Introduction to NGS read mapping

Céline Keime
keime@igbmc.fr

NGS read mapping

- Introduction
- Short read mappers
- Specificity of RNA-seq read mapping
- Alignment and related file formats
- Alignment visualization

NGS read mapping

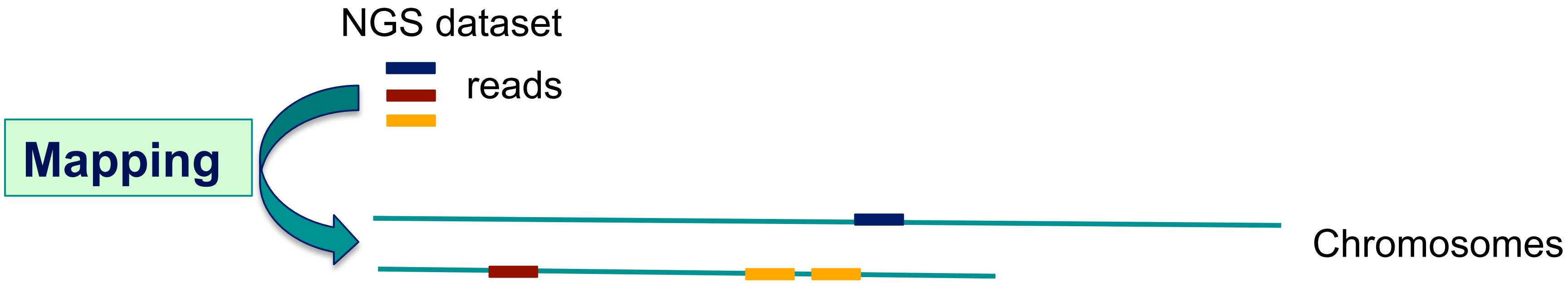
- Introduction
- Short read mappers
- Specificity of RNA-seq read mapping
- Alignment and related file formats
- Alignment visualization

What is mapping ?

- Map reads against a reference genome

  = Predict the locus from which a read originates

  ➔ Find the loci with sufficient similarity



- Sufficient similarity

  ➔ Less mismatches / indels

Alignment	reference genome	CACGTACC	CACGTA_CC	CACGTACC
	reads	CACGTTCC	CACGTATCC	CACGT_CC
		mismatch	indels (insertion/deletion)	

Challenges of short read mapping

- Reference sequence can be large (~3 Gb for human)
- Short reads ➔ several, equally likely places in reference sequence from which they could have been read
  e.g. repetitive regions
- The genome from which reads have been generated may be different from reference genome
  ➔ Need to allow mismatches and indels
- Need to tolerate sequencing errors in reads
- Need to do that for each of the millions of reads !

➔ Too long with traditional mappers such as BLAST or BLAT

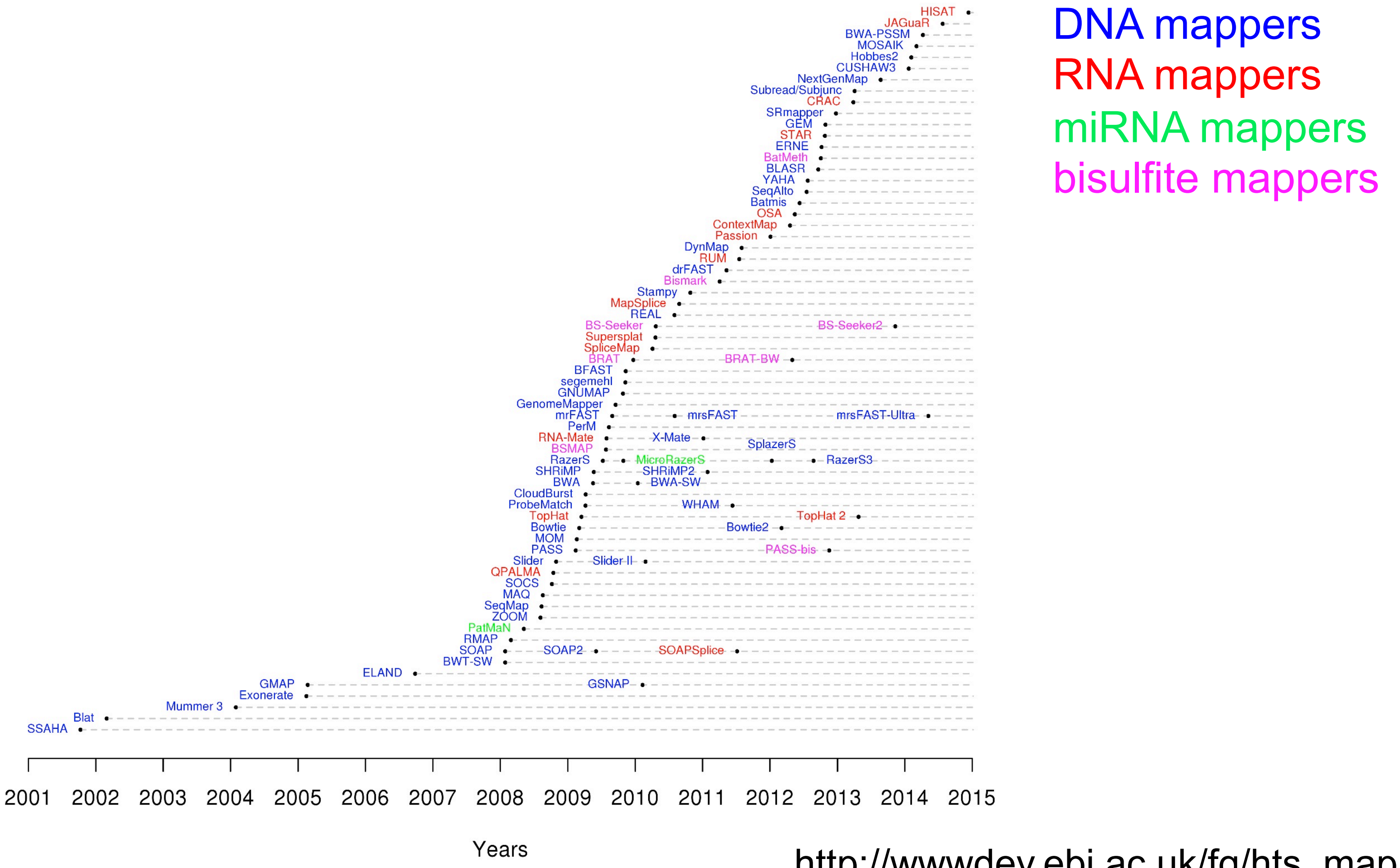
➔ Specialized read mappers with highly efficient algorithms

NGS read mapping

- Introduction
- Short read mappers
- Specificity of RNA-seq read mapping
- Alignment and related file formats
- Alignment visualization

A lot of tools developed …

- More than 90 mapping tools



http://wwwdev.ebi.ac.uk/fg/hts_mappers/

Two main strategies

- Indexing
  - Like the index at the end of a book
    ➔ an index of a large DNA sequence allows one to rapidly find shorter sequences embedded within it
  - 2 strategies : index the reads or the genome
  - e.g. **Maq**

- Transforming
  - Use a technique originally developed for compressing large files called the Burrows-Wheeler transform

    ➔ The transformed human genome fits into less than 2G of memory
  - Align a read character by character to the transformed genome
  - e.g. **Bowtie, BWA**

- More detail (but still brief description) of these strategies in
  Trapnell et al., Nature Biotechnology 2009; 27(5): 455-457

How to choose a mapper ?

- Main criteria to take into account
  - Type of data (DNA, RNA, bisulfite), support of paired-end
  - Read length limits
  - Quality aware
  - Multi-mapping reporting
  - Sensitivity
    - Ability to align a large fraction of reads **with errors and variants**
  - Accuracy
    - If an aligner aligns a large fraction of reads, but most alignments are wrong, this is useless !
  - Speed
  - Memory requirements

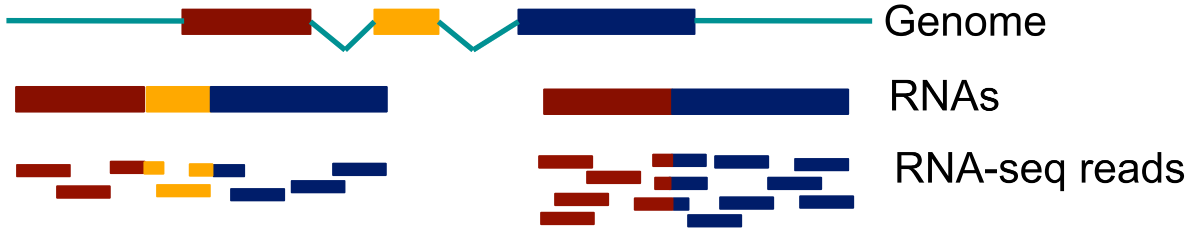
- Several comparative analyses
  - Very interesting to start with :

    Fonseca et al. Bioinformatics 2012;28 (24): 3169-3177.

NGS read mapping

- Introduction
- Short read mappers
- Specificity of RNA-seq read mapping
- Alignment and related file formats
- Alignment visualization

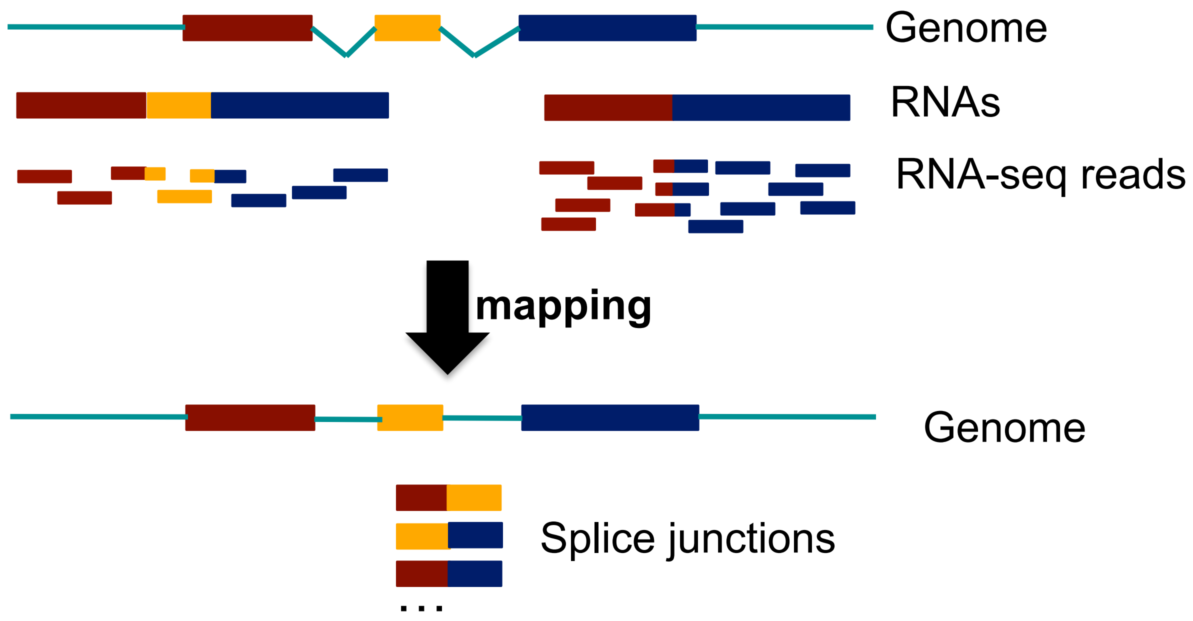
Specificity of RNA-seq reads



➔ In an RNA-seq library, several reads span exon junctions

Map onto the genome and splice junctions ?

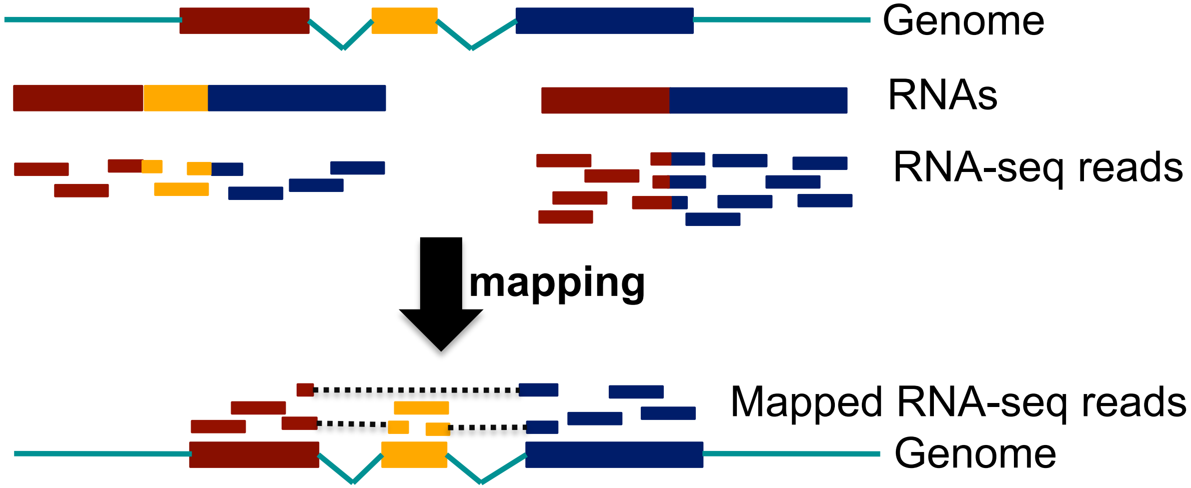
- ERANGE, RNA-Mate



- But
  - Limited to recovering of previously documented splice junctions (known or predicted)

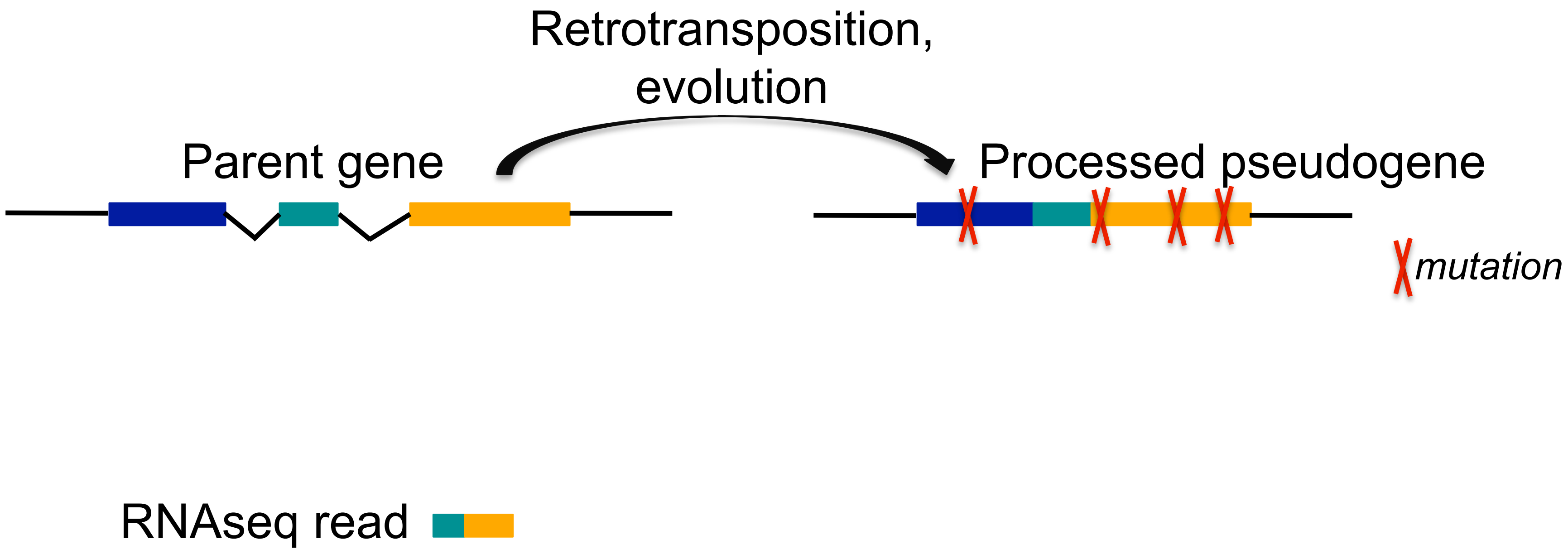
Spliced mapping

- Allows mapping of reads across splice junctions



- Different strategies for spliced mapping
  - 14 mappers developed e.g. Tophat2, GSNAP, MapSplice
  - Comparative analysis
    - Engström et al. Nature Methods 2013;10, 1185–1191

Pseudogenes and spliced mapping



Retrotransposition, evolution

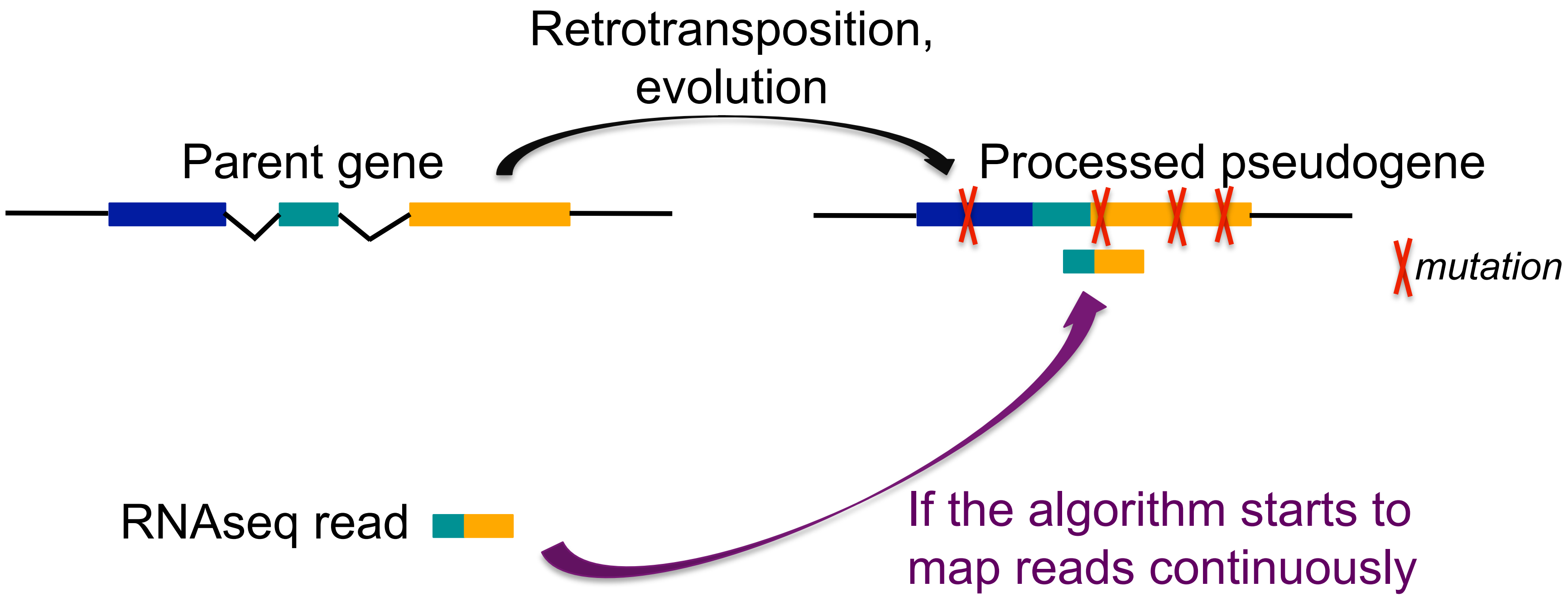
Parent gene

Processed pseudogene

mutation

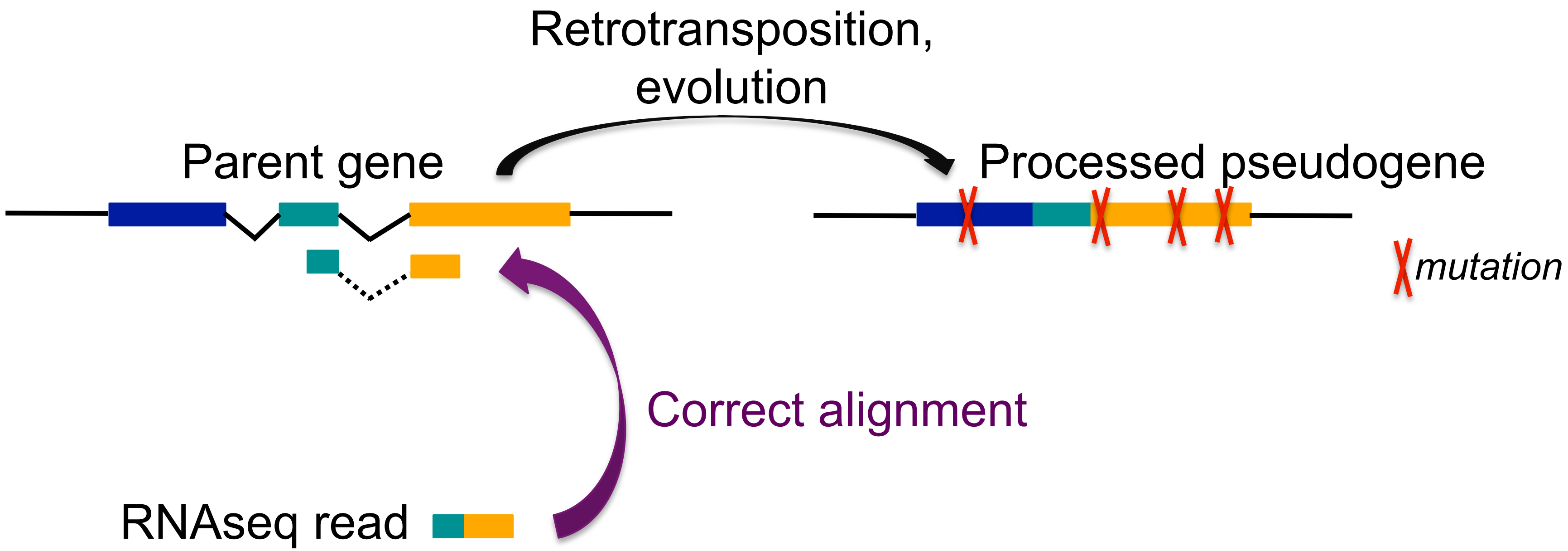
RNAseq read

Pseudogenes and spliced mapping

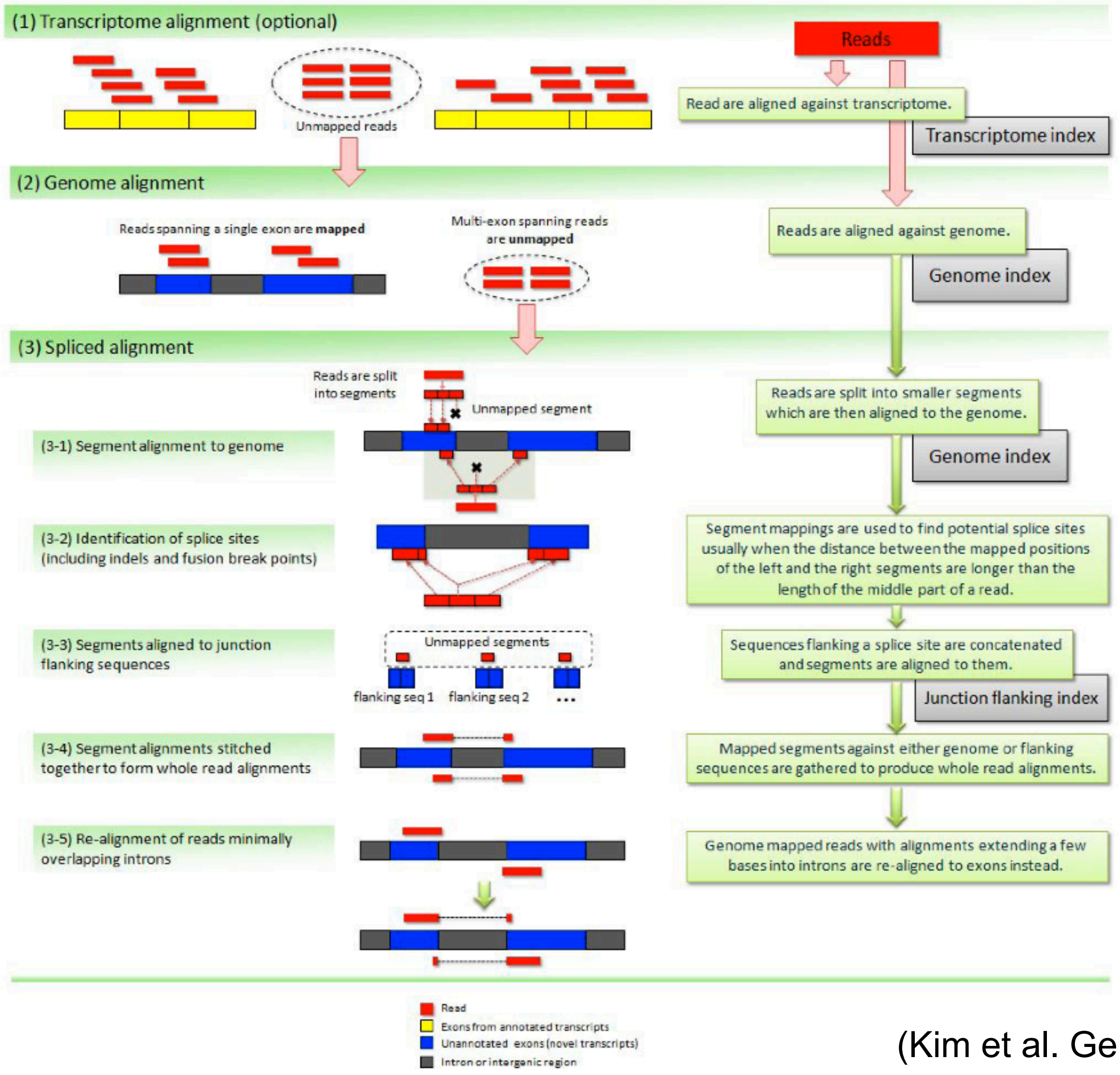


If the algorithm starts to map reads continuously onto the genome

Pseudogenes and spliced mapping



Spliced mapping : Tophat2 pipeline



(1) Transcriptome alignment (optional)

Reads

Read are aligned against transcriptome.

Transcriptome index

(2) Genome alignment

Reads are aligned against genome.

Genome index

(3) Spliced alignment

Reads are split into smaller segments which are then aligned to the genome.

Genome index

(3-1) Segment alignment to genome

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

Segment mappings are used to find potential splice sites usually when the distance between the mapped positions of the left and the right segments are longer than the length of the middle part of a read.

(3-3) Segments aligned to junction flanking sequences

Sequences flanking a splice site are concatenated and segments are aligned to them.

Junction flanking index

(3-4) Segment alignments stitched together to form whole read alignments

Mapped segments against either genome or flanking sequences are gathered to produce whole read alignments.

(3-5) Re-alignment of reads minimally overlapping introns

Genome mapped reads with alignments extending a few bases into introns are re-aligned to exons instead.

Unmapped reads
Reads spanning a single exon are mapped
Multi-exon spanning reads are unmapped
Reads are split into segments
Unmapped segment
Unmapped segments
flanking seq 1
flanking seq 2
...
🟥 Read
🟨 Exons from annotated transcripts
🟦 Unannotated exons (novel transcripts)
⬛ Intron or intergenic region

Unspliced alignment

Spliced alignment

(Kim et al. Genome Biology 2013,14:R36)

NGS read mapping

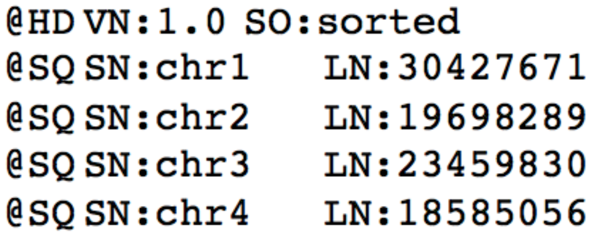
- Introduction
- Short read mappers
- Specificity of RNA-seq read mapping
- Alignment and related file formats
- Alignment visualization

Alignment file format : SAM

- Sequence Alignment/Map format → standard alignment format
- Text file containing all information about an alignment
- SAM format specifications
  - Li et al., Bioinformatics 2009;25(16):2078-9.
  - http://samtools.github.io/hts-specs/SAMv1.pdf

- Header section
  - Generic information regarding the SAM file, not required
  - Each line starts with @ and is tab-delimited
  - @HD : SAM file version, whether the file is sorted
  - @SQ : Name + length of reference sequences used for alignment
  - …

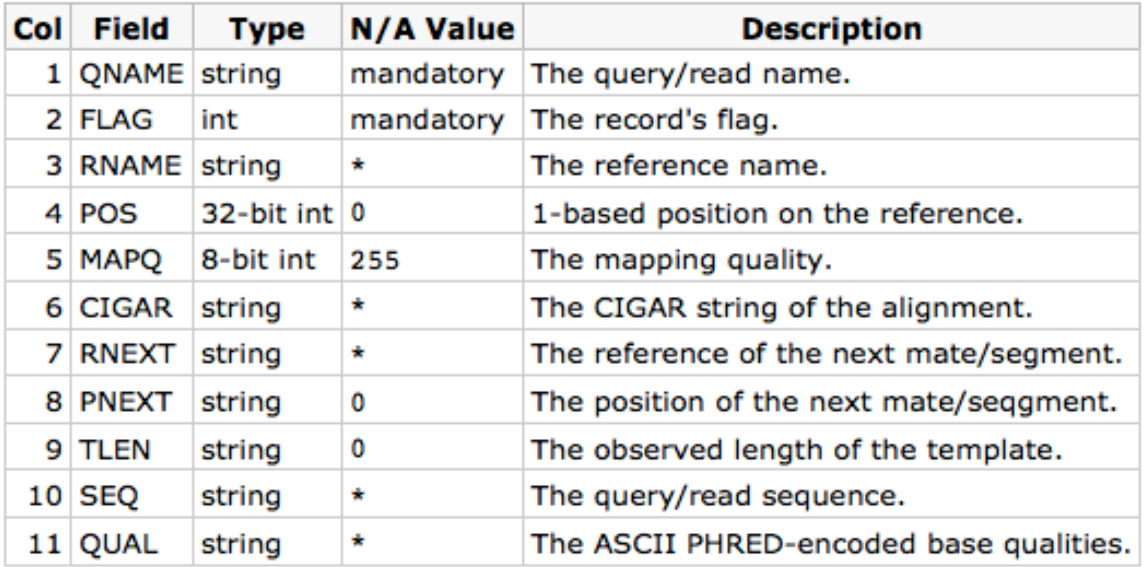
Header section example :



```
@HD VN:1.0 SO:sorted
@SQ SN:chr1    LN:30427671
@SQ SN:chr2    LN:19698289
@SQ SN:chr3    LN:23459830
@SQ SN:chr4    LN:18585056
```

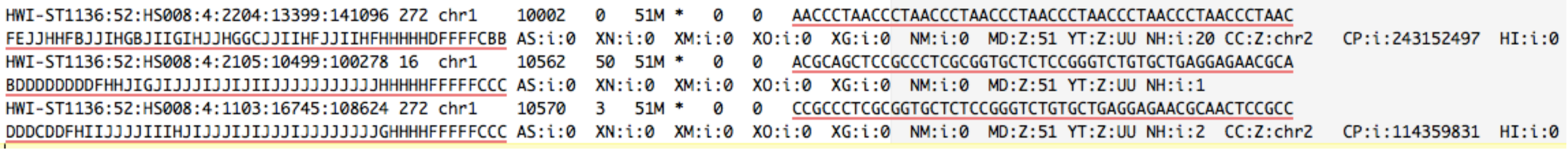
Alignment file format : SAM

- Alignment section : 11 mandatory fields + optional fields
- Mandatory fields :



Col	Field	Type	N/A Value	Description
1	QNAME	string	mandatory	The query/read name.
2	FLAG	int	mandatory	The record's flag.
3	RNAME	string	*	The reference name.
4	POS	32-bit int	0	1-based position on the reference.
5	MAPQ	8-bit int	255	The mapping quality.
6	CIGAR	string	*	The CIGAR string of the alignment.
7	RNEXT	string	*	The reference of the next mate/segment.
8	PNEXT	string	0	The position of the next mate/seqgment.
9	TLEN	string	0	The observed length of the template.
10	SEQ	string	*	The query/read sequence.
11	QUAL	string	*	The ASCII PHRED-encoded base qualities.

Alignment section example :



```
HWI-ST1136:52:HS008:4:2204:13399:141096 272 chr1    10002   0   51M *   0   0   AACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAAC
FEJJHHFBJJIHGBJIIGIHJJHGGCJJIIHFJJIIHFHHHHHDFFFFCBB AS:i:0  XN:i:0  XM:i:0  XO:i:0  XG:i:0  NM:i:0  MD:Z:51 YT:Z:UU NH:i:20 CC:Z:chr2   CP:i:243152497  HI:i:0
HWI-ST1136:52:HS008:4:2105:10499:100278 16  chr1    10562   50  51M *   0   0   ACGCAGCTCCGCCCTCGCGGTGCTCTCCGGGTCTGTGCTGAGGAGAACGCA
BDDDDDDDDFHHJIGJIJJJIJJIJIIJJJJJJJJJJHHHHHFFFFFCCC AS:i:0  XN:i:0  XM:i:0  XO:i:0  XG:i:0  NM:i:0  MD:Z:51 YT:Z:UU NH:i:1
HWI-ST1136:52:HS008:4:1103:16745:108624 272 chr1    10570   3   51M *   0   0   CCGCCCTCGCGGTGCTCTCCGGGTCTGTGCTGAGGAGAACGCAACTCCGCC
DDDCDDFHIIJJJJIIIHJIJJJIJIJJJIJJJJJJJJGHHHHFFFFFCCC AS:i:0  XN:i:0  XM:i:0  XO:i:0  XG:i:0  NM:i:0  MD:Z:51 YT:Z:UU NH:i:2  CC:Z:chr2   CP:i:114359831  HI:i:0
```

Alignment file format : SAM

- Flag (number)

  Describes the alignment

  e.g. reverse strand, not primary alignment, unmapped

  Explain SAM flags in plain English :

  https://broadinstitute.github.io/picard/explain-flags.html

- Mapping quality (number)

  Score indicating whether the read is correctly mapped to this location in the reference genome (different between aligners)

- CIGAR (string)

  Which bases align with the reference (M)
  are deleted from the reference (D)
  correspond to insertions that are not in the reference (I)

Alignment file format : SAM

- CIGAR example
  - Alignment :

```
Reference → C A T A C T _ G A A C T G A C T A A C
     Read →       A C T A G A A _ T G G C T
```

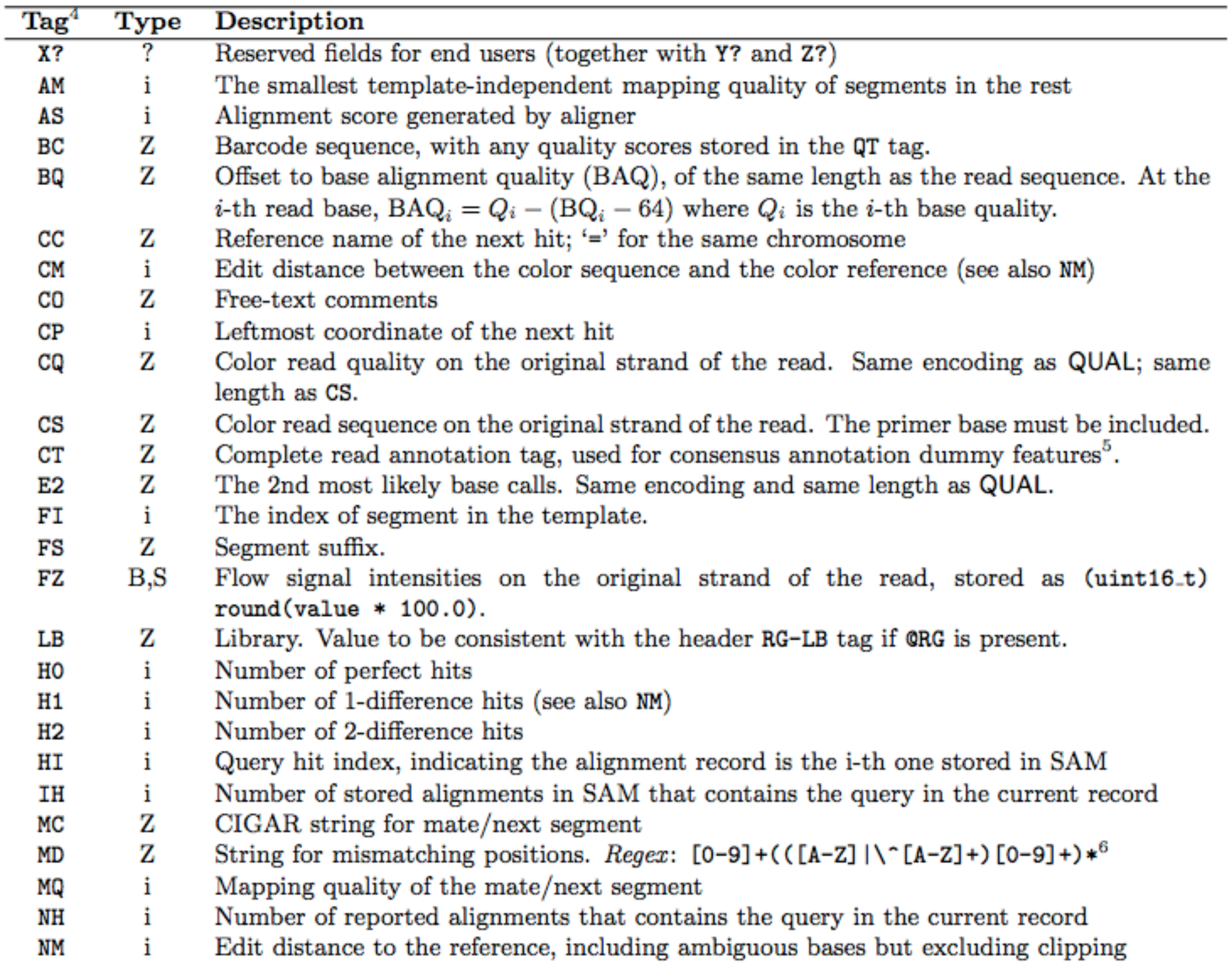
  - CIGAR :

```
3M1I3M1D5M
```

    - 3M : the first 3 bases in the read sequence align with the reference
    - 1I : the next base in the read does not exist in the reference
    - 3M : then 3 bases align with the reference
    - 1D : the next reference base does not exist in the read sequence
    - 5M : then 5 more bases align with the reference
      - Note that among these bases one is different from the reference but it still counts as an M since it aligns to that position

Alignment file format : SAM

- Additional tags (format tag:type:value)



Tag[4]	Type	Description
X?	?	Reserved fields for end users (together with Y? and Z?)
AM	i	The smallest template-independent mapping quality of segments in the rest
AS	i	Alignment score generated by aligner
BC	Z	Barcode sequence, with any quality scores stored in the QT tag.
BQ	Z	Offset to base alignment quality (BAQ), of the same length as the read sequence. At the i-th read base, $\mathrm{BAQ}_i = Q_i - (\mathrm{BQ}_i - 64)$ where Q_i is the i-th base quality.
CC	Z	Reference name of the next hit; '=' for the same chromosome
CM	i	Edit distance between the color sequence and the color reference (see also NM)
CO	Z	Free-text comments
CP	i	Leftmost coordinate of the next hit
CQ	Z	Color read quality on the original strand of the read. Same encoding as QUAL; same length as CS.
CS	Z	Color read sequence on the original strand of the read. The primer base must be included.
CT	Z	Complete read annotation tag, used for consensus annotation dummy features[5].
E2	Z	The 2nd most likely base calls. Same encoding and same length as QUAL.
FI	i	The index of segment in the template.
FS	Z	Segment suffix.
FZ	B,S	Flow signal intensities on the original strand of the read, stored as (uint16_t) round(value * 100.0).
LB	Z	Library. Value to be consistent with the header RG-LB tag if @RG is present.
H0	i	Number of perfect hits
H1	i	Number of 1-difference hits (see also NM)
H2	i	Number of 2-difference hits
HI	i	Query hit index, indicating the alignment record is the i-th one stored in SAM
IH	i	Number of stored alignments in SAM that contains the query in the current record
MC	Z	CIGAR string for mate/next segment
MD	Z	String for mismatching positions. *Regex*: [0-9]+(([A-Z]|\^[A-Z]+)[0-9]+)*[6]
MQ	i	Mapping quality of the mate/next segment
NH	i	Number of reported alignments that contains the query in the current record
NM	i	Edit distance to the reference, including ambiguous bases but excluding clipping

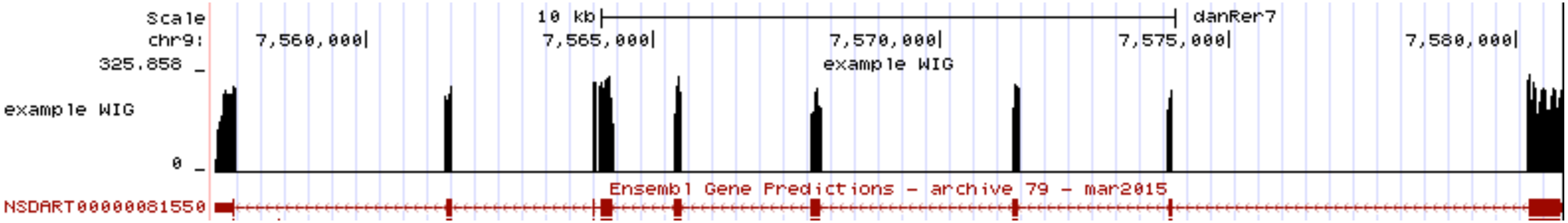
Alignment file format : BAM & samtools

- BAM
  - Binary file
  - Compressed version of SAM format
  - BAM files can be sorted and indexed
    - Makes accessing data very fast
  - BAI (extension .bai) : index for a BAM file
    - sample.bam.bai index for sample.bam file

- Samtools
  - Various utilities for manipulating alignment in SAM format (SAM <> BAM, sorting, indexing, variant calling, calculating statistics on alignments, …)
  - http://www.htslib.org/

Wiggle (WIG) file format

- Tab-delimited text file
- “Summary” generated from an alignment
- For dense continuous data (eg coverage)
- Each line represents a portion of a chromosome
- Columns :
  - Chromosome
  - Start
  - End
  - Value
- More precise definition and examples
  - http://genome.ucsc.edu/goldenPath/help/wiggle.html



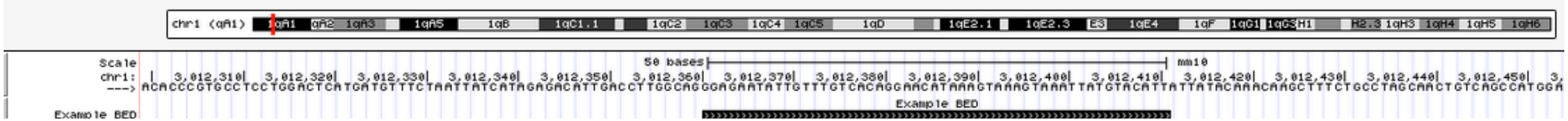
Browser Extensible Data (BED) format

- Tab-delimited text file
- For genomic intervals
- From 3 to 12 columns (always in this order) :
  - Chromosome
  - Start
  - End
  - Name
  - Score
  - Strand (+ or -)
  - …

required

Most common :
6 columns

- More precise definition and examples
  - http://genome.ucsc.edu/FAQ/FAQformat.html#format1
- Manipulation of BED files
  - BEDTools : http://code.google.com/p/bedtools/

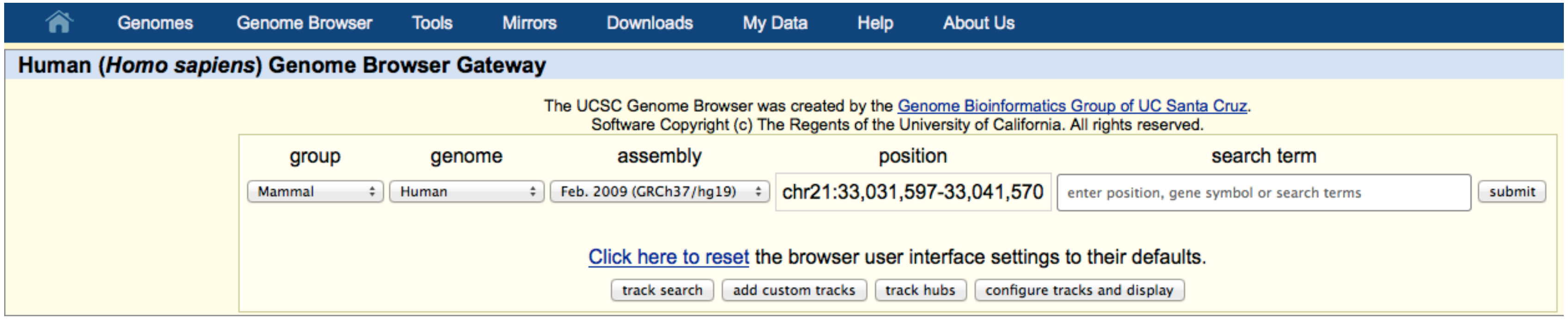


NGS read mapping

- Introduction
- Short read mappers
- Specificity of RNA-seq read mapping
- Alignment and related file formats
- Alignment visualization

Alignment visualization

- Using a Genome Browser
  - UCSC : http://genome.ucsc.edu



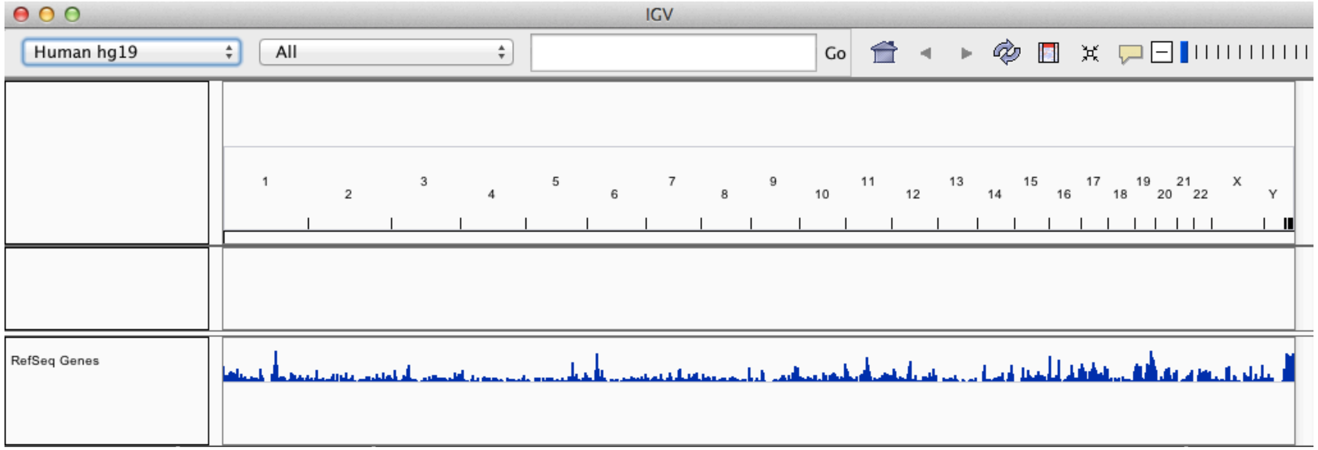
Human (*Homo sapiens*) Genome Browser Gateway

The UCSC Genome Browser was created by the Genome Bioinformatics Group of UC Santa Cruz.
Software Copyright (c) The Regents of the University of California. All rights reserved.

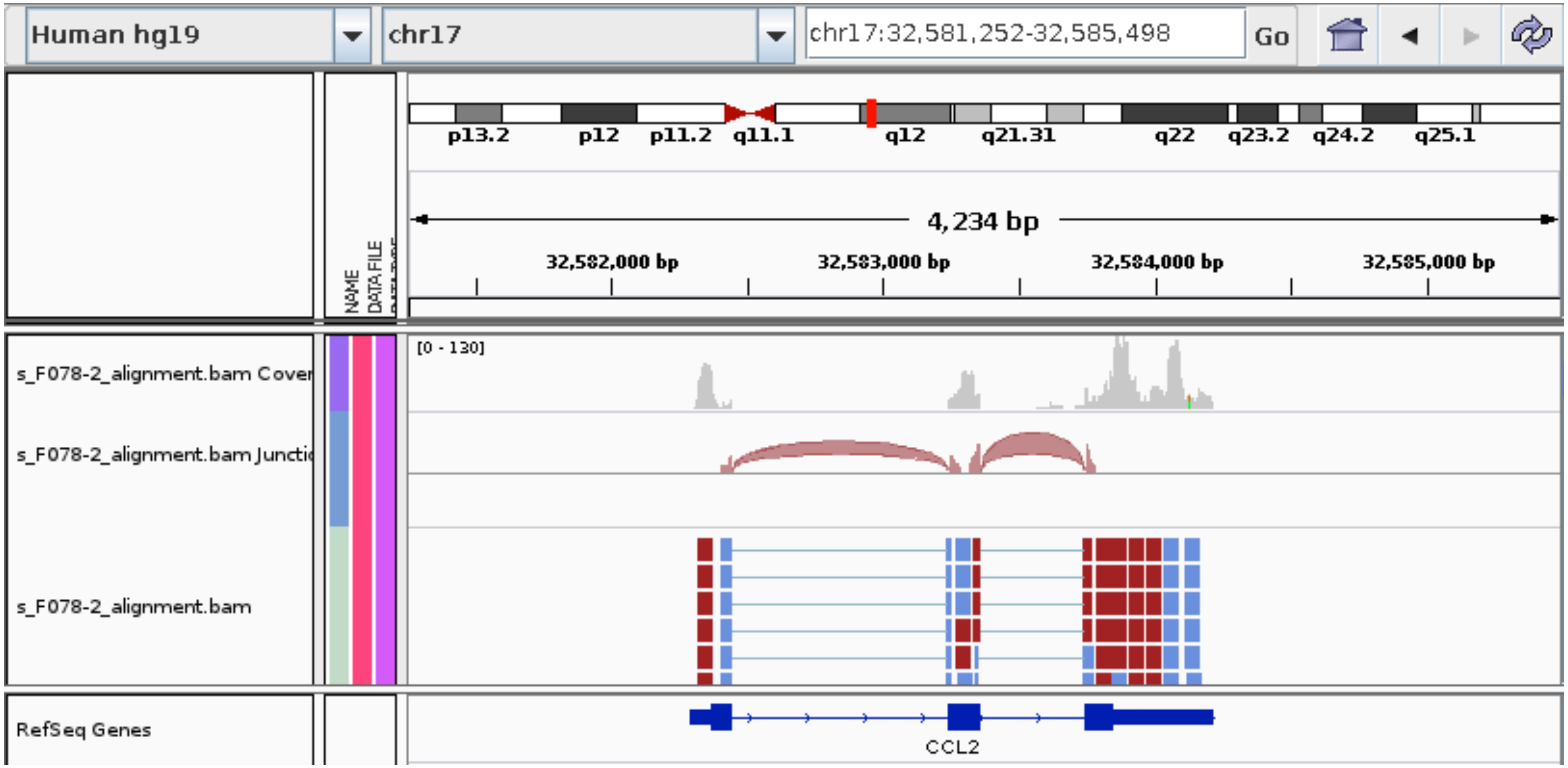
group	genome	assembly	position	search term	
Mammal	Human	Feb. 2009 (GRCh37/hg19)	chr21:33,031,597-33,041,570	enter position, gene symbol or search terms	submit

Click here to reset the browser user interface settings to their defaults.

  - IGV : http://www.broadinstitute.org/igv/



Integrative Genomics Viewer



Exercise

- We will work on the 4 RNA-seq samples from MITF project
- These samples have been aligned on hg19 human genome assembly using Tophat2
  - Summary of results :

Sample ID	Sample name	Total number of reads	% of aligned reads	% of uniquely aligned reads	% of multiple aligned reads
TSB-11_5_S1	siLuc2	44,340,015	96.45	89.06	7.39
TSB-12_6_S1	siLuc3	49,763,265	96.84	89.57	7.28
TSB-13_19_S2	siMitf3	42,595,950	96.48	89.14	7.34
TSB-14_12_S2	siMitf4	39,065,527	96.86	89.46	7.40

- Select the appropriate genome assembly and load the 4 BAM files TSB-*_mrnaseq_noSpikes_alignment.bam into IGV

Exercise

1. A ChIP-seq peak has previously been identified near IDH1 gene. Is this gene differentially expressed between siLuc and siMitf samples ?
2. In the last exon of this gene, can you identify a nucleotide difference in the RNA-seq samples compared to the reference genome ? What is the exact position of this difference ?
3. The same RNA samples have been processed with a different RNA-seq protocol.

   The corresponding BAM file for the first sample is : TSB-11_5_S1_rnaseq_noSpikes_alignment_2ndprotocol.bam

   Load this BAM file into IGV.
   Search for a difference between the two protocols used.

   Advise : right-click on the tracks corresponding to BAM files and look at the “Color alignments by” menu

Exercise

4. Look at the splice junctions identified in ACP5 gene.
Are all these junctions annotated in Refseq ? And in Ensembl ?

Advises :

- File → New session
- View → Preferences → Alignments tab → Splice Junction Track Options panel :
  - Show junction track
  - e.g. Min flanking width=2 / Min junction coverage=10
- File → Load from file and select the 4 BAM files TSB-*_mrnaseq_noSpikes_alignment.bam
- Expand the Refseq track
  - Right-click on the track → Expanded
- You can also perform a Sashimi-plot for a better visualization of these junctions :
  - Right-click on a BAM track → Sashimi plot → Select Gene track : Refseq genes → Select Alignment Tracks : all alignments