stats in `align_summary.txt` now reflect the *reported* mappings under the constraints of the provided Tophat options, instead of reflecting the internally detected alignments. As such, the number of reads with multiple mappings may appear to be incorrectly reported if the user provided options that directly affect the reporting of such multiple mapppings.
- Fixed a bug that caused TopHat to fail when bowtie1 and pre-filtering options were used together.

Site Map

Home
Getting started
Manual
Index and annotation downloads
FAQ
Protocol

News and updates

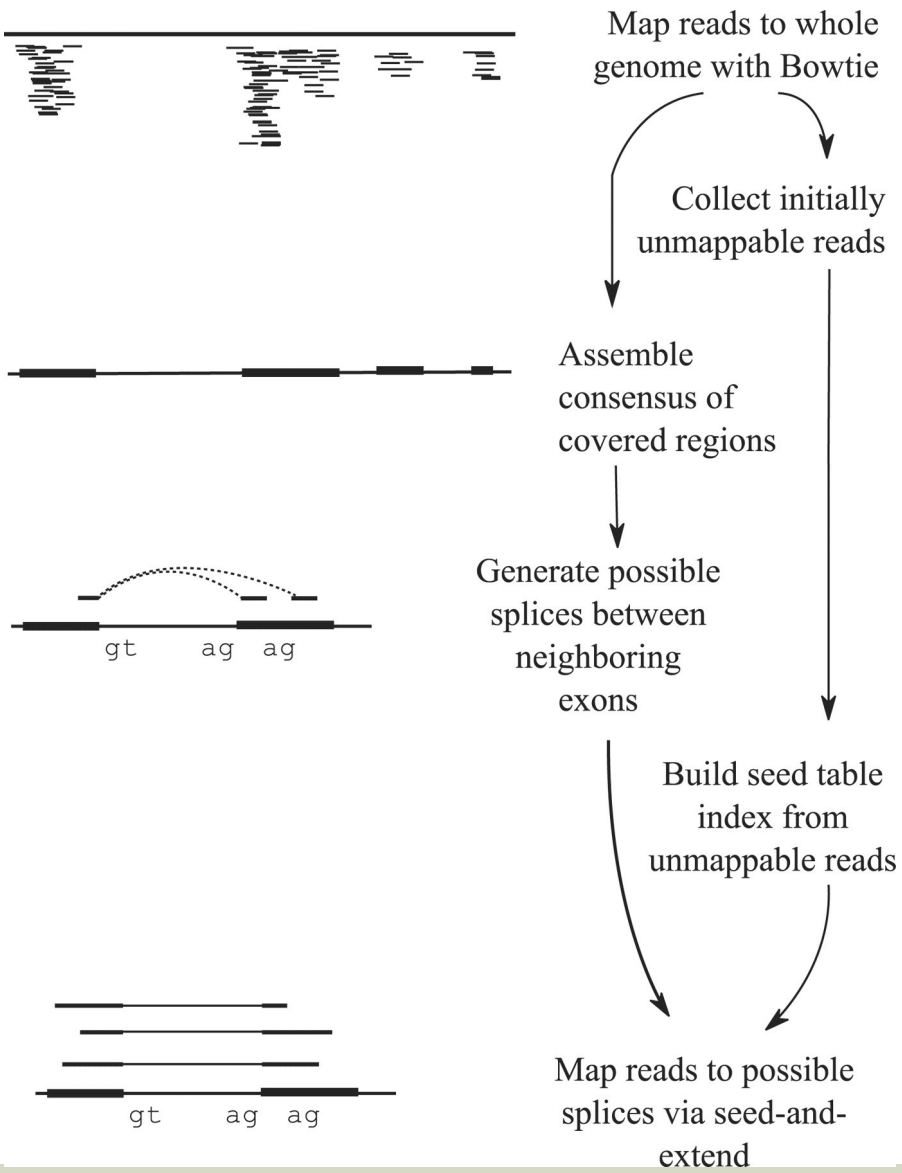
New releases and related tools will be announced through the Bowtie **mailing list**.

Getting Help

Questions and comments about TopHat can be posted on the **Tuxedo Tools Users Google Group**. Please use tophat.cufflinks@gmail.com for private

Trapnell et al., 2009

READ MAPPING



- TopHat (and the other spliced aligner) seeks “islands” in which reads accumulate on the genome.
- Then they seek reads that can join different islands, which corresponds to the exon-exon junctions.
- This way, it is possible to reconstruct the exon-intron structure of a gene and infer the splicing variants expressed by the gene in the analysed sample

READ MAPPING

Reference: Genome or transcriptome?

Genome:

- Requires a decent genome sequence
- Requires spliced alignments
- Can find novel (previously un-annotated) genes
- Can find novel exons/isoforms

Transcriptome:

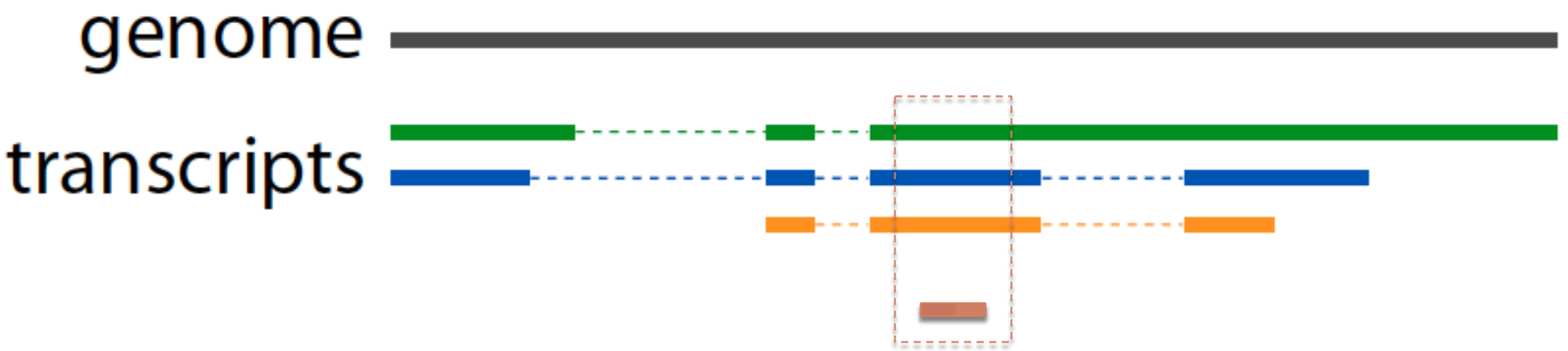
- Sometimes the genome is unknown, but there are cDNA libraries
- Spliced alignment is not necessary
- Multireads are an issue
- Cannot find what is unknown

READ MAPPING

Multireads: reads that can be mapped in different places

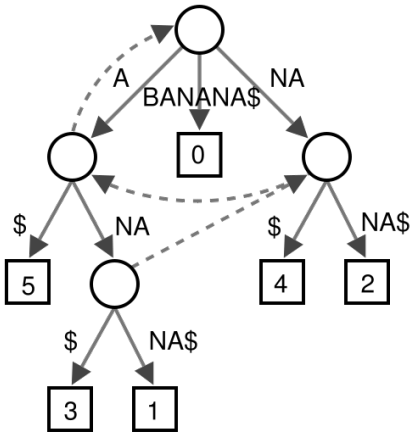
They can origniate from:

- By chance (especially if the read is short)
- Due to paralogs sharing domains (when mapping to the genome)
- Due to splicing variants sharing an exon or part of it (when mapping to the transcriptome)



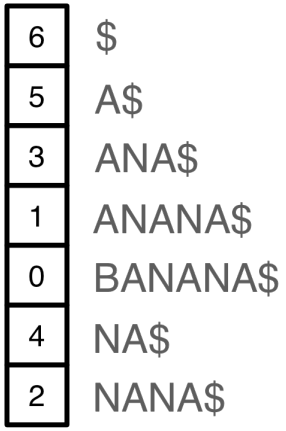
GENOME INDEXING

Genomes are too large for direct alignment of reads onto them. They need to be converted into a compact and quickly accessible form



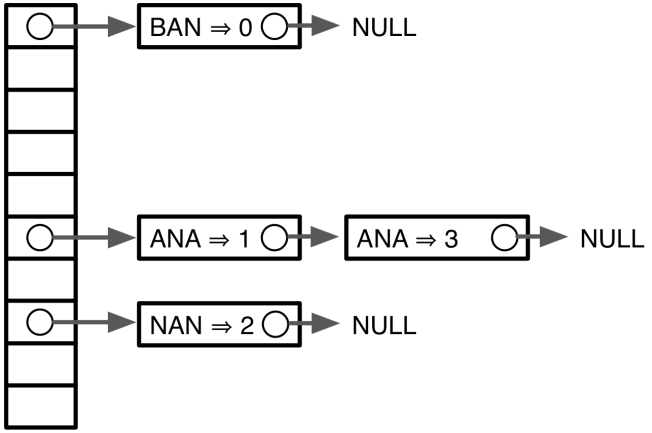
Suffix tree

> 35 GBs



Suffix array

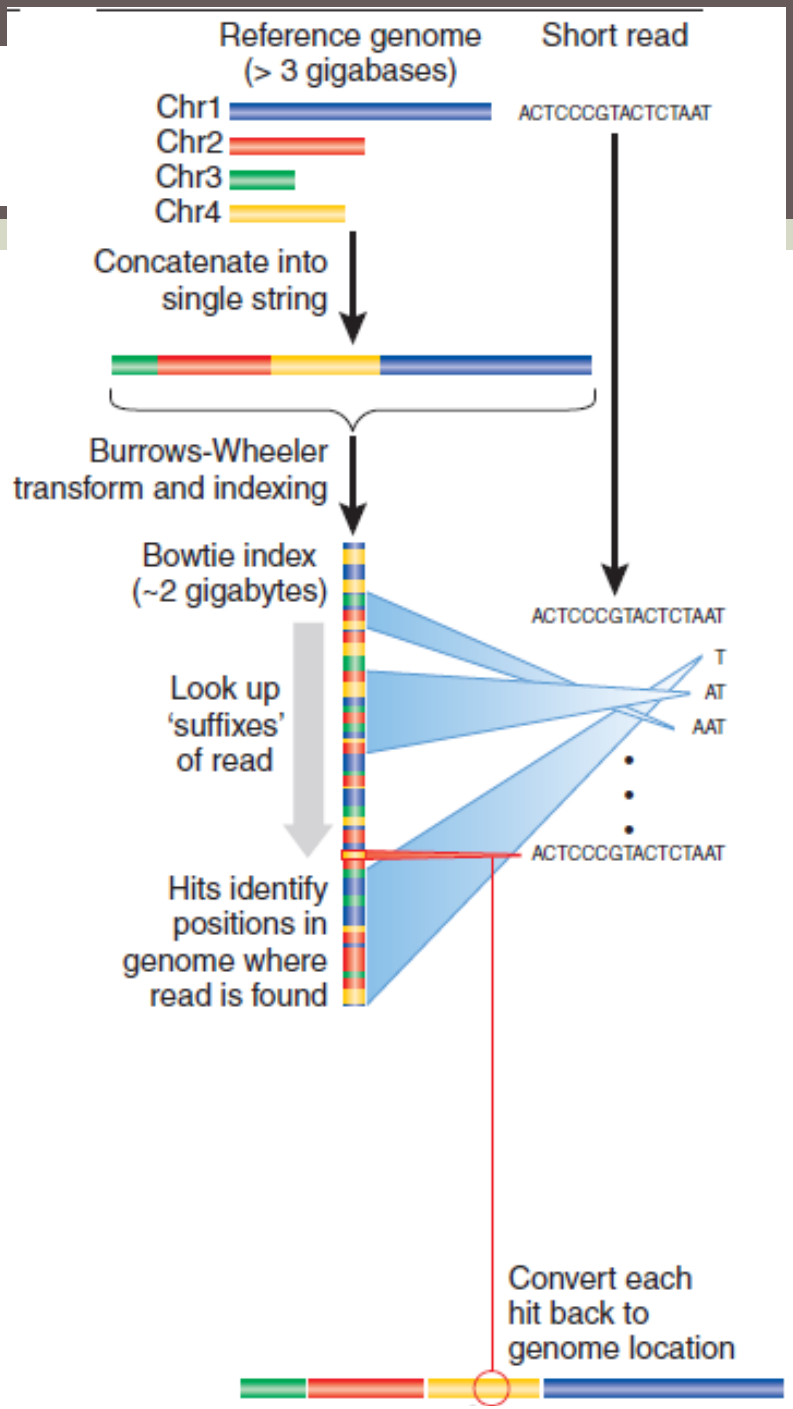
> 12 GBs



Seed hash tables

Many variants, incl. spaced seeds

> 12 GBs



The bowtie algorithm implements a Burrows-Wheeler transform (BWT) to index and compress the genome

Trapnell & Salzberg, Nature Biotech 2009

GENOME INDEXING

Transformation			
Input	**All Rotations**	**Sort the Rows**	**Output**
^BANANA@	^BANANA@ @^BANANA A@^BANAN NA@^BANA ANA@^BAN NANA@^BA ANANA@^B BANANA@^	ANANA@^**B** ANA@^BA**N** A@^BANA**N** BANANA@**^** NANA@^B**A** NA@^BAN**A** ^BANANA@ @^BANAN**A**	BNN^AA@A

[Source: Wikipedia]

GENOME INDEXING

Inverse Transformation

Input

```
BNN^AA@A
```

Add 1	Sort 1	Add 2	Sort 2
B N N ^ A A @ A	A A A B N N ^ @	BA NA NA ^B AN AN @^ A@	AN AN A@ BA NA NA ^B @^
Add 3	**Sort 3**	**Add 4**	**Sort 4**
BAN NAN NA@ ^BA ANA ANA @^B A@^	ANA ANA A@^ BAN NAN NA@ ^BA @^B	BANA NANA NA@^ ^BAN ANAN ANA@ @^BA A@^B	ANAN ANA@ A@^B BANA NANA NA@^ ^BAN @^BA
Add 5	**Sort 5**	**Add 6**	**Sort 6**
BANAN NANA@ NA@^B ^BANA ANANA ANA@^ @^BAN A@^BA	ANANA ANA@^ A@^BA BANAN NANA@ NA@^B ^BANA @^BAN	BANANA NANA@^ NA@^BA ^BANAN ANANA@ ANA@^B @^BANA A@^BAN	ANANA@ ANA@^B A@^BAN BANANA NANA@^ NA@^BA ^BANAN @^BANA
Add 7	**Sort 7**	**Add 8**	**Sort 8**
BANANA@ NANA@^B NA@^BAN ^BANANA ANANA@^ ANA@^BA @^BANAN A@^BANA	ANANA@^ ANA@^BA A@^BANA BANANA@ NANA@^B NA@^BAN ^BANANA @^BANAN	BANANA@^ NANA@^BA NA@^BANA ^BANANA@ ANANA@^B ANA@^BAN @^BANANA A@^BANAN	ANANA@^B ANA@^BAN A@^BANAN BANANA@^ NANA@^BA NA@^BANA ^BANANA@ @^BANANA

Output

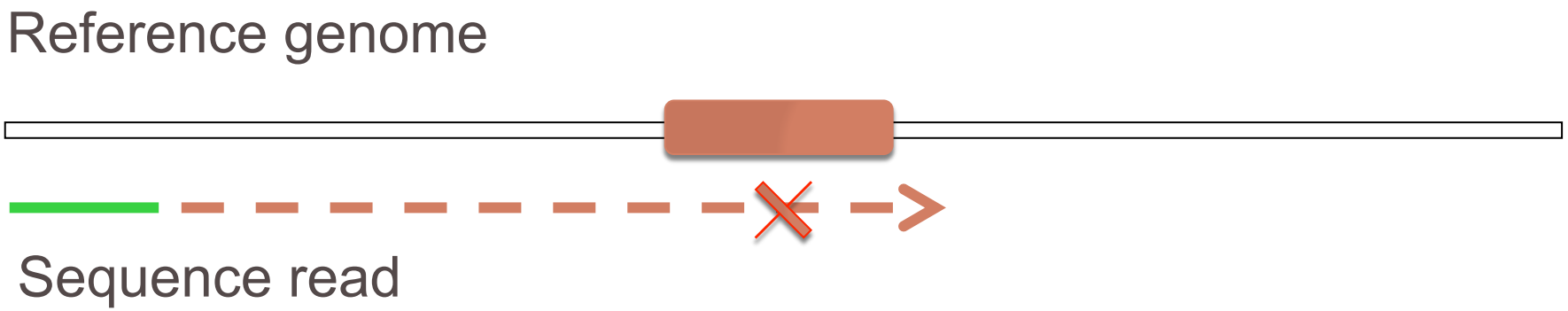
```
^BANANA@
```

[Source: Wikipedia]

GENOME INDEXING

- BWT(T) is reversible, meaning that it is always possible to reconstruct the original sequence
- It can be also compressed, since identical characters tend to end up close in the transformed sequence (for example the human genome can be stored in around 1-2 Gb)
- It allows the fast lookup of sub-sequences (for example a read)
- Once a good match is found, you can quickly know its coordinates in the original sequences (i.e. its genomic coordinates)
- It can be adapted for imperfect matches

READ MAPPING

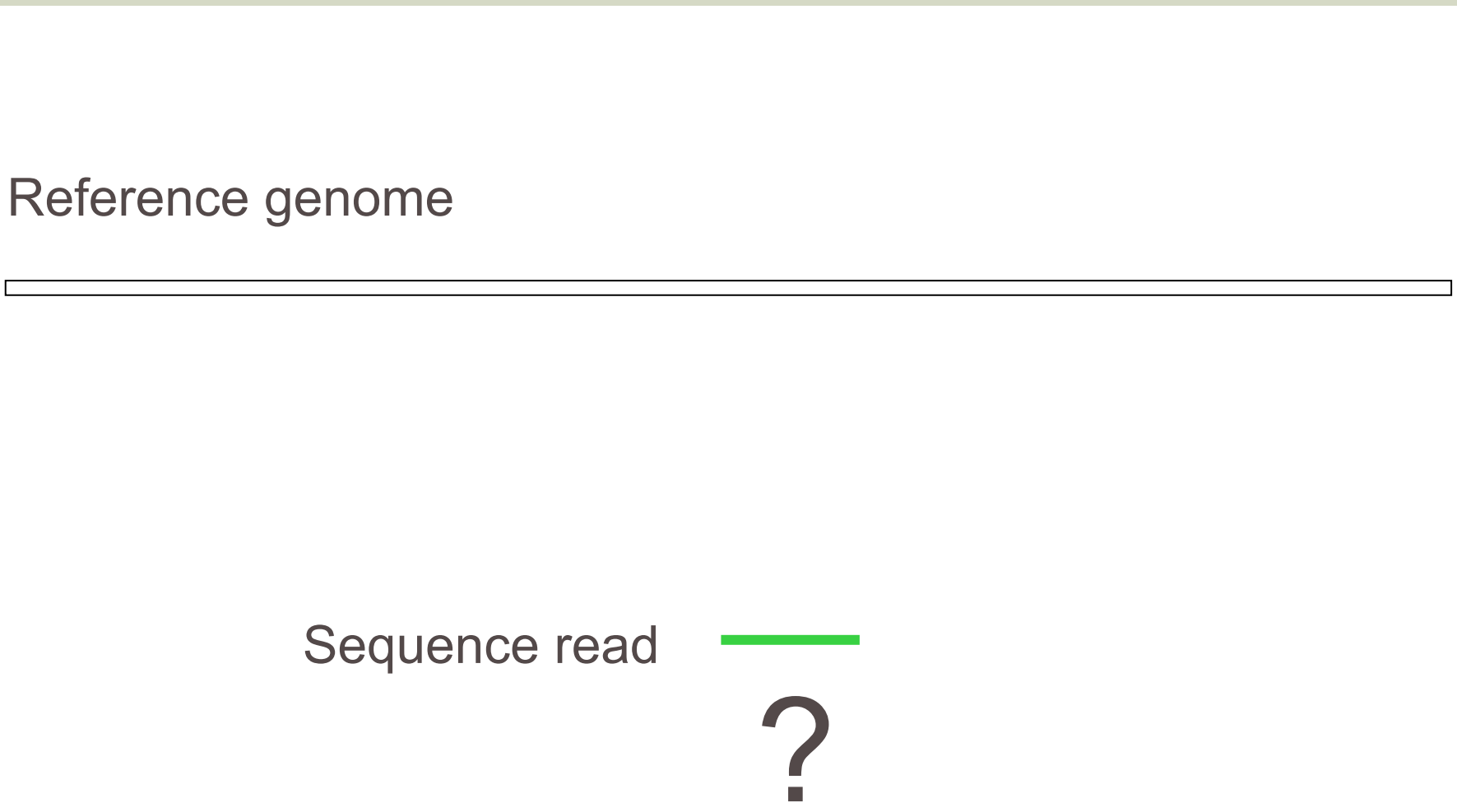


- The read sequence could be not identical to the corresponding transcribed locus in the reference genome, because of sequencing errors and variations between the reference genome sequence and the genomic sequence of the sample

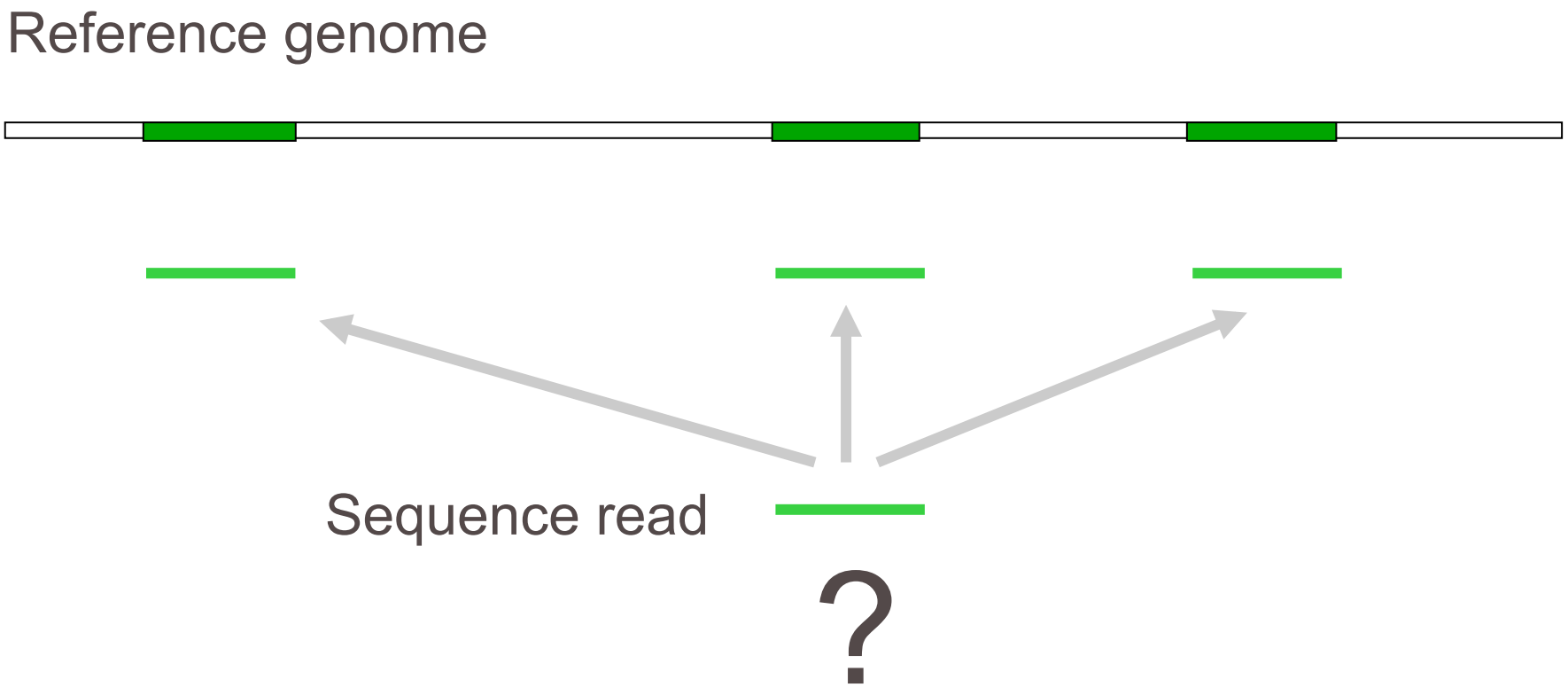
READ MAPPING

- A possibly large number of reads cannot be mapped in an unambiguous way, meaning that there could be more than one genome locus to which the read can be mapped equally well (imagine the sky in the puzzle analogy)
- The shorter the read is, the more likely is that there is some other genomic region other than its origin locus to which the read can be mapped
- This is even more true if we tolerate imperfect matching during the mapping (and we must do it to compensate for sequencing errors and genetic variation)
- Paired ended sequencing can help in solving such mapping ambiguities

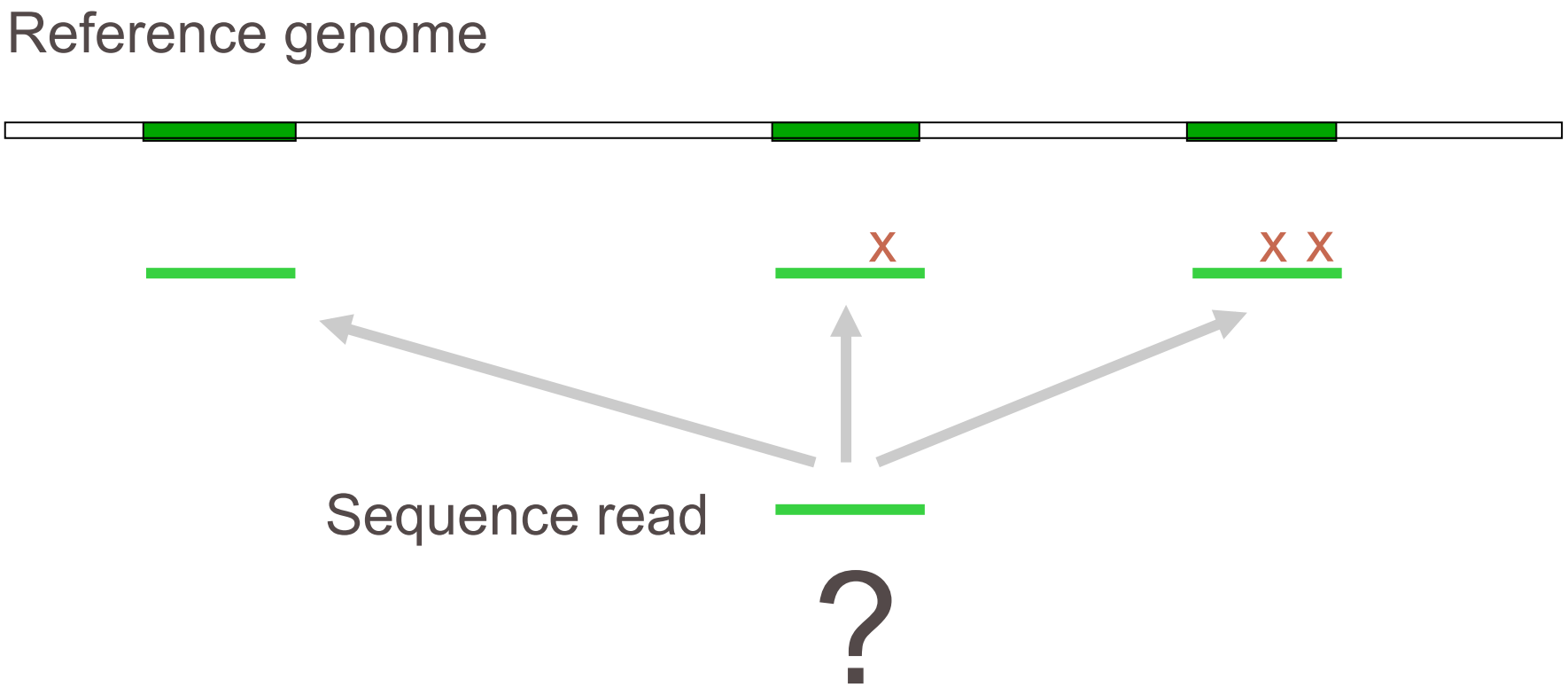
READ MAPPING



READ MAPPING

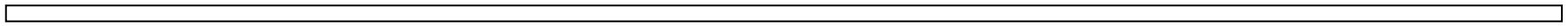


READ MAPPING



READ MAPPING

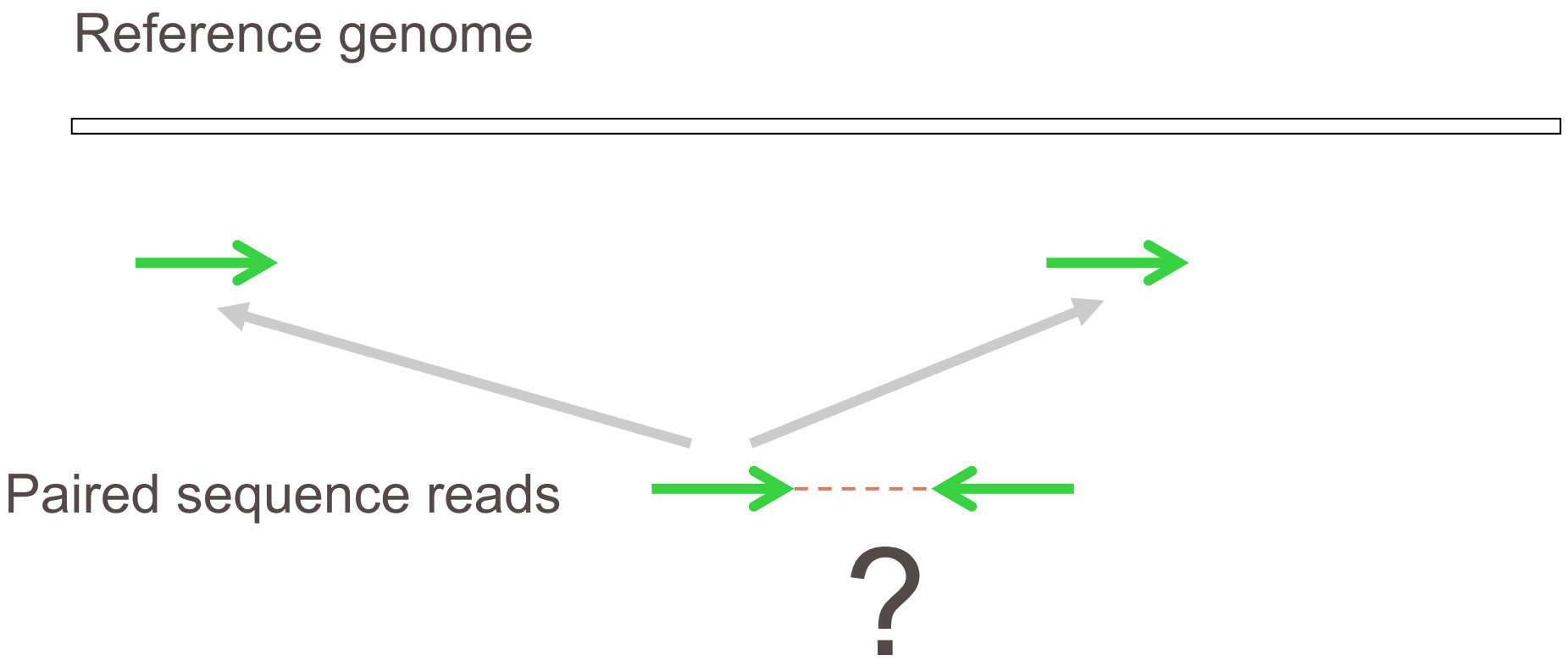
Reference genome



Paired sequence reads

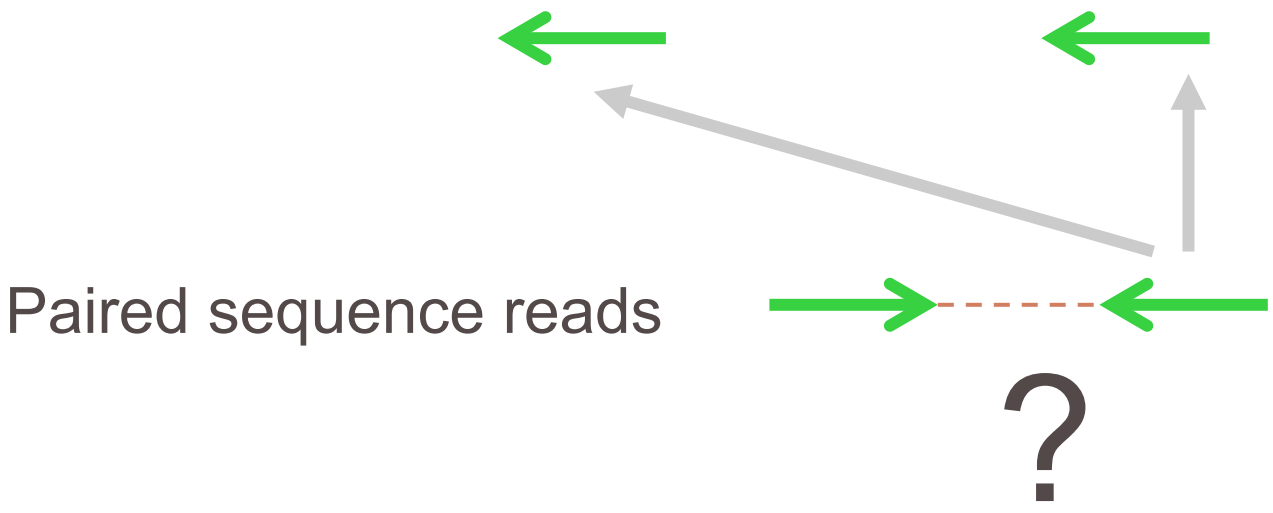
?

READ MAPPING

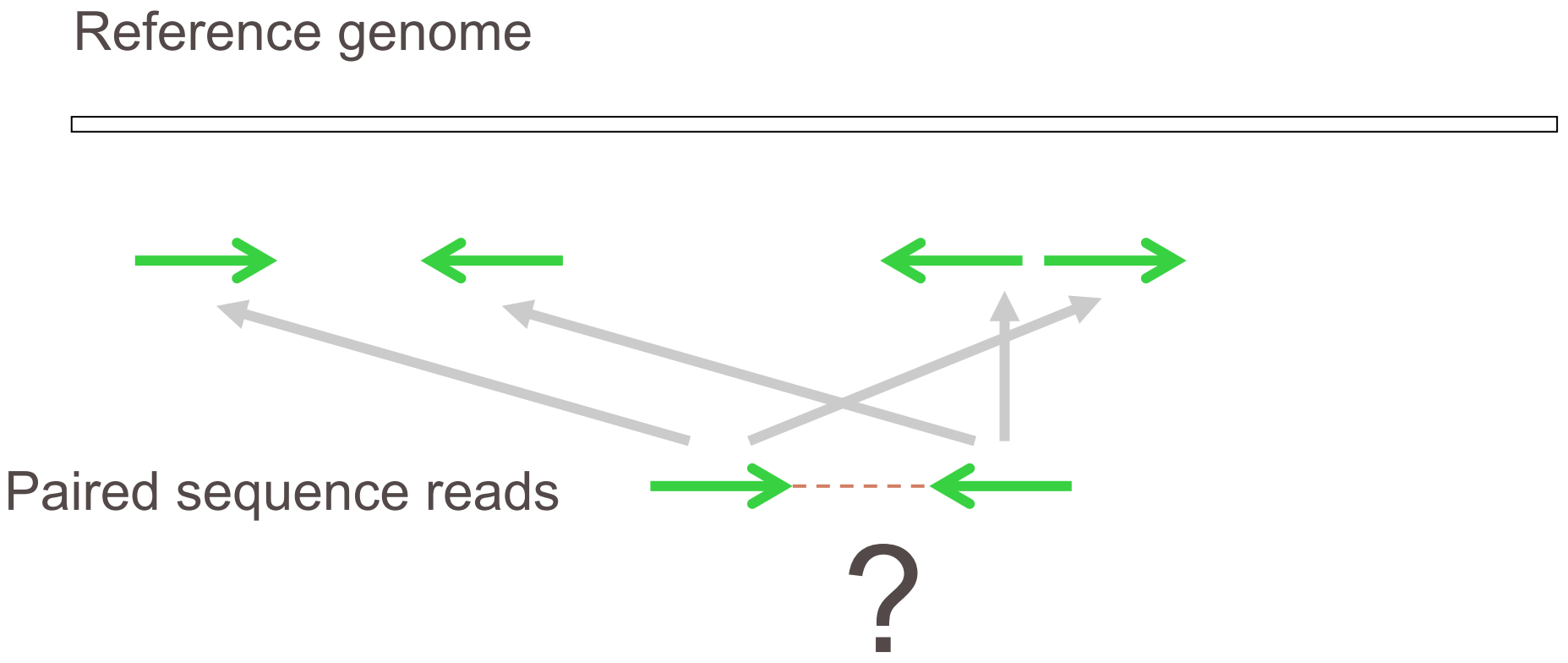


READ MAPPING

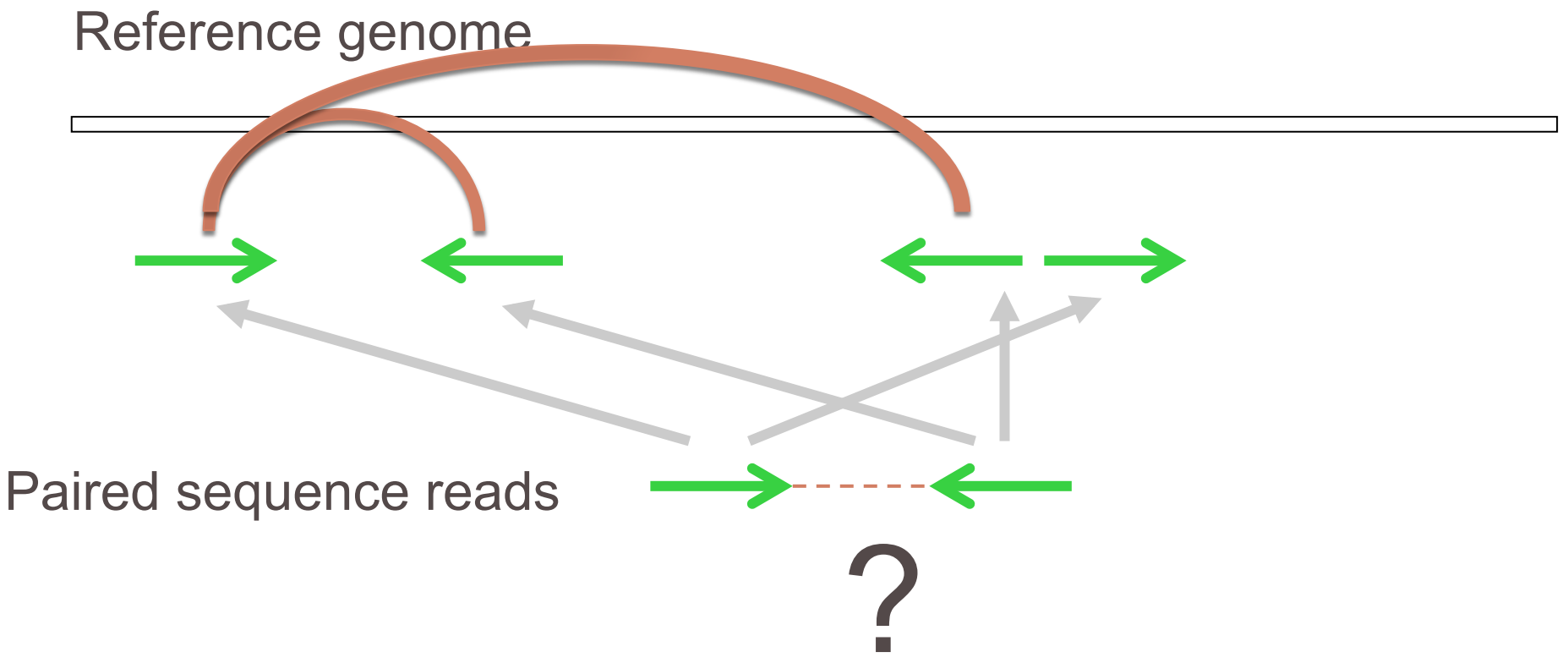
Reference genome



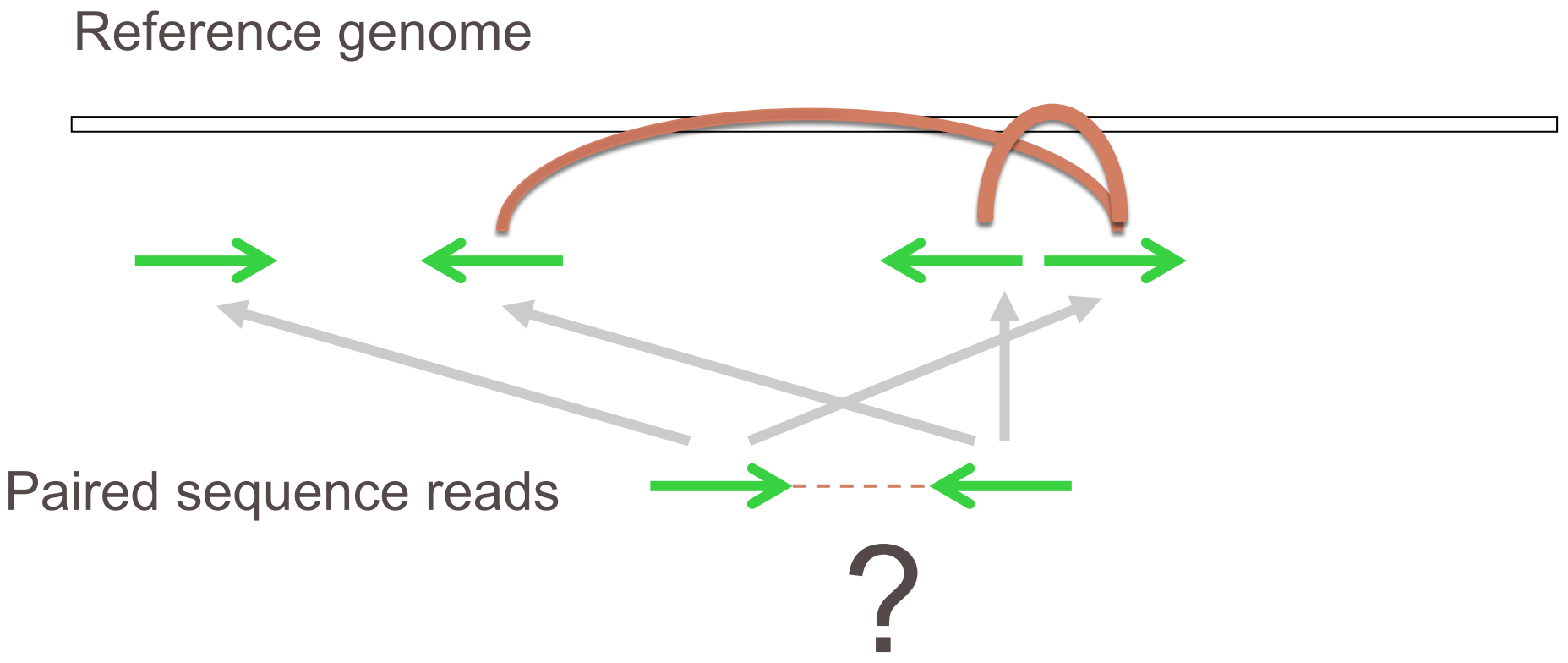
READ MAPPING



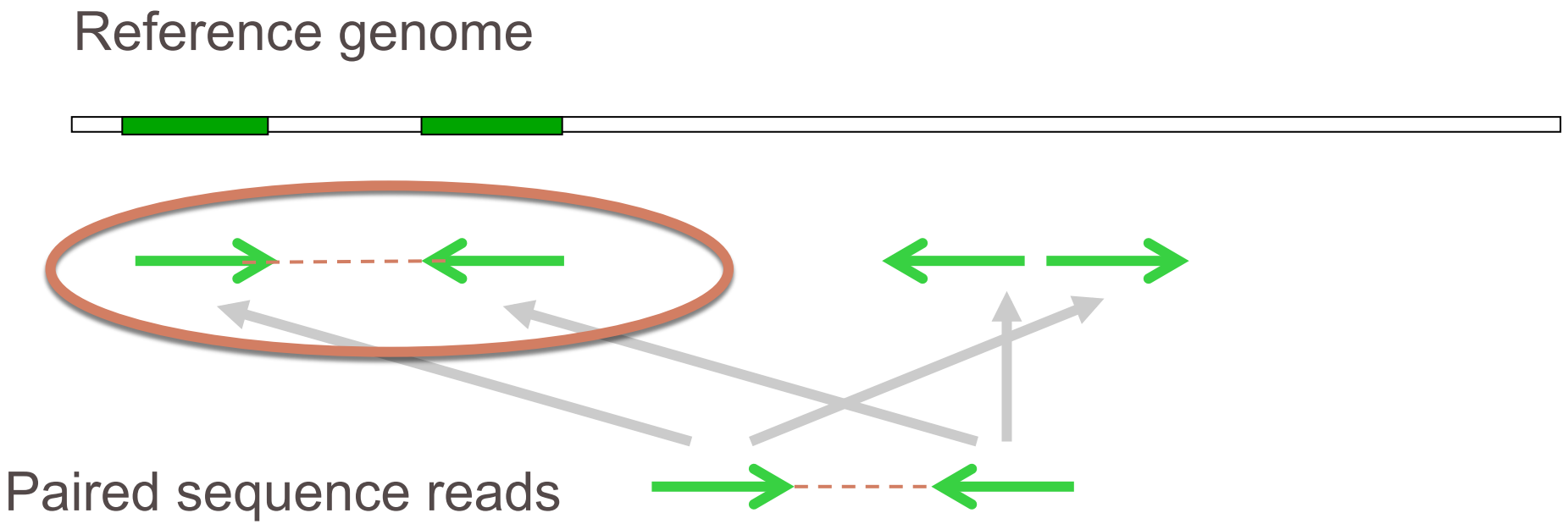
READ MAPPING



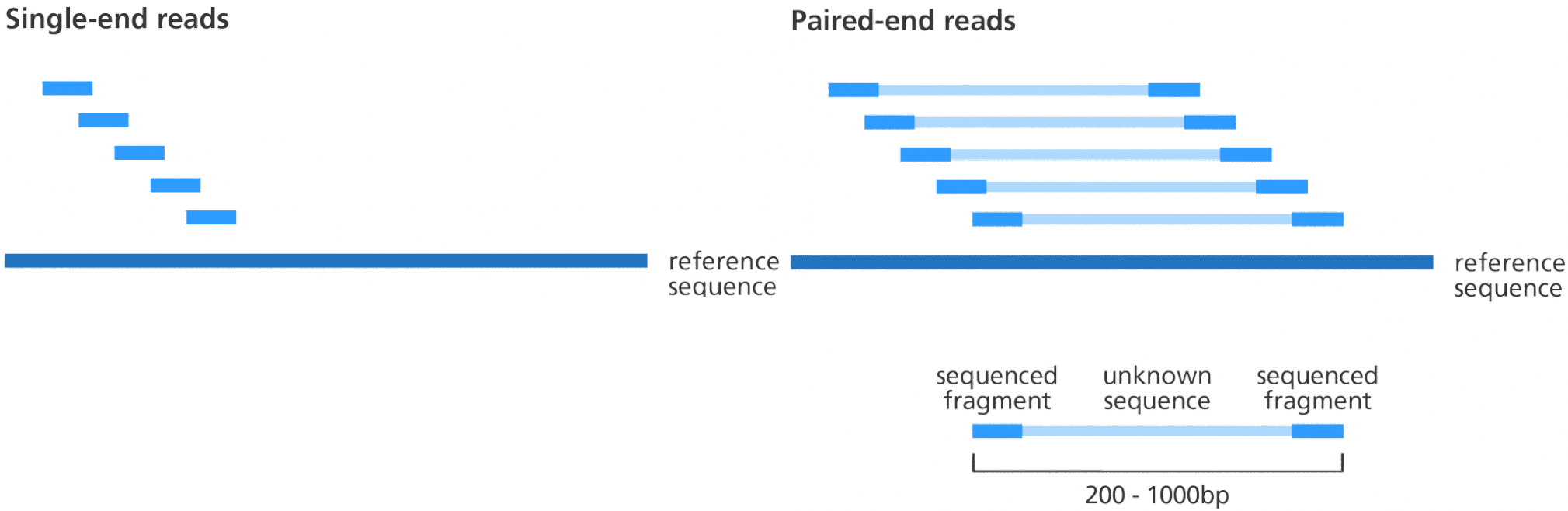
READ MAPPING



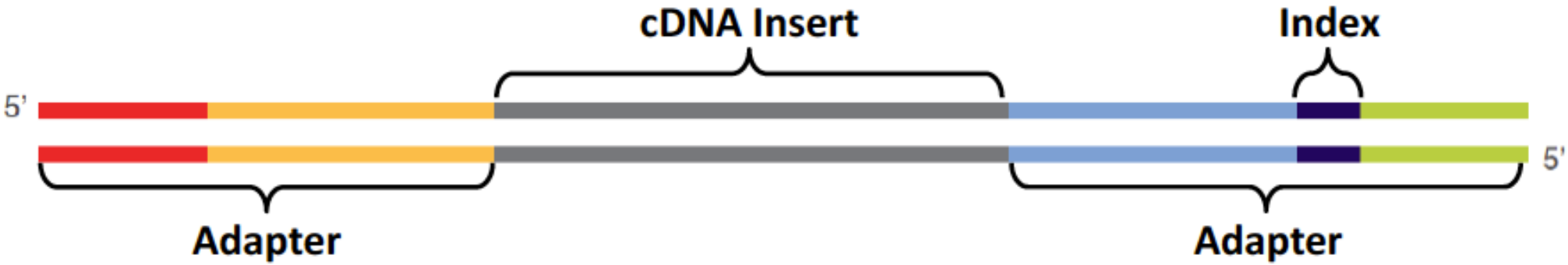
READ MAPPING



INSERT SIZE

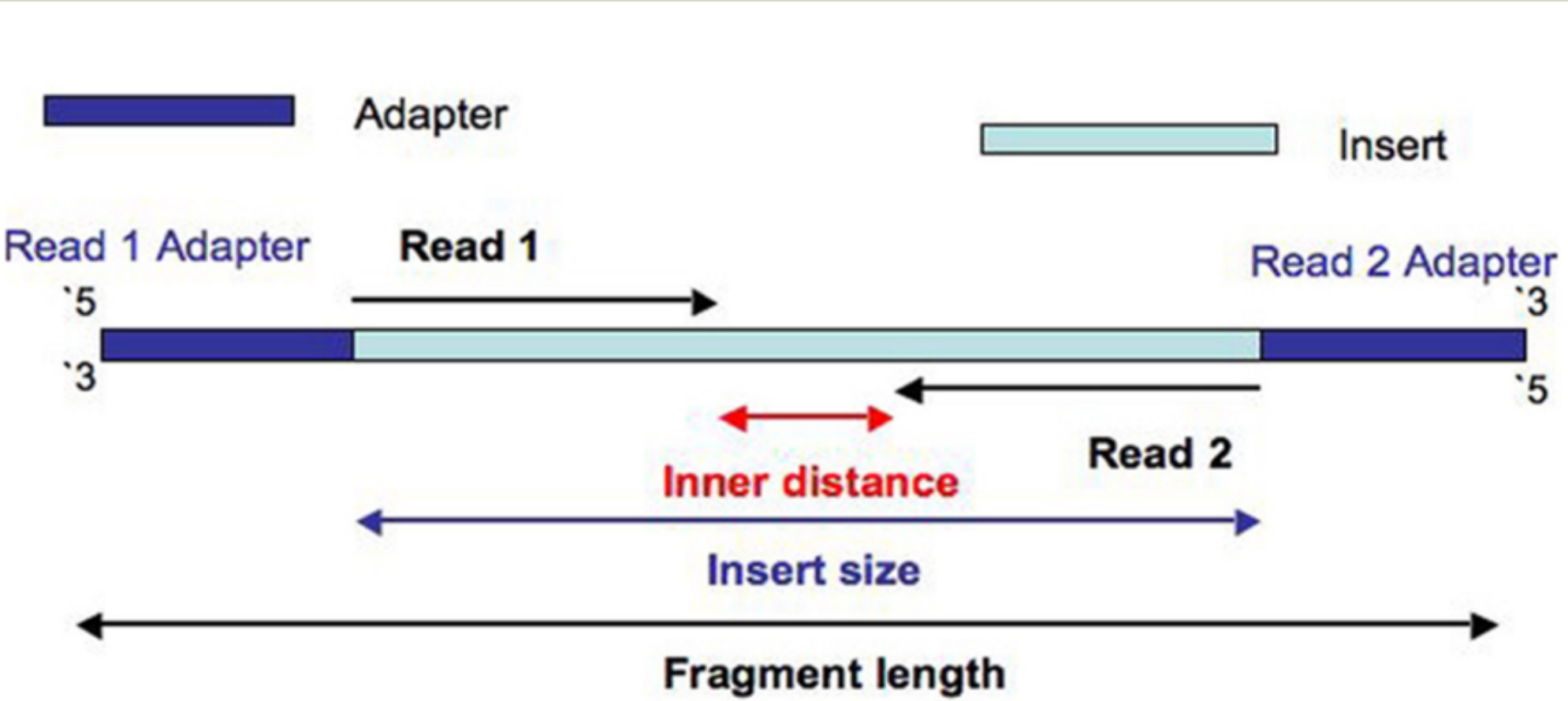

Single-end reads
reference sequence
Paired-end reads
reference sequence
sequenced fragment
unknown sequence
sequenced fragment
200 - 1000bp

INSERT SIZE

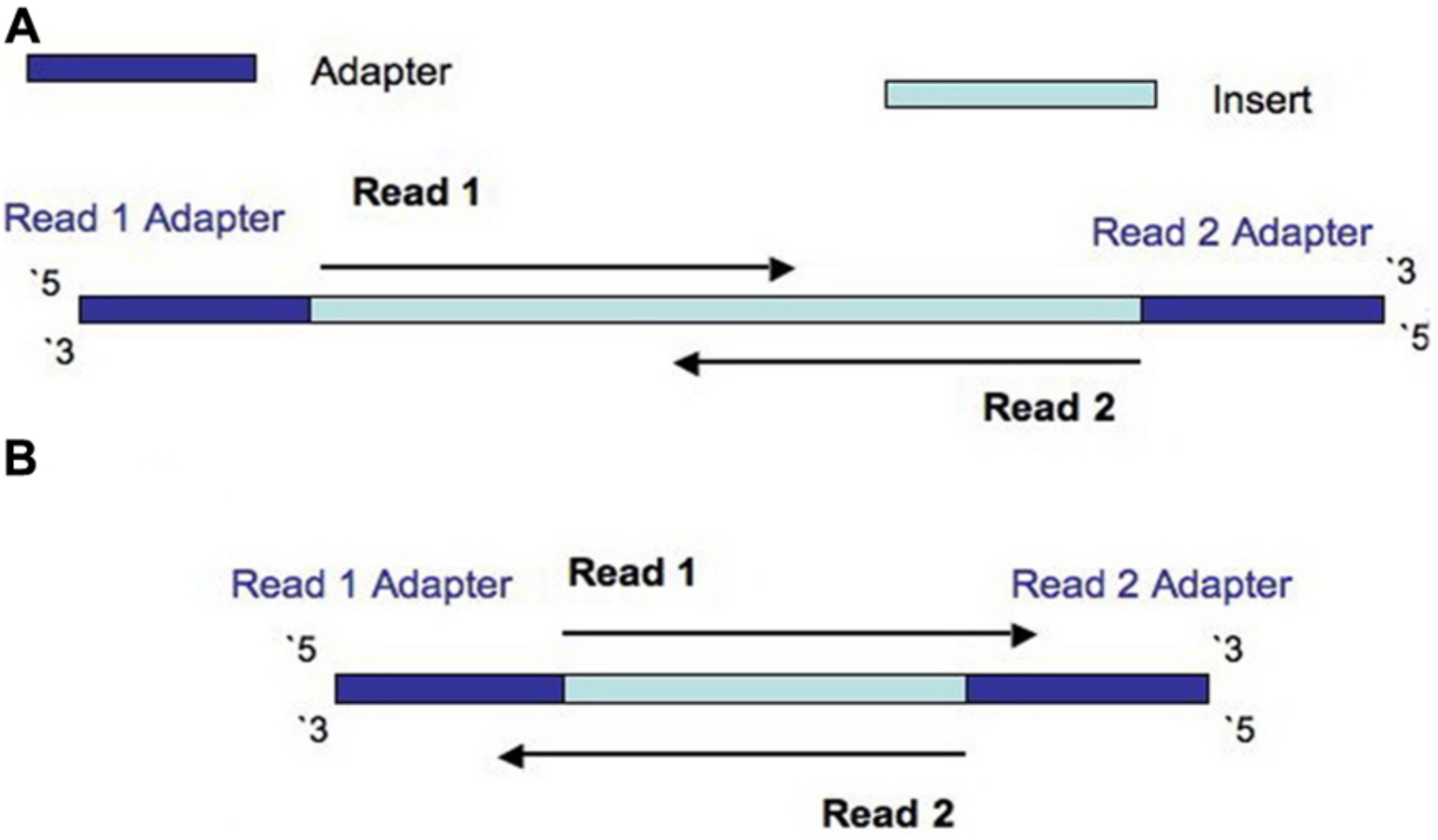


- In order for this to work, we need to know, at least approximately, how far the paired reads can be
- Since during sample preparation fragments were selected by size, all cDNAs should have similar size, which should be known by whomever performed the preparation
- For public datasets one can estimate the expected distance between paired reads, by mapping reads (the whole dataset or a subset) to the genome, computing the distribution of the distance between paired reads, its mean and its standard deviation
- These parameters can be then given to the mapping software for a more accurate mapping

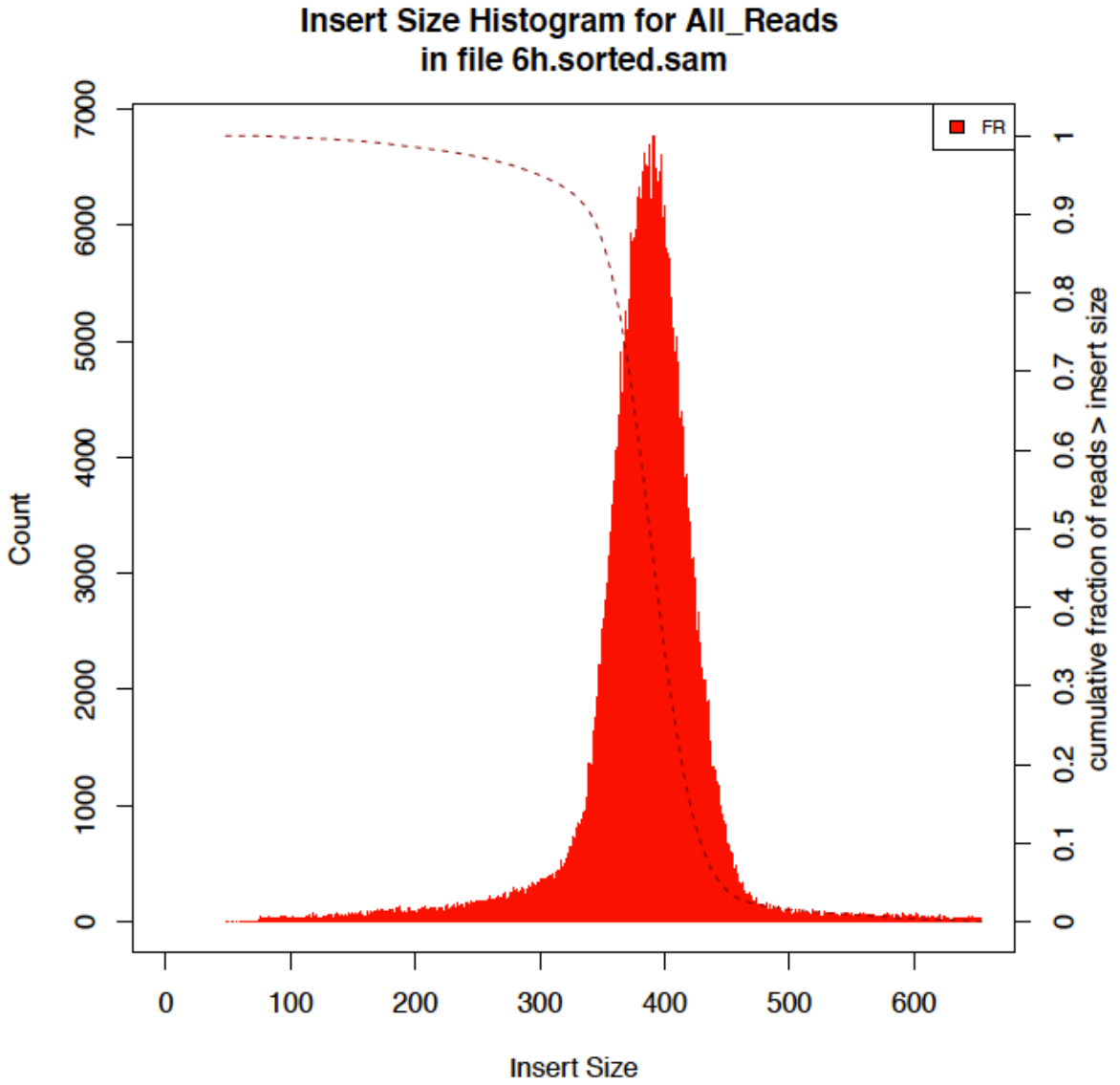
INSERT SIZE


Adapter
Insert
Read 1 Adapter
Read 1
Read 2 Adapter
`5
`3
`3
`5
Read 2
Inner distance
Insert size
Fragment length

INSERT SIZE

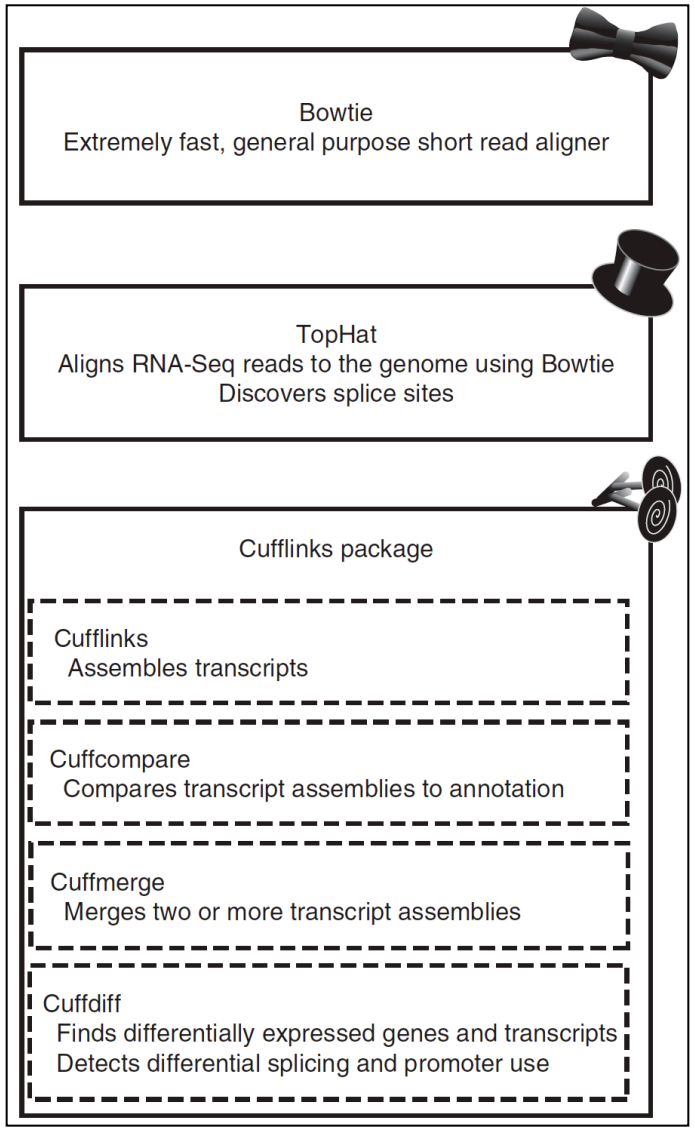

A
Adapter
Insert
Read 1
Read 1 Adapter
Read 2 Adapter
`5
`3
`3
`5
Read 2
B
Read 1 Adapter
Read 1
Read 2 Adapter
`5
`3
`3
`5
Read 2

INSERT SIZE


Insert Size Histogram for All_Reads
in file 6h.sorted.sam
FR
Count
0
1000
2000
3000
4000
5000
6000
7000
0
100
200
300
400
500
600
Insert Size
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
cumulative fraction of reads > insert size

READ MAPPING

Options for DGE analysis (tuxedo suite)



Bowtie/Bowtie2 use Burrows-Wheeler indexing for aligning reads (not splice-aware)

TopHat/TopHat2 utilizes either Bowtie or Bowtie2 to align reads in a splice-aware manner and aids the discovery of new splice junctions

The Cufflinks package has 4 components, the 2 major ones are listed below:

Cufflinks does reference-based or unguided transcriptome assembly

Cuffdiff does **DGE** (Differential Gene Expression) analysis in a simple pairwise comparison, and a series of pairwise comparisons in a time-course experiment

READ MAPPING

Old Tuxedo suite

- **Bowtie/Bowtie2** - general aligner
- **Tophat/Tophat2** - splice-aware aligner
- **Cufflinks suite** - Transcriptome assembly, gene counting, and DGE calculation for simple models
- This suite is well-known, but there are a few issues with methods used
- No longer being supported!

NEW Tuxedo suite

- **HISAT2** - DNA & RNA aligner in one
- **StringTie** - Transcriptome assembly, & gene counting
- **Ballgown** - Improved DGE calculation for simple models, and visualization
- This suite is very new and attempts to improve on the old suite's issues
- However, not well-tested on anything but human at the moment

READ MAPPING

Simulation-based comprehensive benchmarking of RNA-seq aligners

Giacomo Baruzzo[1,5], Katharina E Hayer[2,5], Eun Ji Kim[2], Barbara Di Camillo[1], Garret A FitzGerald[2,3] & Gregory R Grant[2,4]

[1]Department of Information Engineering, University of Padova, Padua, Italy. [2]Institute for Translational Medicine and Therapeutics (ITMAT), University of Pennsylvania, Philadelphia, Pennsylvania, USA. [3]Department of Systems Pharmacology and Translational Therapeutics, University of Pennsylvania, Philadelphia, Pennsylvania, USA. [4]Department of Genetics, University of Pennsylvania, Philadelphia, Pennsylvania, USA. [5]These authors contributed equally to this work. Correspondence should be addressed to G.R.G. (ggrant@upenn.edu).

RECEIVED 18 APRIL; ACCEPTED 15 NOVEMBER; PUBLISHED ONLINE 12 DECEMBER 2016; DOI:10.1038/NMETH.4106

READ MAPPING

- Several read mappers were tested on simulated reads generated for human and *Plasmodium falciparum* genes and genomes
- For each genome, three datasets were built:
  - T1: low polymorphism rate, low indel rate, 0.005 sequencing error rate (similar to what one would expect from a typical RNA-Seq experiment on human)
  - T2: : medium polymorphism rate, medium indel rate, 0.01 sequencing error rate
  - T3: : high polymorphism rate, high indel rate, 0.02 sequencing error rate

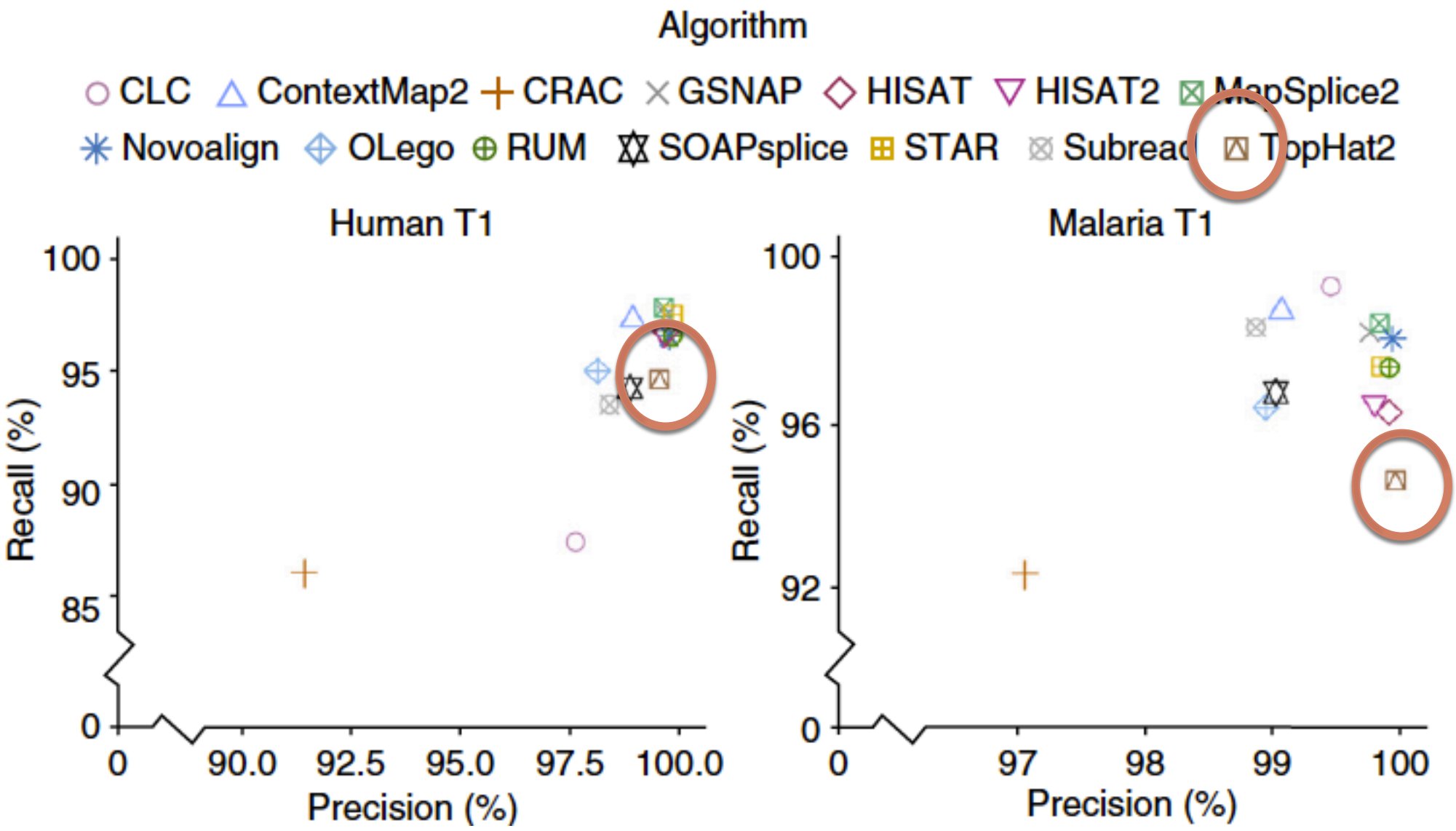
Baruzzo et al., 2016

READ MAPPING

- Alignment accuracy was measured at base level and read level. At base level, the mapping can only be right (the read base is mapped where it should be) or wrong. At read level, a read is considered well mapped if they are not multimappers and at least one base is aligned correctly.
- Accuracy is estimated through two metrics, precision and recall
  - Precision: fraction of all bases (or reads) that were aligned correctly
  - Recall: fraction of aligned bases (or reads) that were aligned correctly

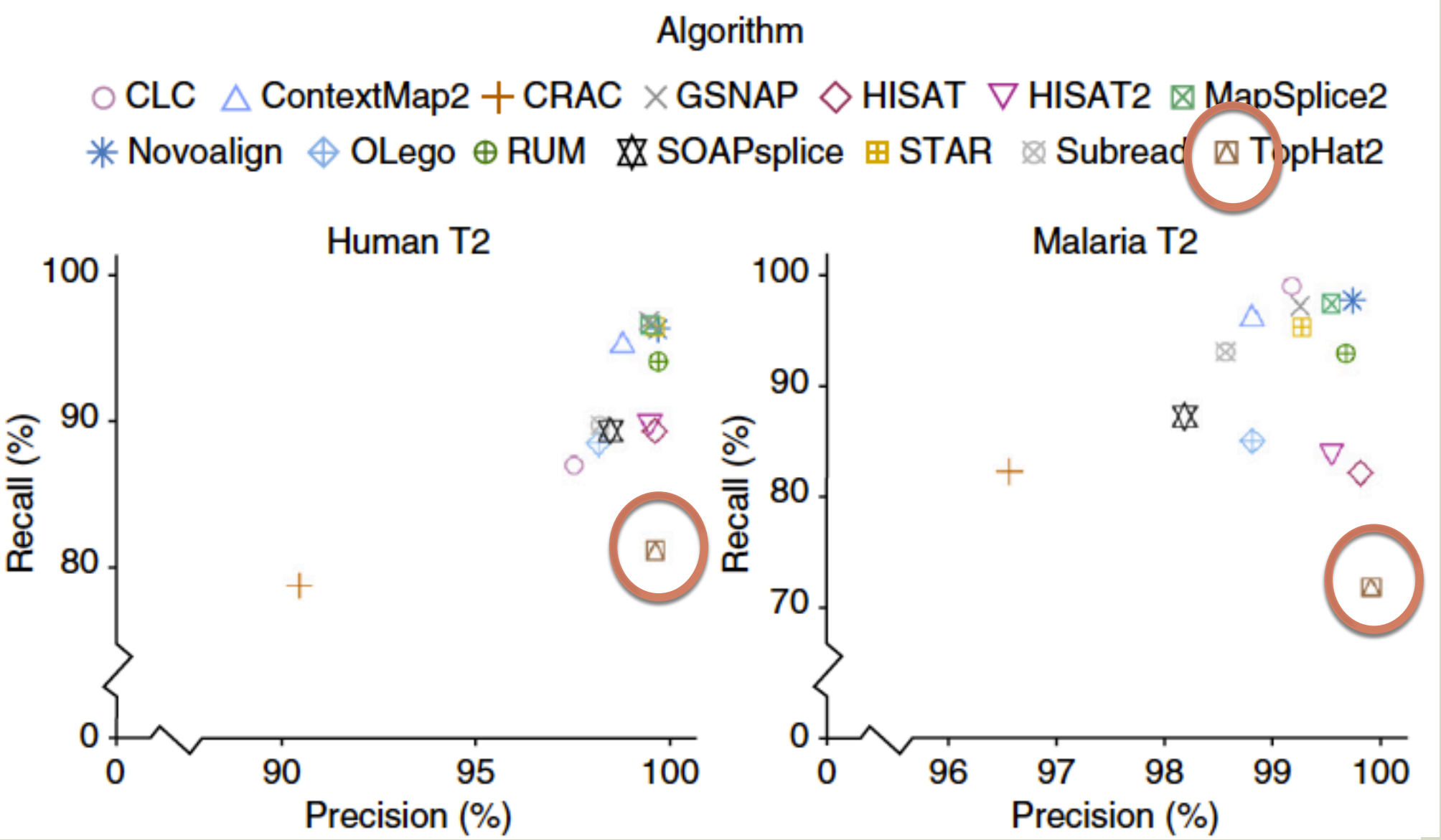
Baruzzo et al., 2016

READ MAPPING



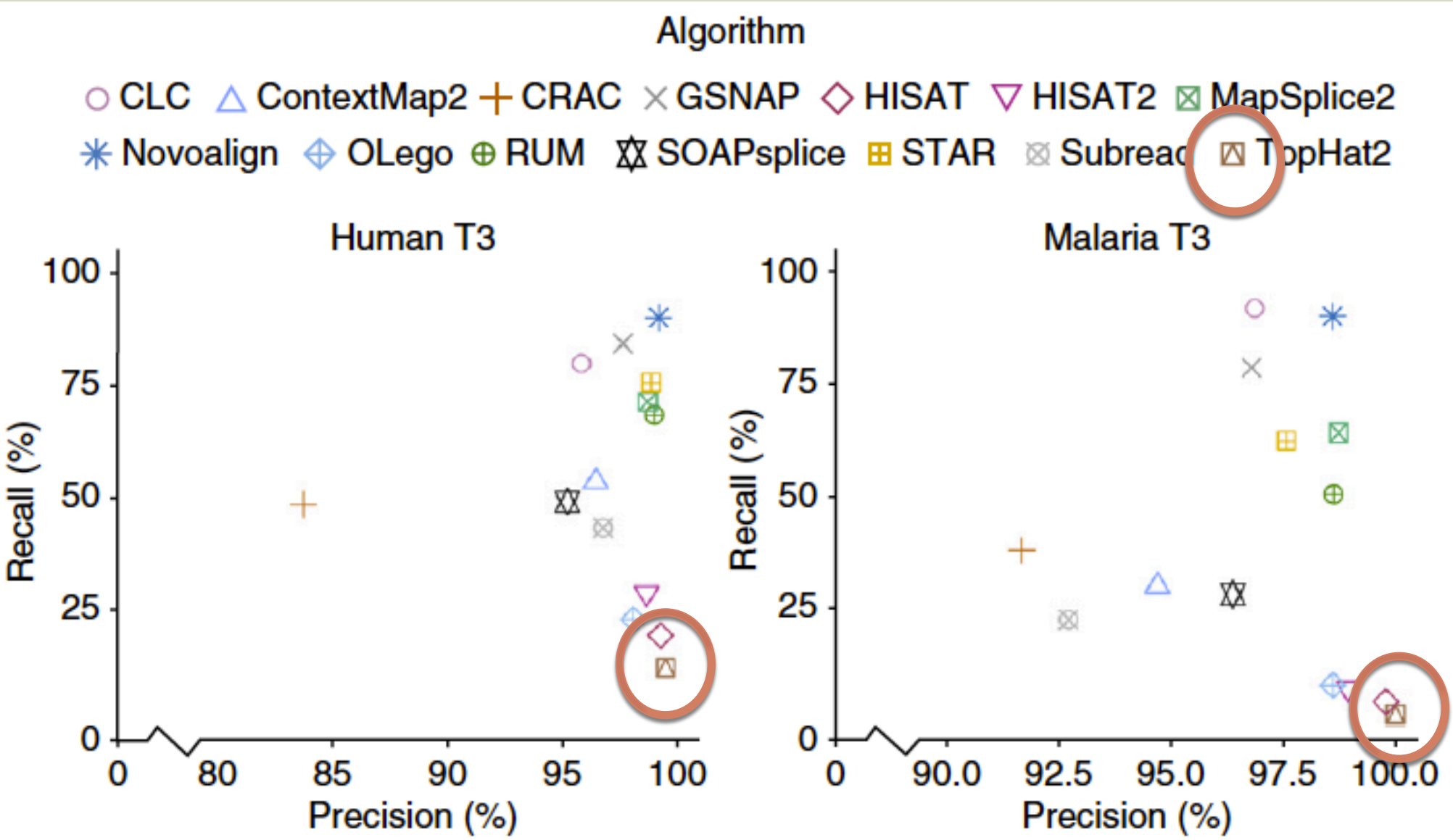
Baruzzo et al., 2016

READ MAPPING



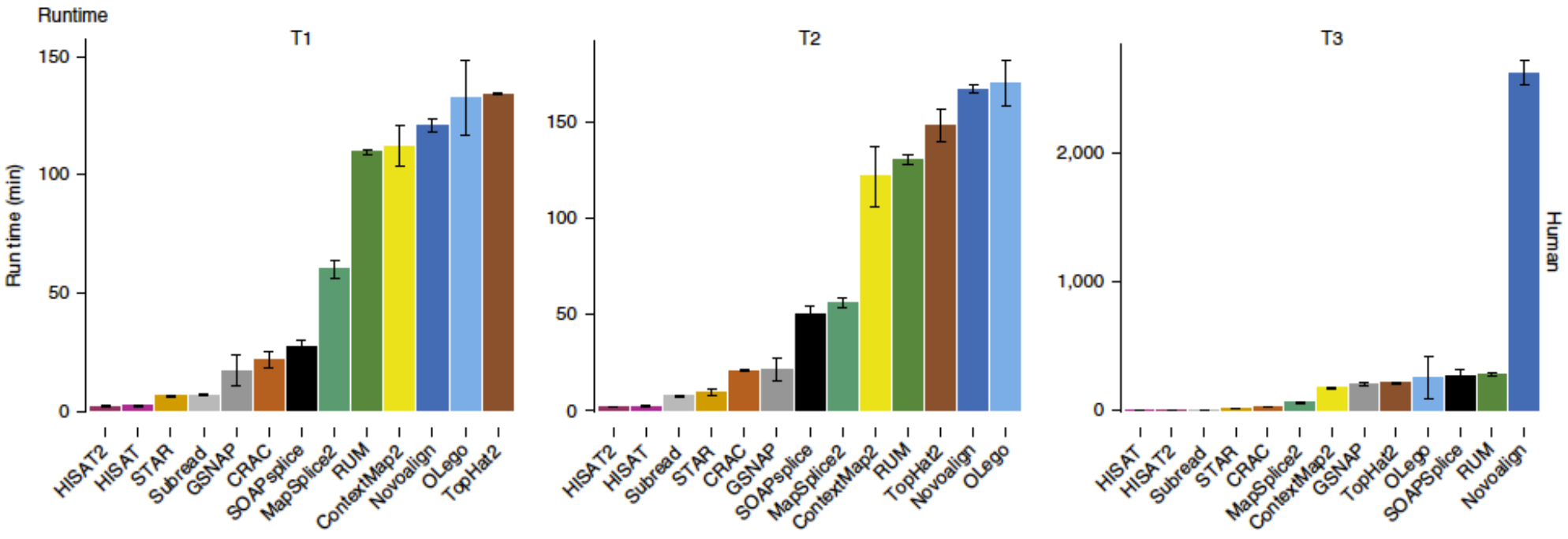
Baruzzo et al., 2016

READ MAPPING



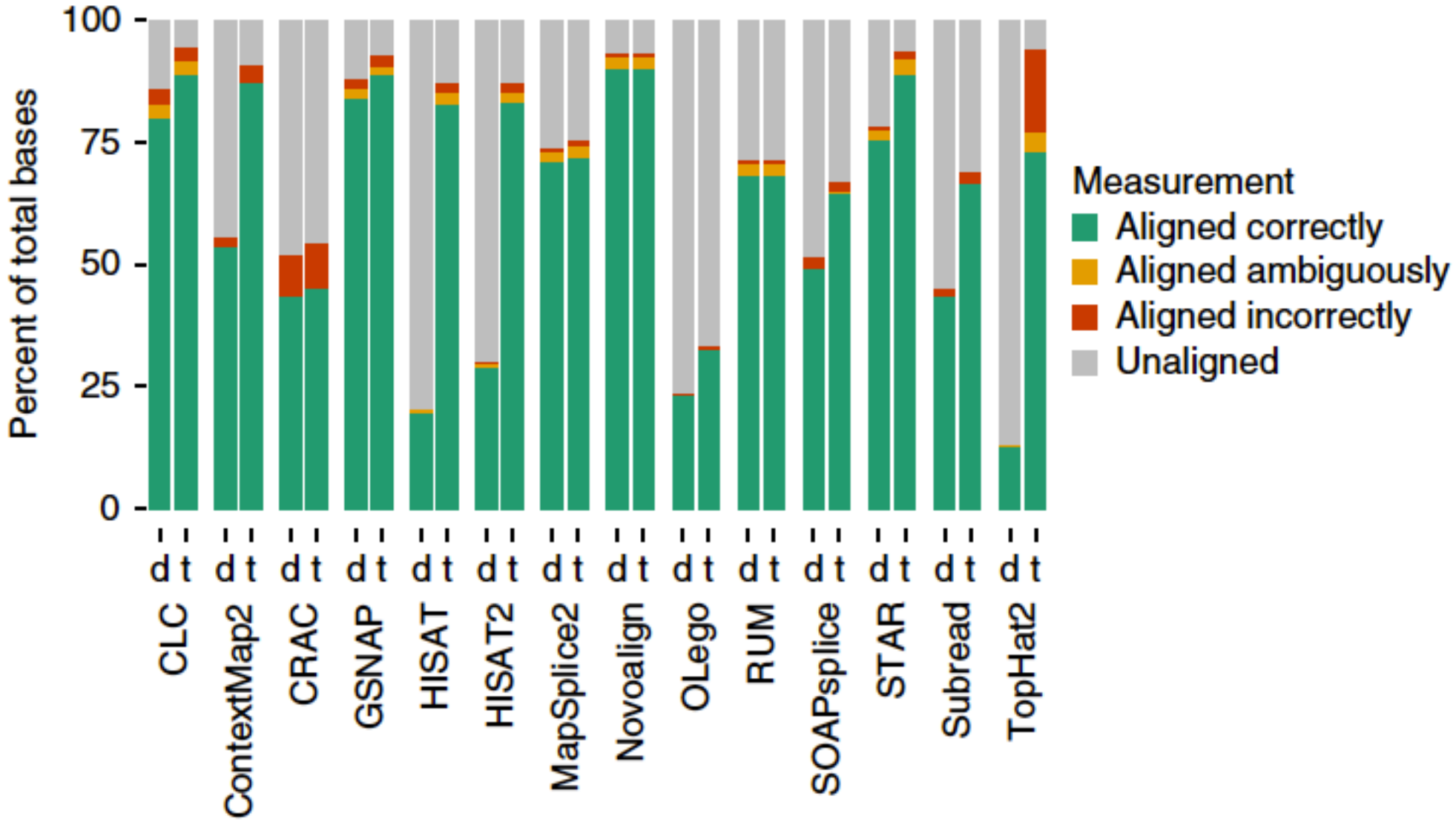
Baruzzo et al., 2016

READ MAPPING



Baruzzo et al., 2016

READ MAPPING



Baruzzo et al., 2016

READ MAPPING

"Despite its popularity, TopHat2 is consistently among the worst performers on both human and malaria T2 and T3 libraries"
Baruzzo et al., Nature Methods 2016

So why use TopHat?

- It's still the most popular, hence it's easier to get help if something went wrong
- It's easy to install and run
- It's still good on "normal" datasets
- If you play with parameters it can work well also on difficult datasets
- It's what referees expect

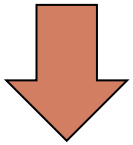
Nevertheless, remember that any mapper (for example STAR, that we'll use later) can be plugged in a pipeline, since they all report alignment using the same format (the BAM/SAM format)

Baruzzo et al., 2016

THE SAM (SEQUENCE ALIGNMENT MAP) FORMAT

```
Coor     12345678901234  5678901234567890123456789012345
ref      AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT

+r001/1        TTAGATAAAGGATA*CTG
+r002         aaaAGATAA*GGATA
+r003       gcctaAGCTAA
+r004                       ATAGCT..............TCAGC
-r003                             ttagctTAGGC
-r001/2                                        CAGCGCCAT
```



```
@HD VN:1.3 SO:coordinate
@SQ SN:ref LN:45
r001 163 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002   0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA    *
r003   0 ref  9 30 5H6M       *  0   0 AGCTAA    *   NM:i:1
r004   0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC       *
r003  16 ref 29 30 6H5M       *  0   0 TAGGC     *   NM:i:0
r001  83 ref 37 30 9M         =  7 -39 CAGCGCCAT         *
```

```
@HD VN:1.3 SO:coordinate
@SQ SN:ref LN:45
r001 163 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002   0 ref  9 30 3S6M1P1I4M * 0   0 AAAAGATAAGGATA    *
r003   0 ref  9 30 5H6M       * 0   0 AGCTAA   *   NM:i:1
r004   0 ref 16 30 6M14N5M    * 0   0 ATAGCTTCAGC       *
r003  16 ref 29 30 6H5M       * 0   0 TAGGC    *   NM:i:0
r001  83 ref 37 30 9M         = 7 -39 CAGCGCCAT         *
```

Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	`[!-?A-~]{1,255}`	Query template NAME
2	FLAG	Int	$[0,2^{16}-1]$	bitwise FLAG
3	RNAME	String	`\*|[!-()+-<>-~][!-~]*`	Reference sequence NAME
4	POS	Int	$[0,2^{29}-1]$	1-based leftmost mapping POSition
5	MAPQ	Int	$[0,2^{8}-1]$	MAPping Quality
6	CIGAR	String	`\*|([0-9]+[MIDNSHPX=])+`	CIGAR string
7	RNEXT	String	`\*|=|[!-()+-<>-~][!-~]*`	Ref. name of the mate/next segment
8	PNEXT	Int	$[0,2^{29}-1]$	Position of the mate/next segment
9	TLEN	Int	$[-2^{29}+1,2^{29}-1]$	observed Template LENgth
10	SEQ	String	`\*|[A-Za-z=.]+`	segment SEQuence
11	QUAL	String	`[!-~]+`	ASCII of Phred-scaled base QUALity+33

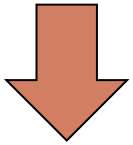
```
@HD VN:1.3 SO:coordinate
@SQ SN:ref LN:45
r001 163 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002   0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA    *
r003   0 ref  9 30 5H6M       *  0   0 AGCTAA   *   NM:i:1
r004   0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC       *
r003  16 ref 29 30 6H5M       *  0   0 TAGGC    *   NM:i:0
r001  83 ref 37 30 9M         =  7 -39 CAGCGCCAT         *
```

CIGAR string

Op	BAM	Description
M	0	alignment match (can be a sequence match or mismatch)
I	1	insertion to the reference
D	2	deletion from the reference
N	3	skipped region from the reference
S	4	soft clipping (clipped sequences present in SEQ)
H	5	hard clipping (clipped sequences NOT present in SEQ)
P	6	padding (silent deletion from padded reference)
=	7	sequence match
X	8	sequence mismatch

```
Coor     12345678901234  5678901234567890123456789012345
ref      AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT

+r001/1        TTAGATAAAGGATA*CTG
```



```
r001 163 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
```

CIGAR string:
8 aligned nucleotides (8M), 2 insertions (2I), 4 aligned nucleotides (4M), 1 deletions (1D), 3 aligned nucleotides

Op	BAM	Description
M	0	alignment match (can be a sequence match or mismatch)
I	1	insertion to the reference
D	2	deletion from the reference
N	3	skipped region from the reference
S	4	soft clipping (clipped sequences present in SEQ)
H	5	hard clipping (clipped sequences NOT present in SEQ)
P	6	padding (silent deletion from padded reference)
=	7	sequence match
X	8	sequence mismatch