

CRAC, a software for analyzing RNA-seq reads

N. Philippe

T. Commes

É. Rivals

M. Salson

CRAC, a software for analyzing RNA-seq reads



N. Philippe



T. Commes

Institut de recherche en biothérapie
Inserm – Univ. Montpellier 1 – CHU

É. Rivals

M. Salson

CRAC, a software for analyzing RNA-seq reads



N. Philippe



T. Commes



É. Rivals

Institut de recherche en biothérapie
Inserm – Univ. Montpellier 1 – CHU

LIRMM – Équipe MAB
CNRS – Univ. Montpellier 2

M. Salson

CRAC, a software for analyzing RNA-seq reads



N. Philippe

Institut de biologie computationnelle



T. Commes

Institut de recherche en biothérapie
Inserm – Univ. Montpellier 1 – CHU

M. Salson



É. Rivals

LIRMM – Équipe MAB
CNRS – Univ. Montpellier 2

CRAC, a software for analyzing RNA-seq reads



N. Philippe

Institut de biologie computationnelle



T. Commes



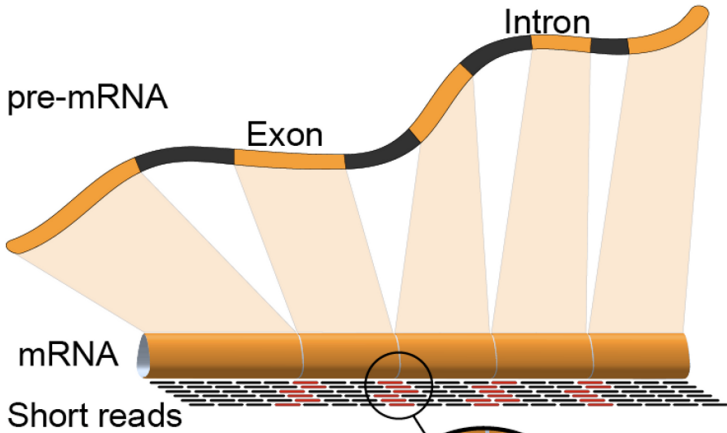
É. Rivals

Institut de recherche en biothérapie
Inserm – Univ. Montpellier 1 – CHU

LIRMM – Équipe MAB
CNRS – Univ. Montpellier 2

M. Salson

Équipe Bonsai
LIFL (CNRS – Univ. Lille 1)
Inria

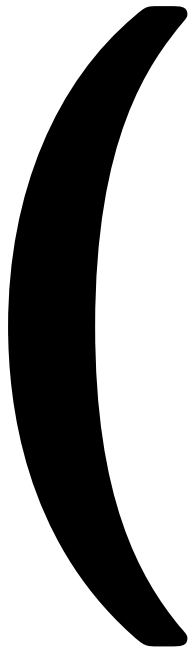


Short read is split by
intron when aligning
to reference Genome

CGACTAGCTAGCATCAGCATCTATTATAGCATCATCGACTATACGACTATCGATCATCGTATATATAGGAGATCACAG



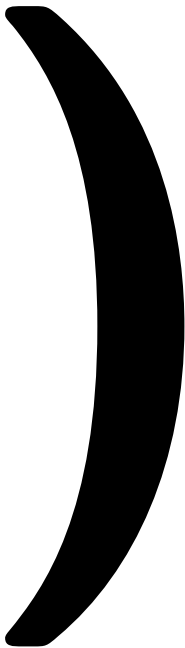




Underlying hypothesis

Underlying hypothesis

1 read $\longrightarrow$ 1 location







In our bibliography

In our bibliography

Using reads to annotate the genome: influence of length, background distribution, and sequence errors on prediction capacity

**Nicolas Philippe^{1,2}, Anthony Boureux², Laurent Bréhélin¹, Jorma Tarhio³,
Thérèse Commes² and Éric Rivals^{1,*}**

22

22

for the human genome

22: not too small, not too large

Genome

22: not too small, not too large

$\overleftrightarrow{22}$

Genome

22: not too small, not too large

22



A diagram illustrating a relationship between a point on a horizontal line and a point above it. A thick black horizontal line represents a 'Genome'. A point on this line is labeled '1 loc.'. A curved arrow points from this point to a point above it. A horizontal double-headed arrow above the curved arrow is labeled '22', indicating the horizontal distance between the point on the line and the point above it.

Genome

1 loc.

22: not too small, not too large

22



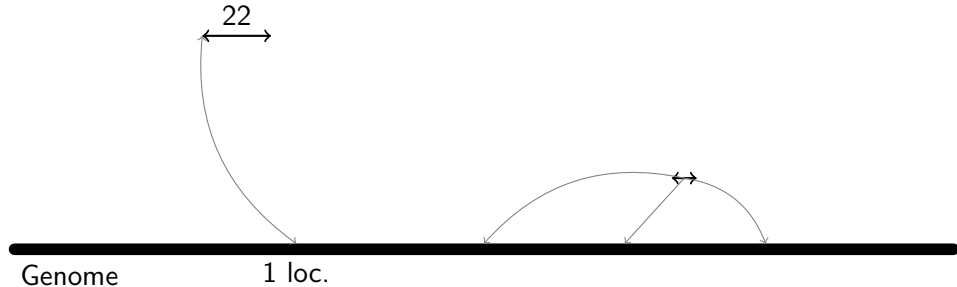
A diagram illustrating a genomic region. A thick horizontal black line represents the genome. A point on this line is labeled "1 loc.". A curved arrow points from this point to a point above the line. A horizontal double-headed arrow above the point is labeled "22". Another horizontal double-headed arrow is shown further to the right on the line.

Genome

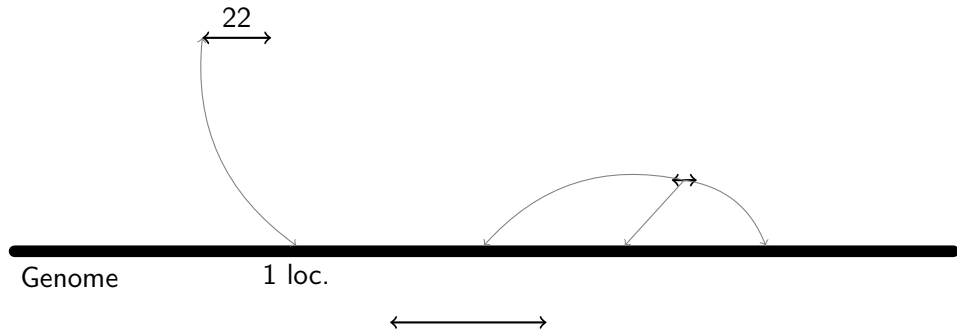
1 loc.

$\leftrightarrow$

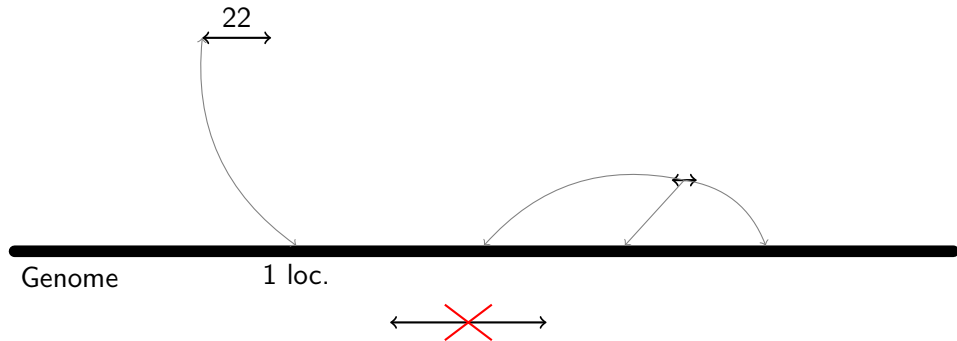
22: not too small, not too large



22: not too small, not too large



22: not too small, not too large



Underlying hypothesis

Underlying hypothesis

Resequencing

100



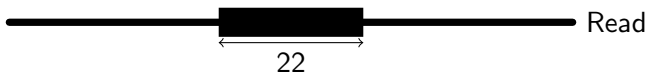
Read

100

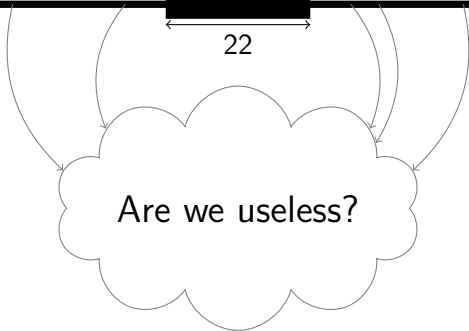
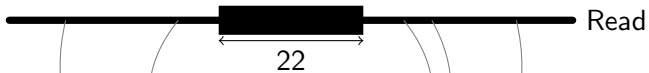


Read

← 100 →



← 100 →



More data

More info

In an ideal world

Genome

Read



We're not in an ideal world

Genome

Read





We're not in an ideal world

Genome

Read



Because!

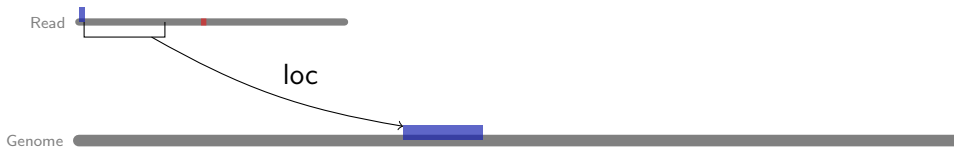
Because!

We can explain what's going on (sometimes)

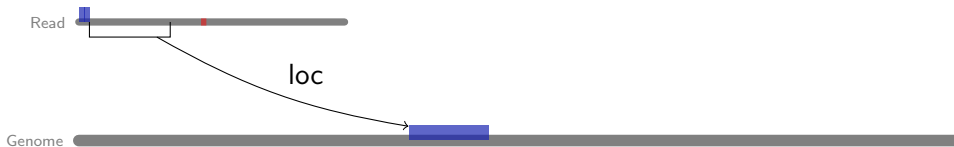
Using 22-mers to decipher a read



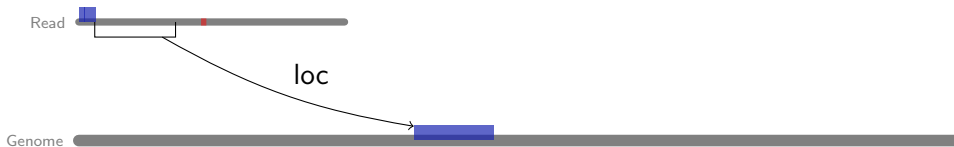
Using 22-mers to decipher a read



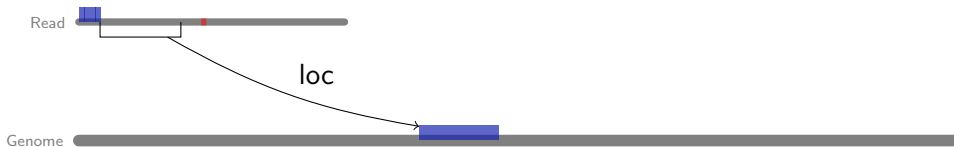
Using 22-mers to decipher a read



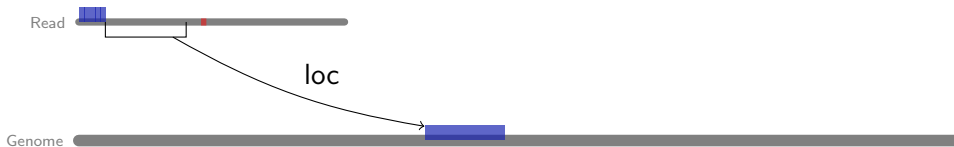
Using 22-mers to decipher a read



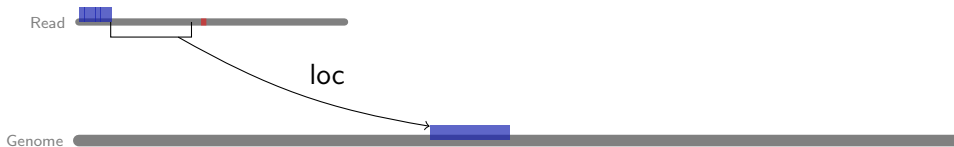
Using 22-mers to decipher a read



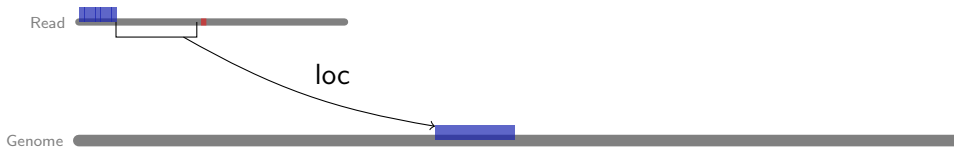
Using 22-mers to decipher a read



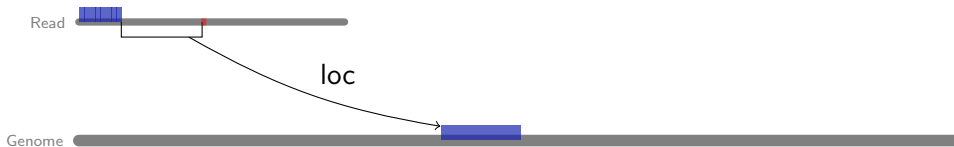
Using 22-mers to decipher a read



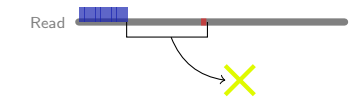
Using 22-mers to decipher a read



Using 22-mers to decipher a read



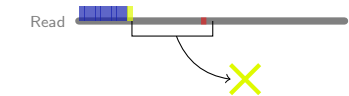
Using 22-mers to decipher a read



Genome

A long horizontal grey bar representing the genome.

Using 22-mers to decipher a read



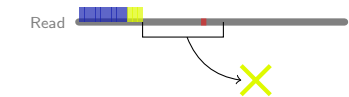
Genome

A long horizontal grey bar representing the genome.

Using 22-mers to decipher a read



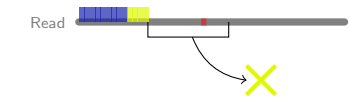
Using 22-mers to decipher a read



Genome

A long horizontal grey bar representing the genome.

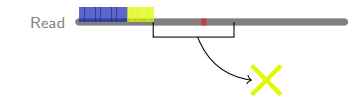
Using 22-mers to decipher a read



Genome

A long horizontal grey bar representing the genome.

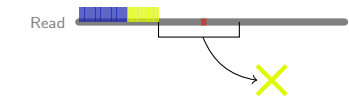
Using 22-mers to decipher a read



Genome

A long horizontal grey bar representing the genome.

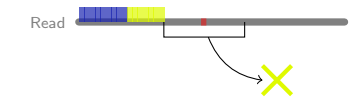
Using 22-mers to decipher a read



Genome

A long horizontal grey bar representing the genome.

Using 22-mers to decipher a read



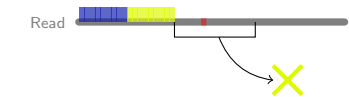
Genome

A long horizontal grey bar representing the genome.

Using 22-mers to decipher a read



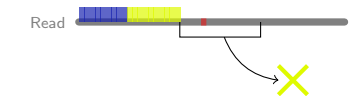
Using 22-mers to decipher a read



Genome

The diagram shows a horizontal grey bar representing the genome. A white box highlights a segment of the genome, and an arrow points from this box to a large yellow 'X' mark, indicating a mismatch or error in the alignment.

Using 22-mers to decipher a read



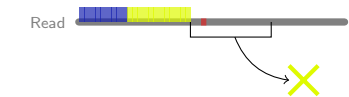
Genome

A long horizontal grey bar representing the genome.

Using 22-mers to decipher a read



Using 22-mers to decipher a read



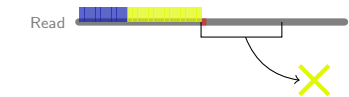
Genome

A long horizontal grey bar representing the genome.

Using 22-mers to decipher a read



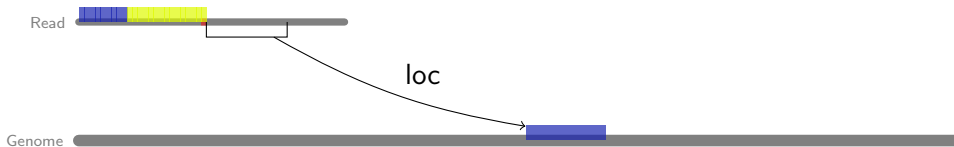
Using 22-mers to decipher a read



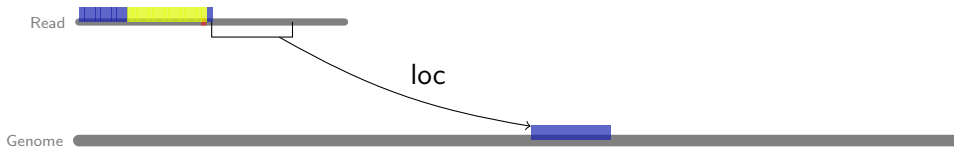
Genome

A long horizontal grey bar representing the genome.

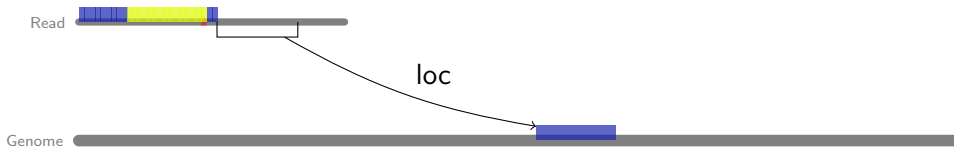
Using 22-mers to decipher a read



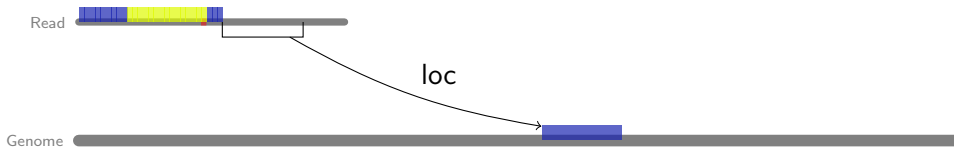
Using 22-mers to decipher a read



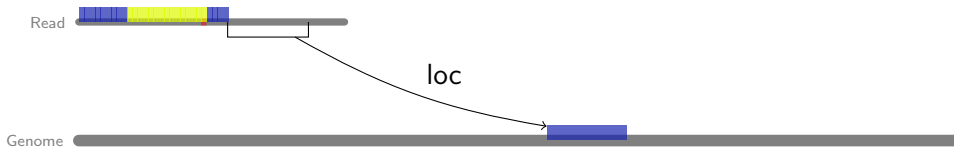
Using 22-mers to decipher a read



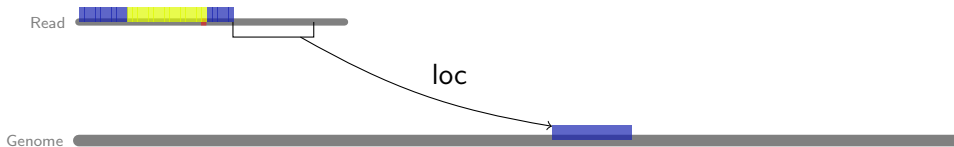
Using 22-mers to decipher a read



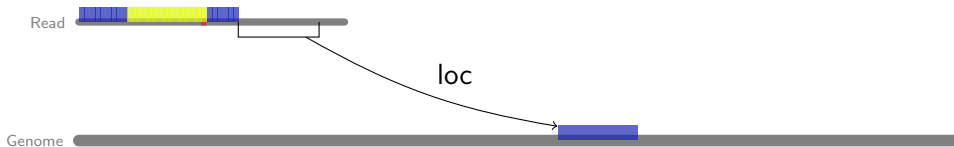
Using 22-mers to decipher a read



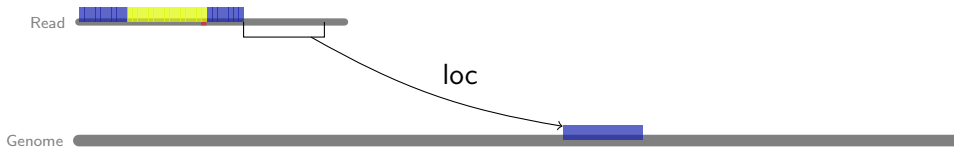
Using 22-mers to decipher a read



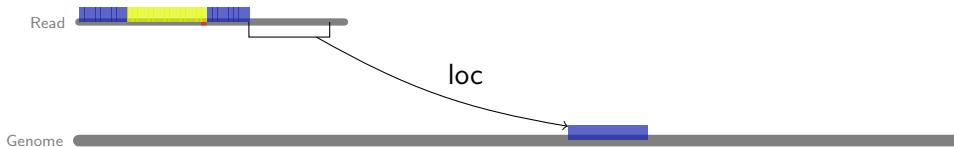
Using 22-mers to decipher a read



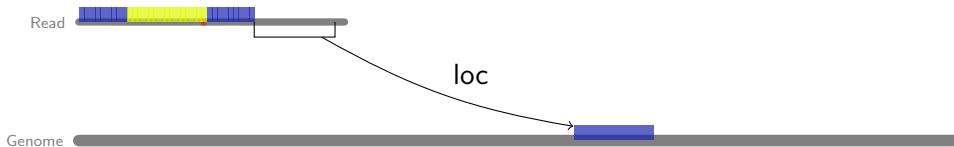
Using 22-mers to decipher a read



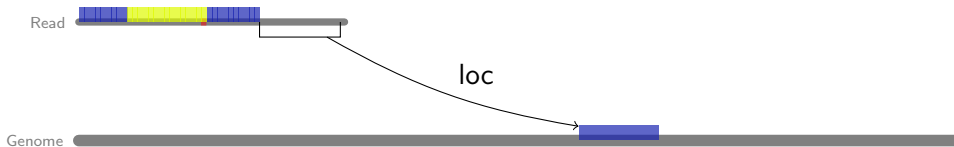
Using 22-mers to decipher a read



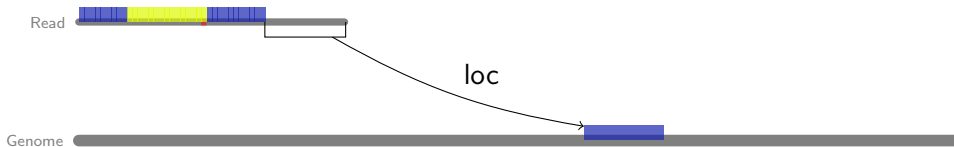
Using 22-mers to decipher a read



Using 22-mers to decipher a read



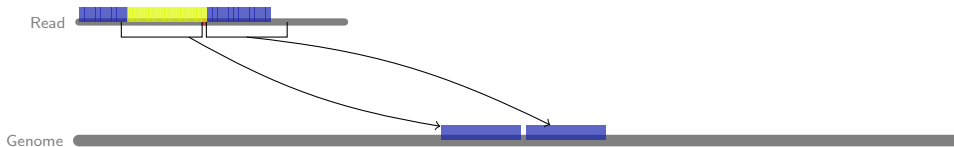
Using 22-mers to decipher a read



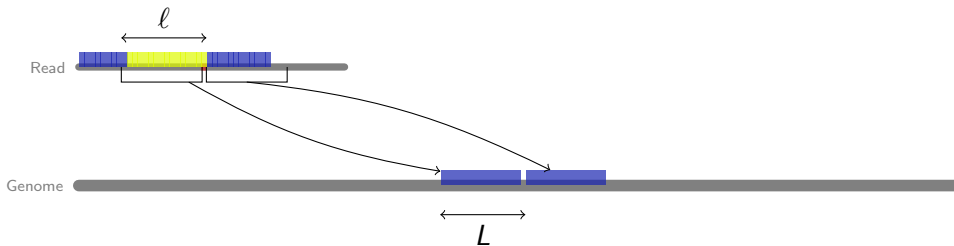
Using 22-mers to decipher a read



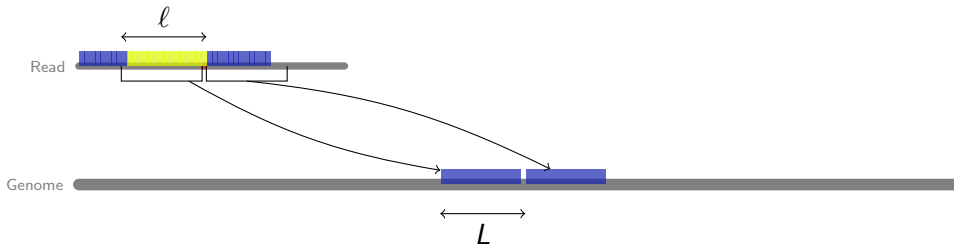
Using 22-mers to decipher a read



Using 22-mers to decipher a read

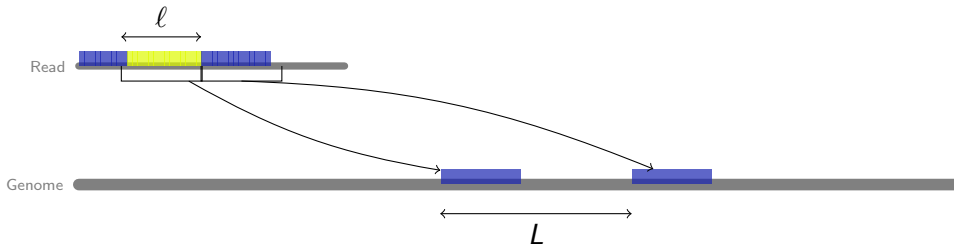


Using 22-mers to decipher a read



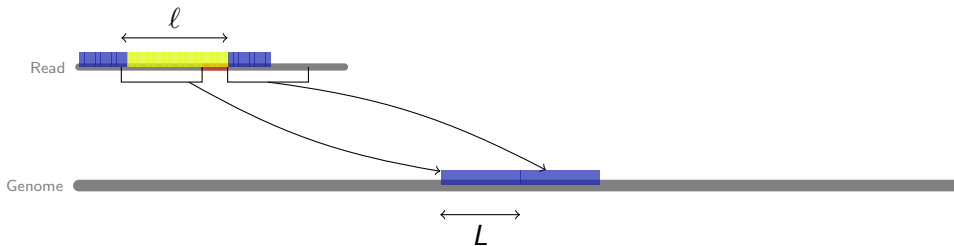
Substitution

Using 22-mers to decipher a read



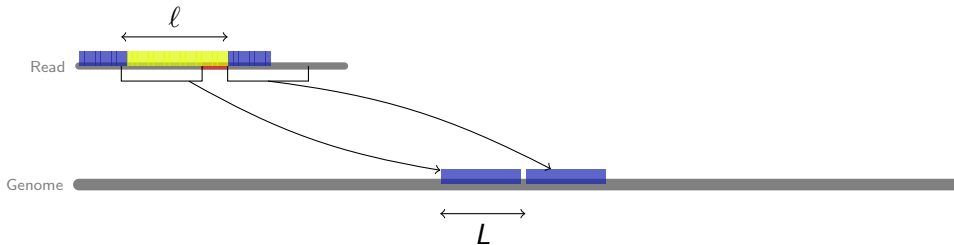
Deletion

Using 22-mers to decipher a read



Insertion

Using 22-mers to decipher a read

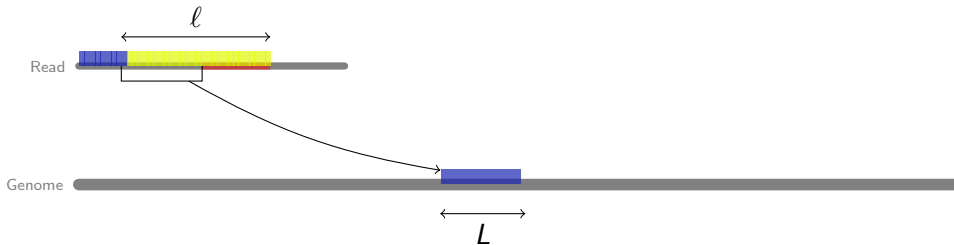


=

<

>

Using 22-mers to decipher a read



We're not in an ideal world

Some vision problems. . .



1cm

Error

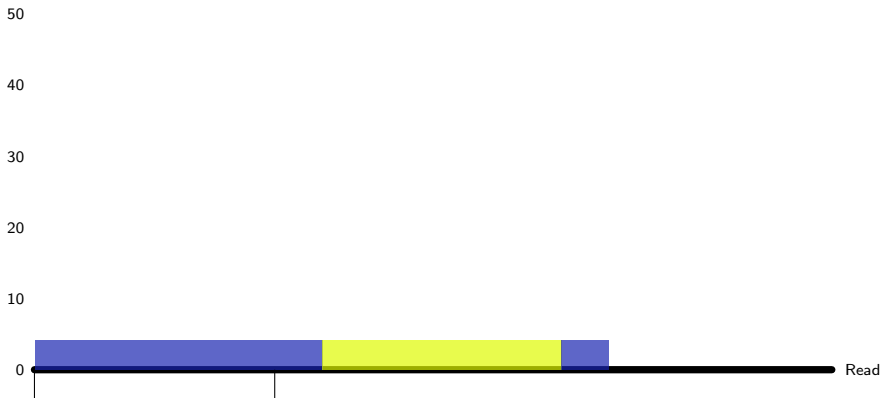
or

Variation?

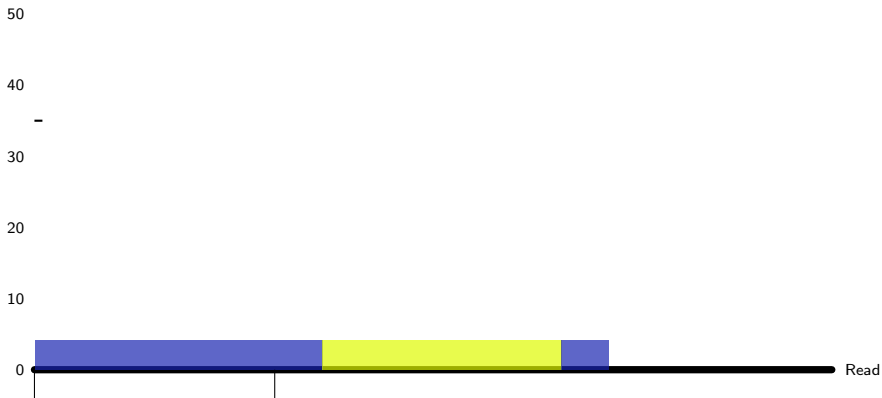
Error or variation: using read collection



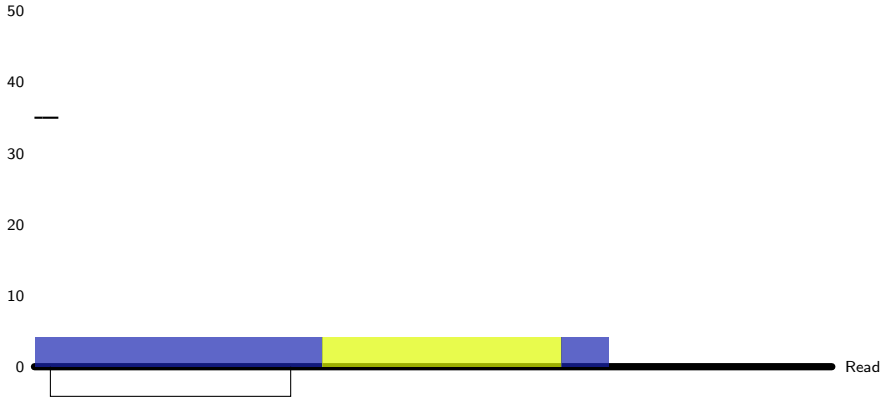
Error or variation: using read collection



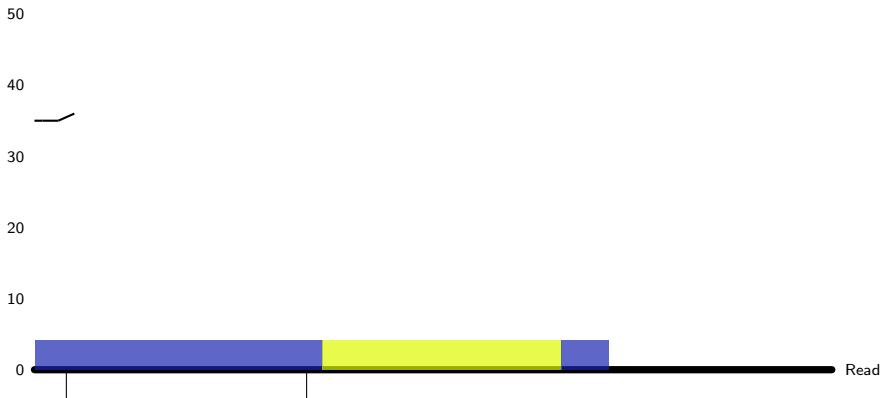
Error or variation: using read collection



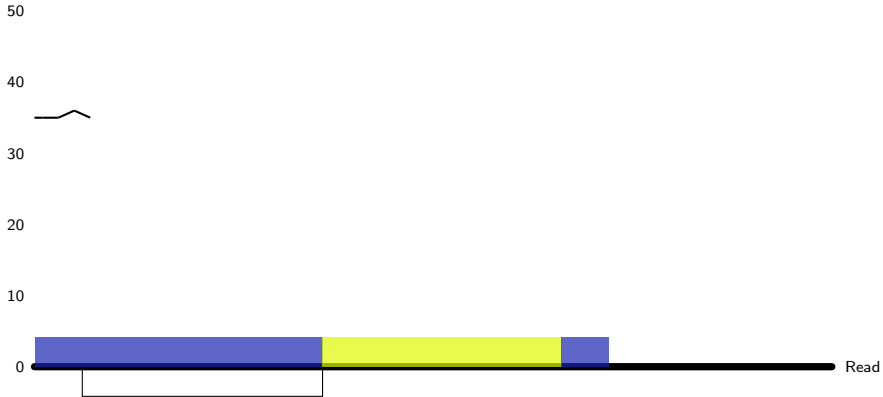
Error or variation: using read collection



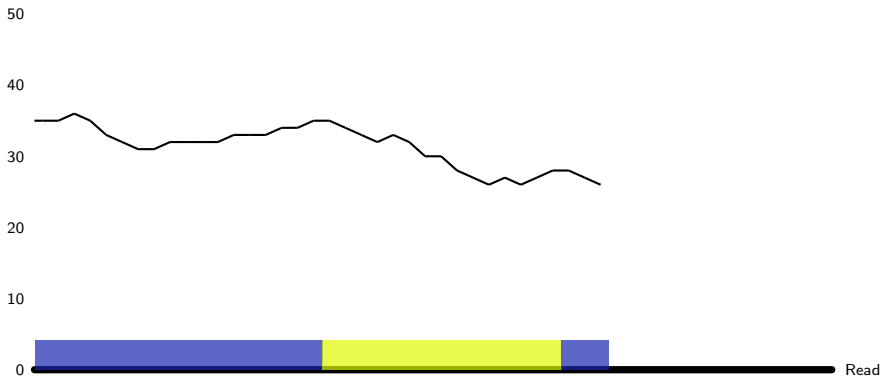
Error or variation: using read collection



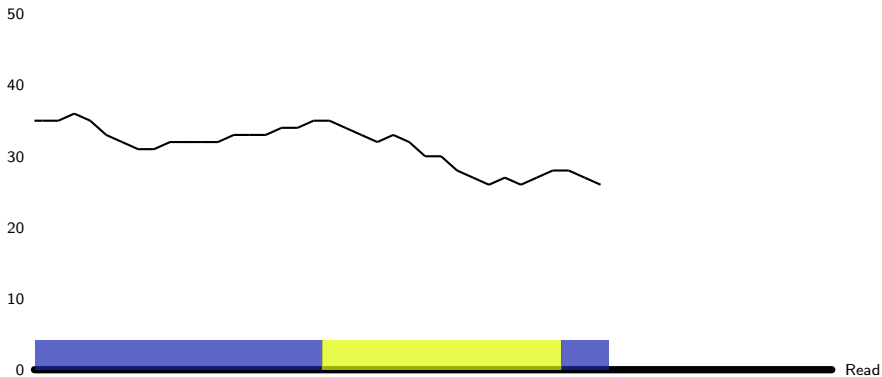
Error or variation: using read collection



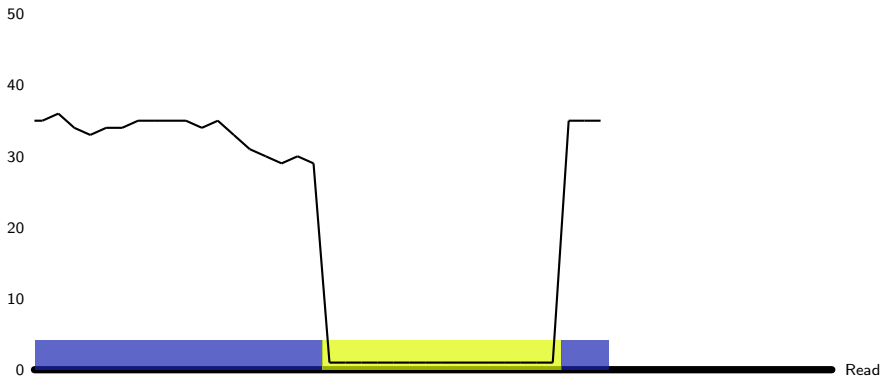
Error or variation: using read collection



Error or variation: using read collection



Error or variation: using read collection



We're not in an ideal world

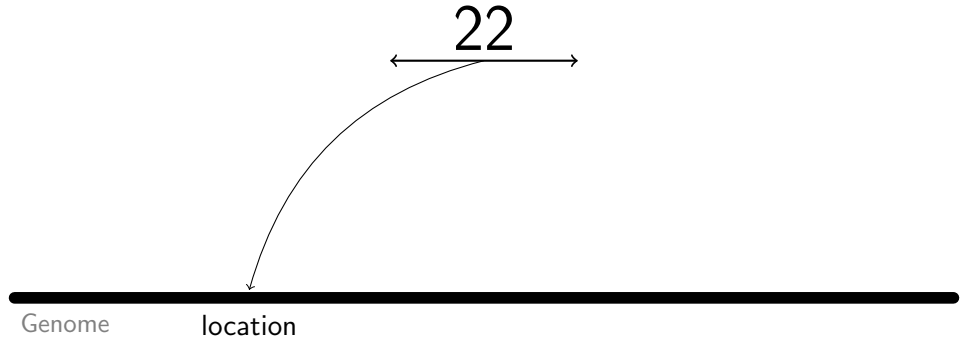
22, not such a magic number

22

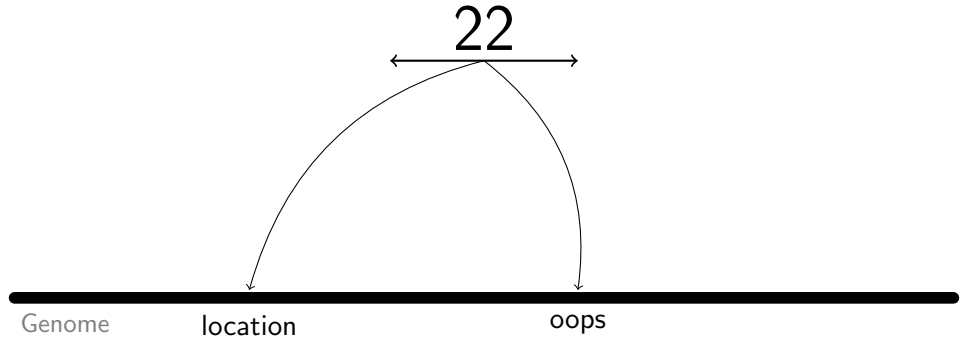


Genome

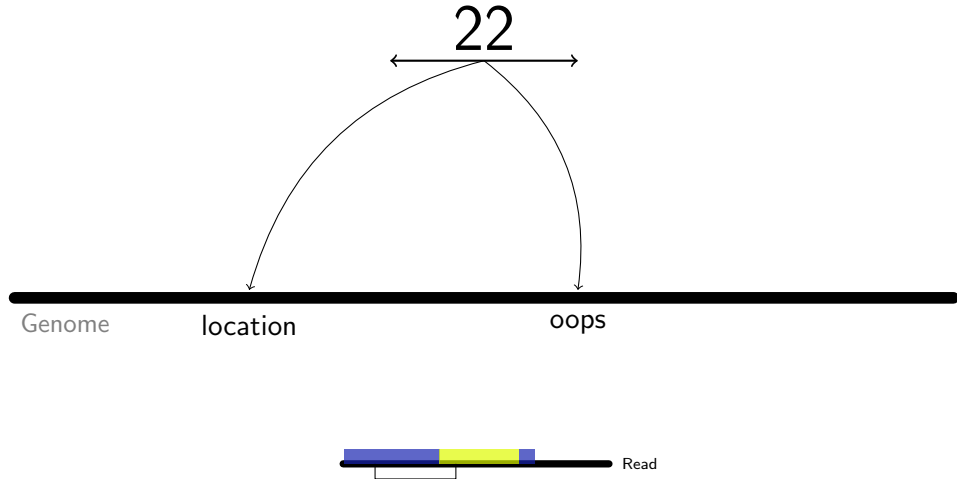
22, not such a magic number



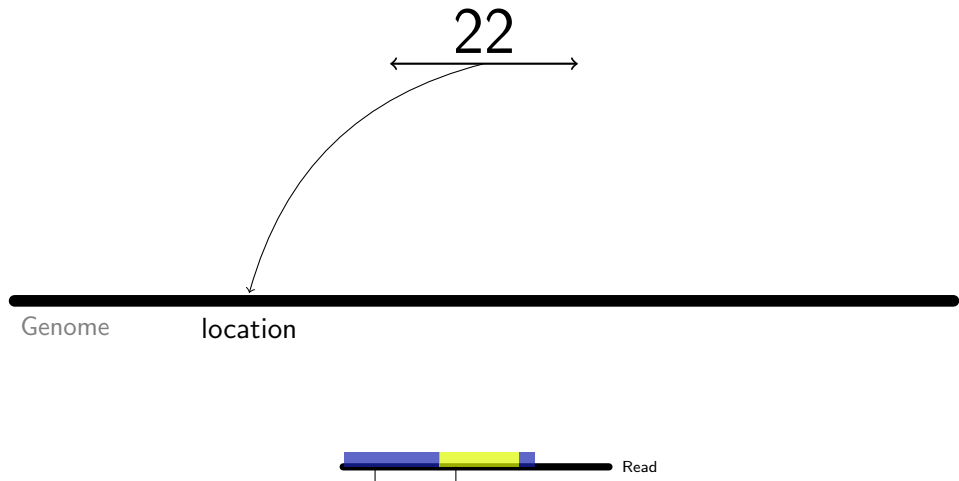
22, not such a magic number



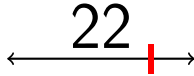
22, not such a magic number



22, not such a magic number



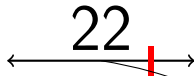
22, not such a magic number



Genome

22, not such a magic number

22



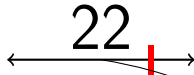
A horizontal line with a red vertical tick mark. A curved arrow starts from the tick mark and points downwards and to the right towards the text 'unexpected location'.

Genome

unexpected location

22, not such a magic number

22



Genome

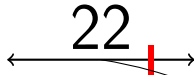
unexpected location



Read

22, not such a magic number

22



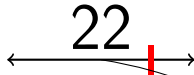
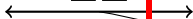
Genome

unexpected location



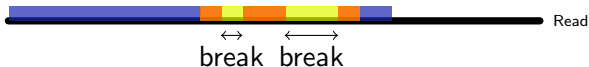
22, not such a magic number

22

A horizontal black line with a red vertical tick mark. A double-headed arrow is positioned above the tick mark, with the number '22' centered above it.

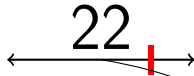
Genome

unexpected location



22, not such a magic number

22

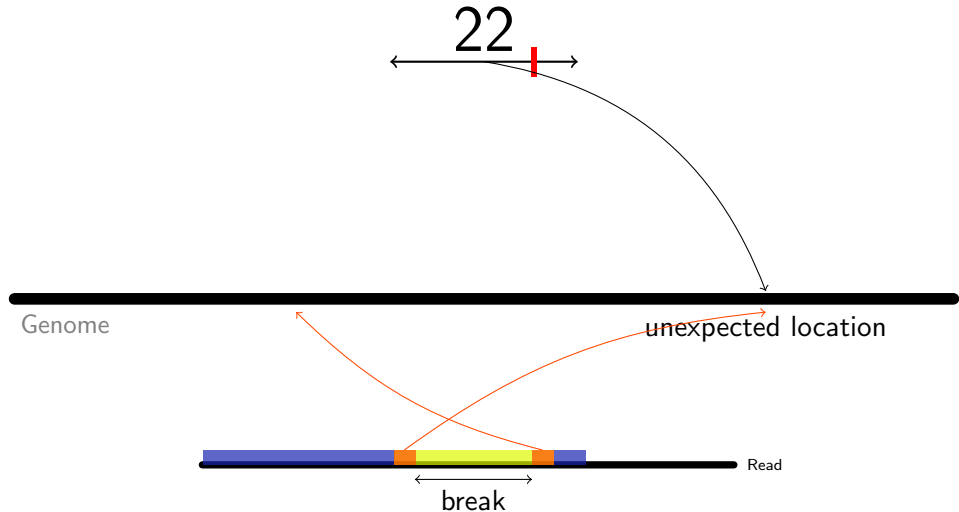


Genome

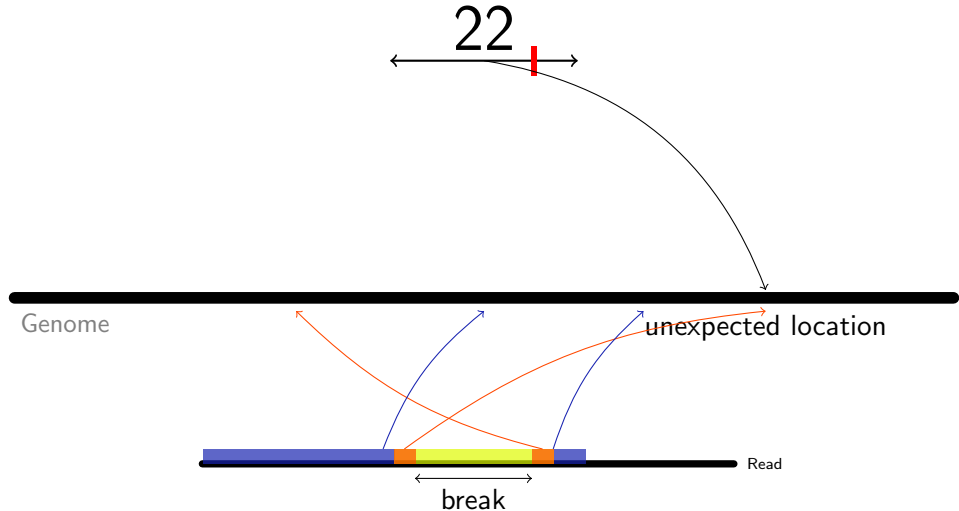
unexpected location



22, not such a magic number

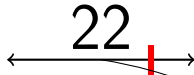


22, not such a magic number



22, not such a magic number

22



Genome

unexpected location



Read

CRAC: a global RNA-seq data analysis

CRAC: a global RNA-seq data analysis

Read mapping

CRAC: a global RNA-seq data analysis

Read mapping

Substitutions

CRAC: a global RNA-seq data analysis

Read mapping

Substitutions

Short indels

CRAC: a global RNA-seq data analysis

Read mapping

Substitutions

Short indels

Splicing events

CRAC: a global RNA-seq data analysis

Read mapping

Substitutions

Short indels

Splicing events

Fusions

CRAC: a global RNA-seq data analysis

Read mapping

Substitutions

Short indels

Splicing events

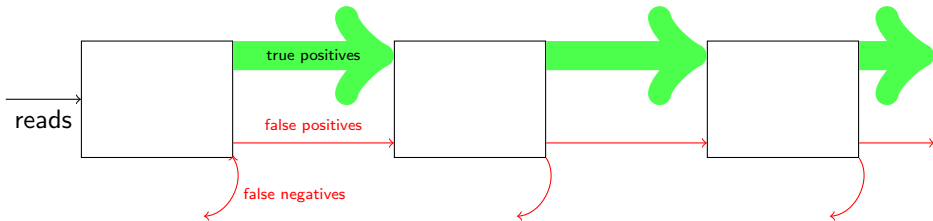
Fusions

(Sequencing errors)

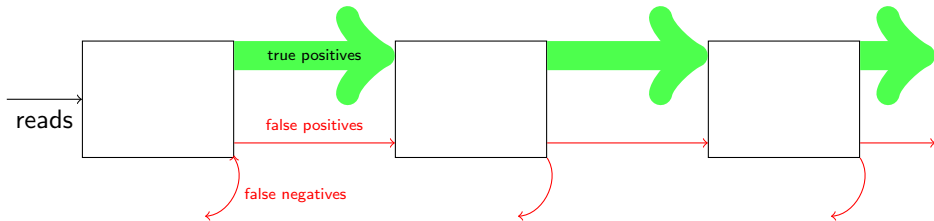
Keep it simple and stupid?



Keep it simple and stupid?



Keep it simple and stupid?

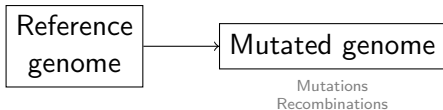


One tool may be better than many...

Simulated RNA-Seq data

Reference
genome

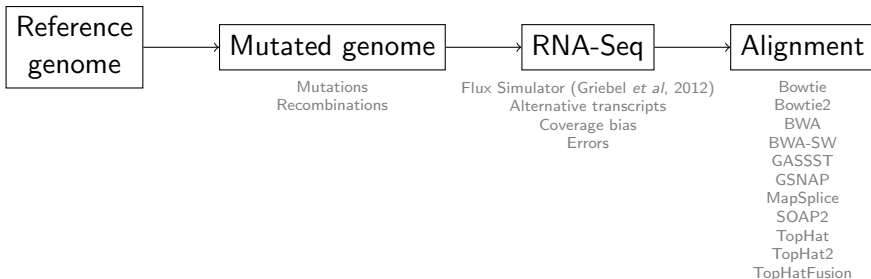
Simulated RNA-Seq data



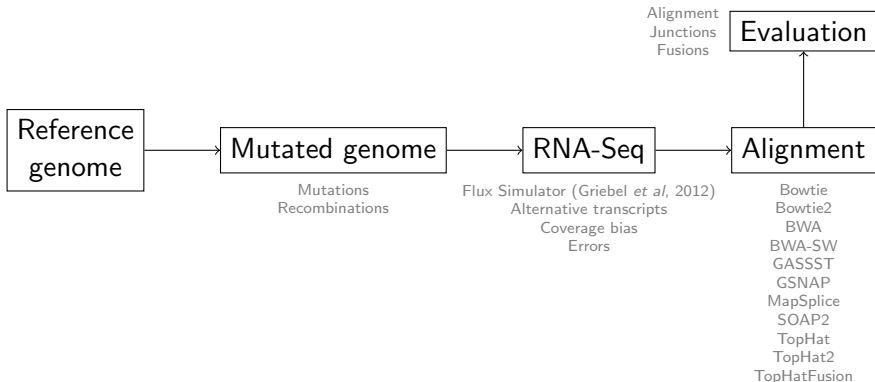
Simulated RNA-Seq data



Simulated RNA-Seq data



Simulated RNA-Seq data



Datasets

45M reads

75bp

29k SNVs

2.5k insertions

2.5k deletions

647 fusions

13M errors

Datasets

45M reads

75bp

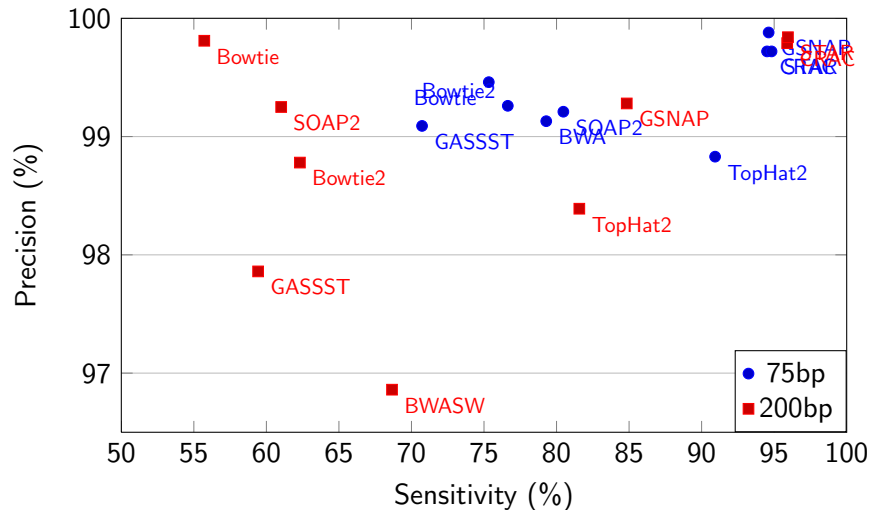
29k SNVs
2.5k insertions
2.5k deletions
647 fusions
13M errors

48M reads

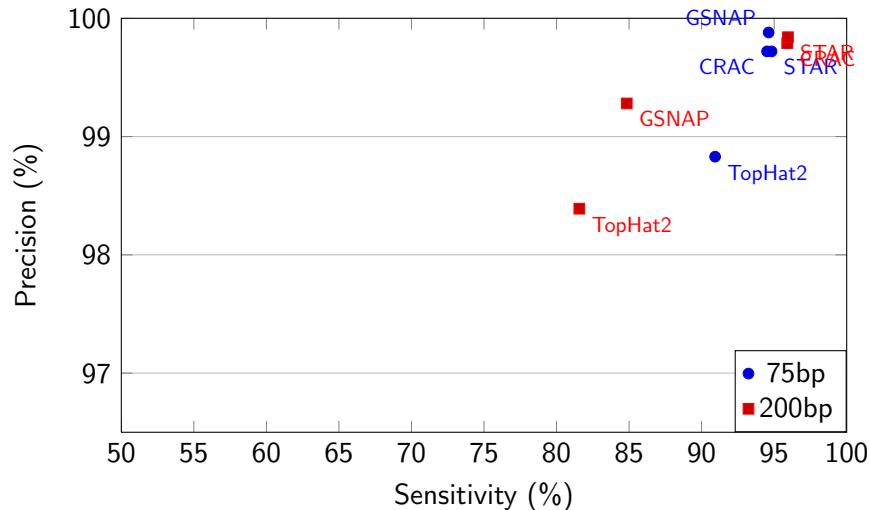
200bp

53k SNVs
5k insertions
5k deletions
914 fusions
39M errors

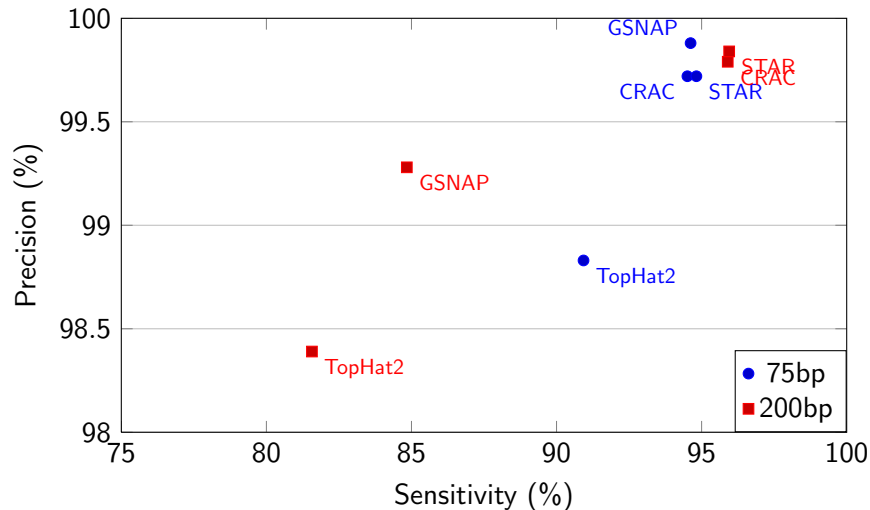
Alignment (single occurrences)



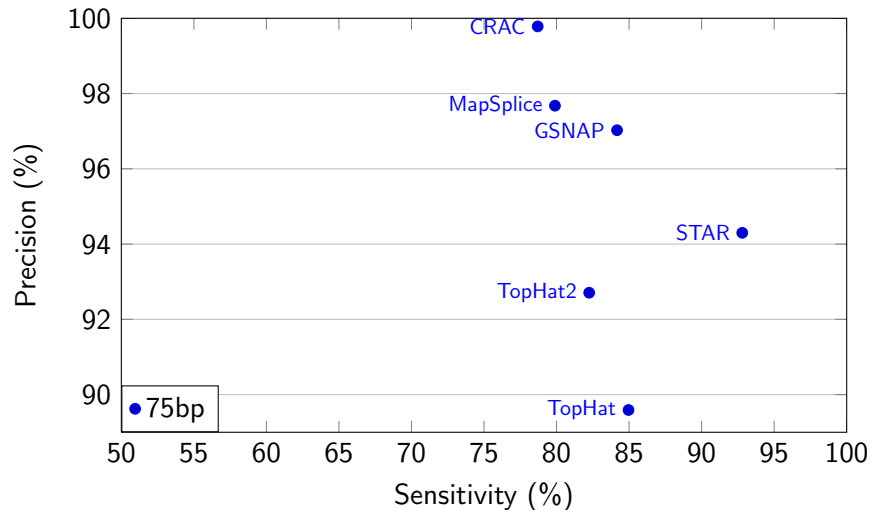
Alignment (single occurrences)



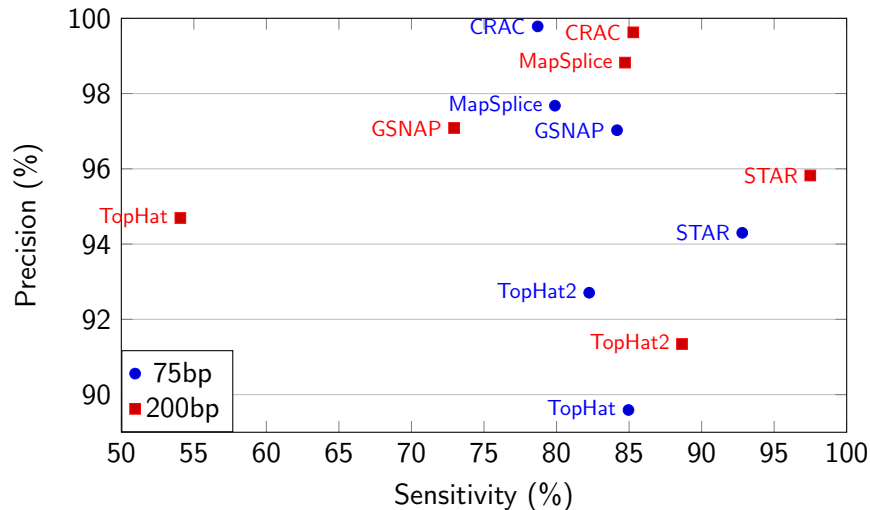
Alignment (single occurrences)



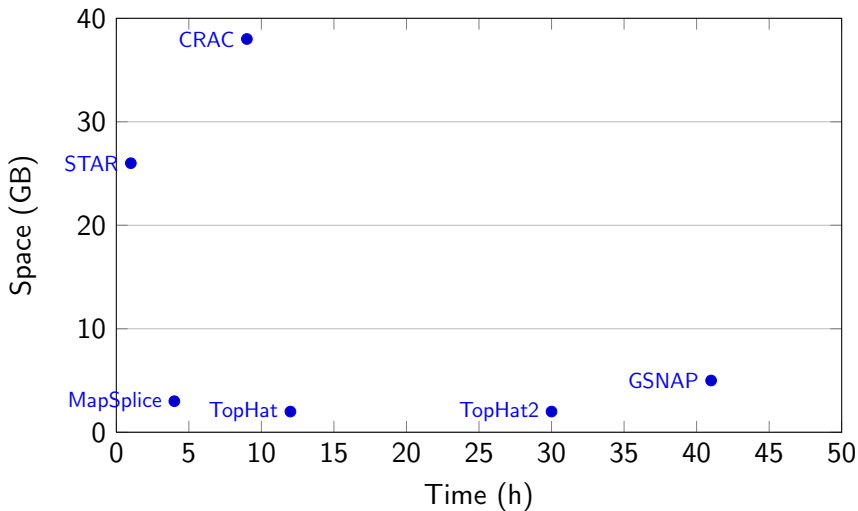
Exon-exon junction predictions



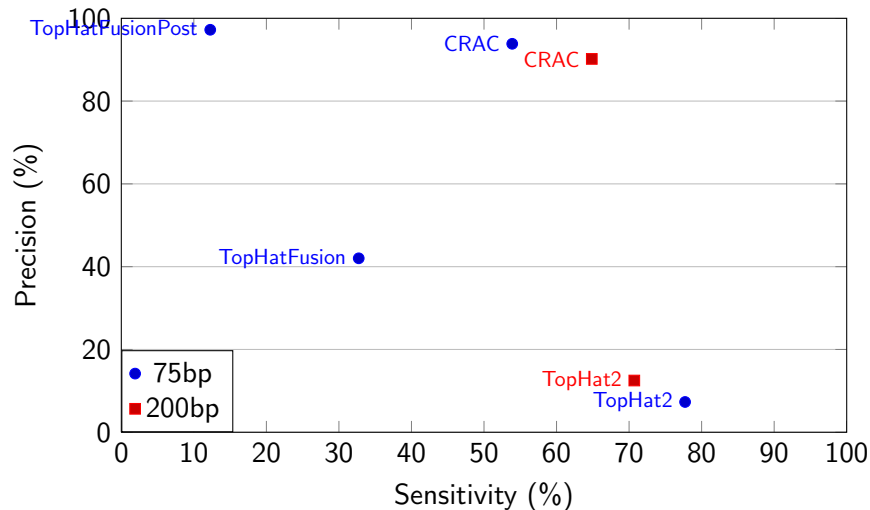
Exon-exon junction predictions



How many time? How many space?



Fusion predictions



Junction prediction in real data

ERR030856

RNA-Seq, Hi-Seq 2000

100bp stranded

76 M reads

Junction prediction in real data

ERR030856

RNA-Seq, Hi-Seq 2000

100bp stranded

76 M reads

Predicted
junctions

CRAC

GSNAP

MapSplice

TopHat

TopHat2

Junction prediction in real data

ERR030856

RNA-Seq, Hi-Seq 2000

100bp stranded

76 M reads

Predicted
junctions

Extraction

25 pb before and after
each junction

CRAC
GSNAP
MapSplice
TopHat
TopHat2

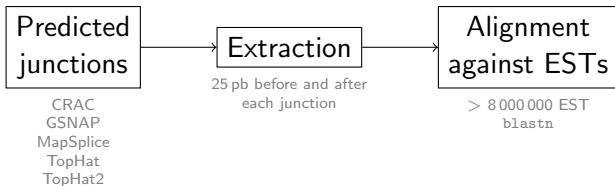
Junction prediction in real data

ERR030856

RNA-Seq, Hi-Seq 2000

100bp stranded

76 M reads



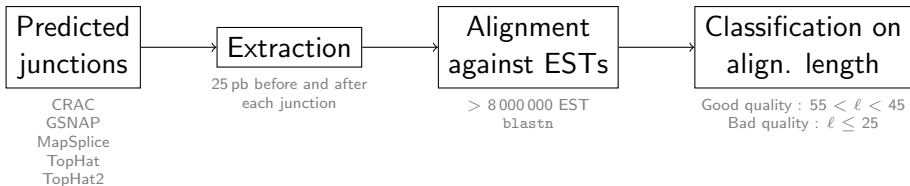
Junction prediction in real data

ERR030856

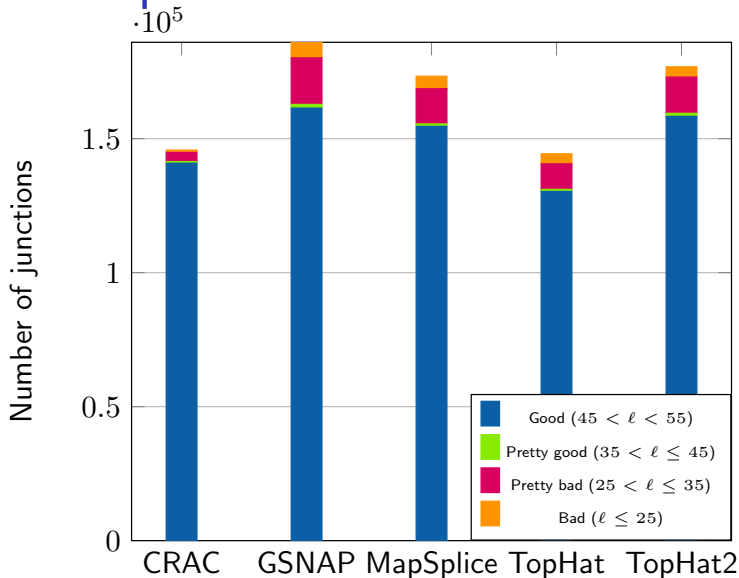
RNA-Seq, Hi-Seq 2000

100bp stranded

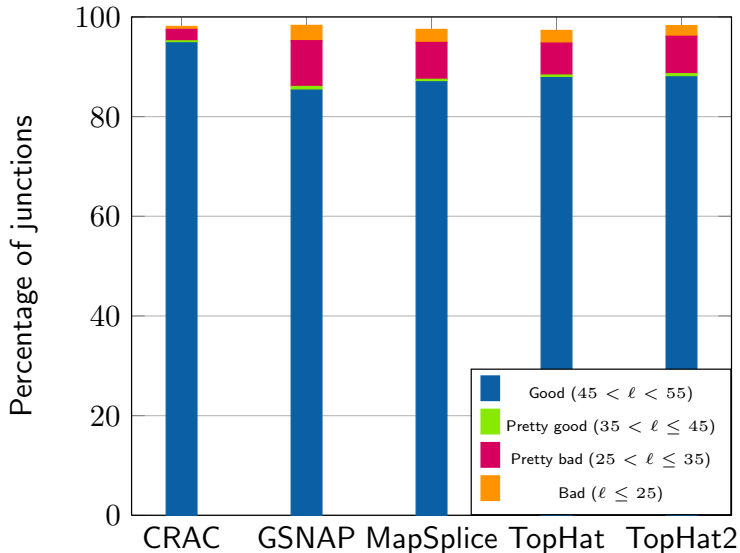
76 M reads



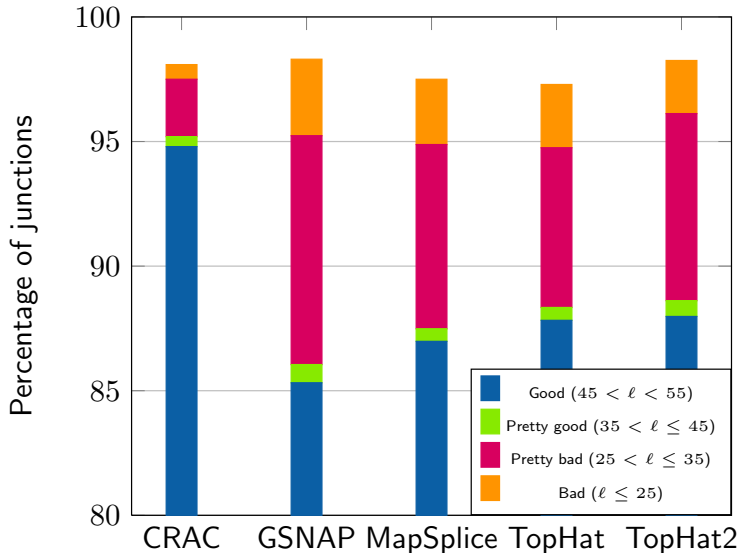
Junction prediction in real data



Junction prediction in real data



Junction prediction in real data



Fusions in real data

Real data

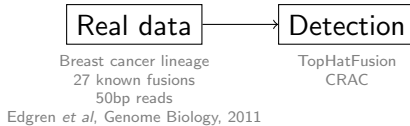
Breast cancer lineage

27 known fusions

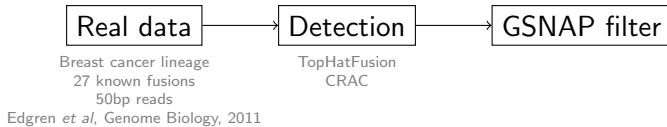
50bp reads

Edgren *et al*, Genome Biology, 2011

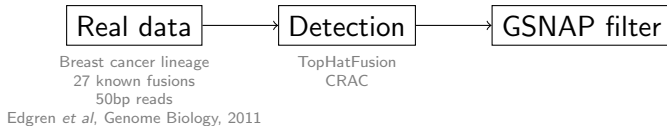
Fusions in real data



Fusions in real data



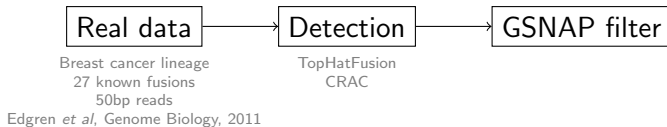
Fusions in real data



CRAC

455

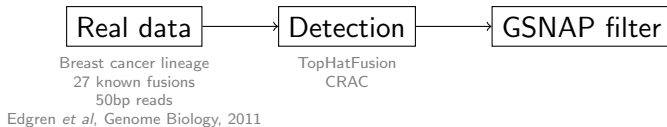
Fusions in real data



CRAC
455

TopHatFusion
193 163

Fusions in real data



CRAC

455

Validated: 20

TopHatFusion

193 163

Validated: 21



rac.

RNA-SEQ Mapping & Analysis

Jérôme Audoux



rac.

RNA-SEQ Mapping & Analysis

Jérôme Audoux

**Re-
sequencing**



rac.

RNA-SEQ Mapping & Analysis

Jérôme Audoux

**Re-
sequencing**

**Unique
location**



rac.

RNA-SEQ Mapping & Analysis

Jérôme Audoux

**Re-
sequencing**

**Unique
location**

**Long
reads**



crac.

RNA-SEQ Mapping & Analysis

Jérôme Audoux

<http://crac.gforge.inria.fr>

**Re-
sequencing**

**Unique
location**

**Long
reads**