# NGS read mapping

Céline Keime
keime@igbmc.fr

# NGS read mapping

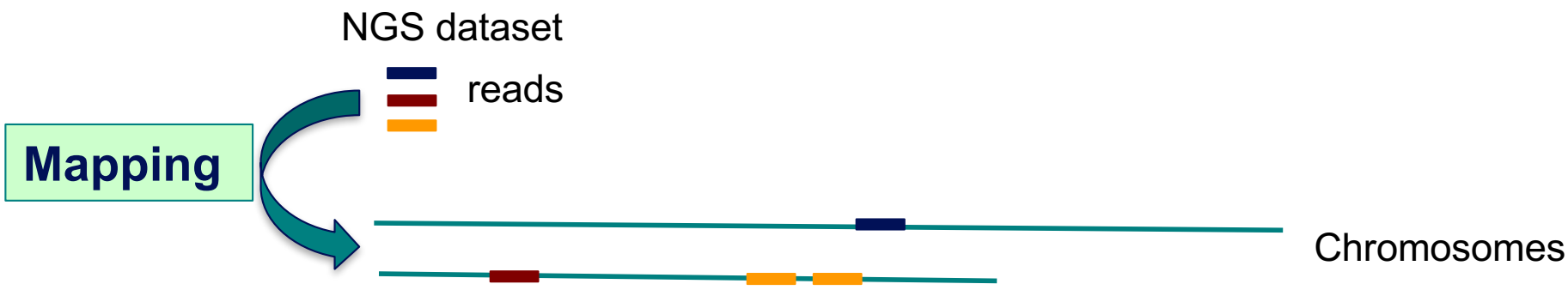
- Introduction to NGS read mapping
- Short read mappers
- Specificity of RNA-seq read mapping
- *Exercise 1 : Mapping of RNA-seq data using Galaxy*
- Alignment and related file formats
- Alignment visualization
- *Exercise 1 : Interpretation of results*
- *Exercise 2 : Whole dataset alignment visualization*
- Quality control of RNA-seq data based on alignments

# NGS read mapping

- Introduction to NGS read mapping
- Short read mappers
- Specificity of RNA-seq read mapping
- *Exercise 1 : Mapping of RNA-seq data using Galaxy*
- Alignment and related file formats
- Alignment visualization
- *Exercise 1 : Interpretation of results*
- *Exercise 2 : Whole dataset alignment visualization*
- Quality control of RNA-seq data based on alignments

# 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 the 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

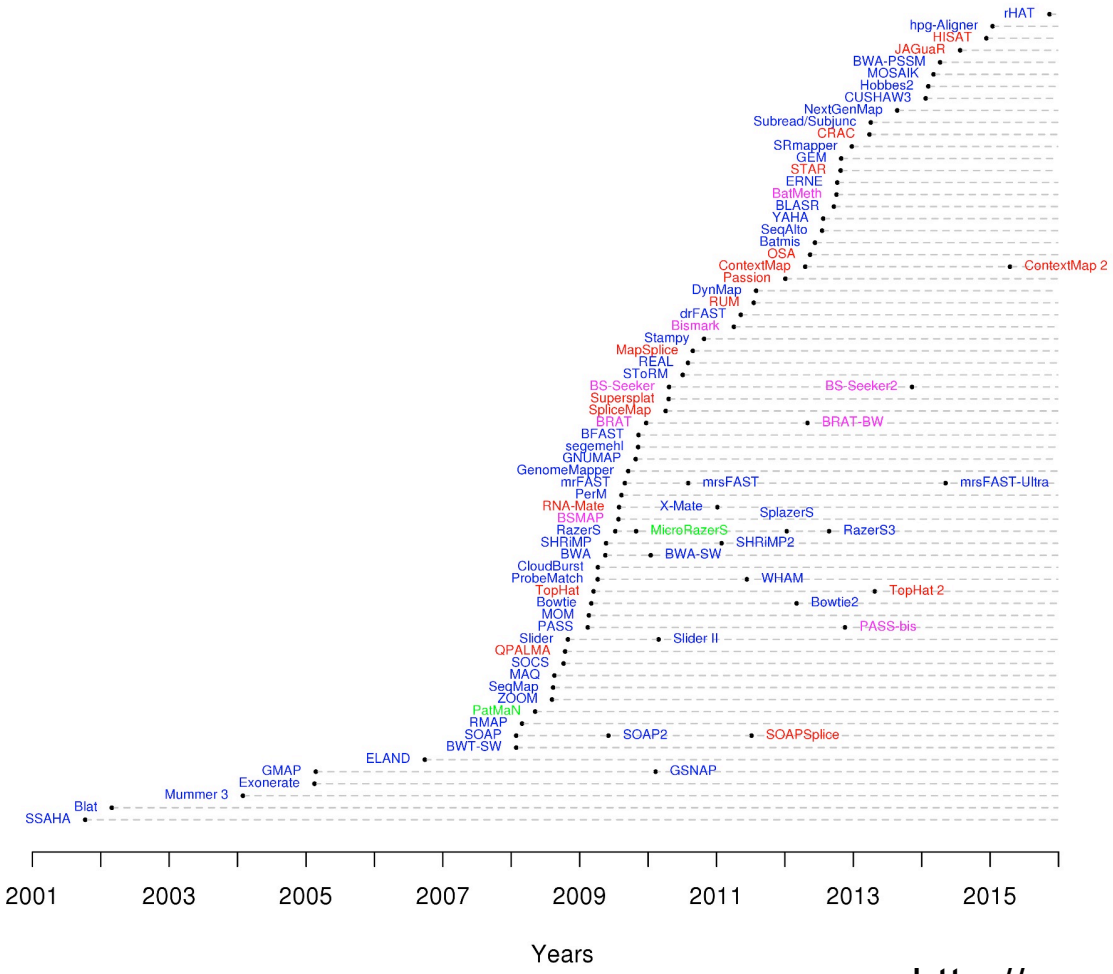
➔ Specialized read mappers with highly efficient algorithms

# NGS read mapping

- Introduction to NGS read mapping
- Short read mappers
- Specificity of RNA-seq read mapping
- *Exercise 1 : Mapping of RNA-seq data using Galaxy*
- Alignment and related file formats
- Alignment visualization
- *Exercise 1 : Interpretation of results*
- *Exercise 2 : Whole dataset alignment visualization*
- Quality control of RNA-seq data based on alignments

# A lot of tools developed …

- More than 90 mapping tools




DNA mappers
RNA mappers
miRNA mappers
bisulfite mappers

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

- Transforming
  - Uses a technique originally developed for compressing large files called the Burrows-Wheeler transform
    ➔ The transformed human genome fits into 2GB of memory
  - Align a read character by character to the transformed genome

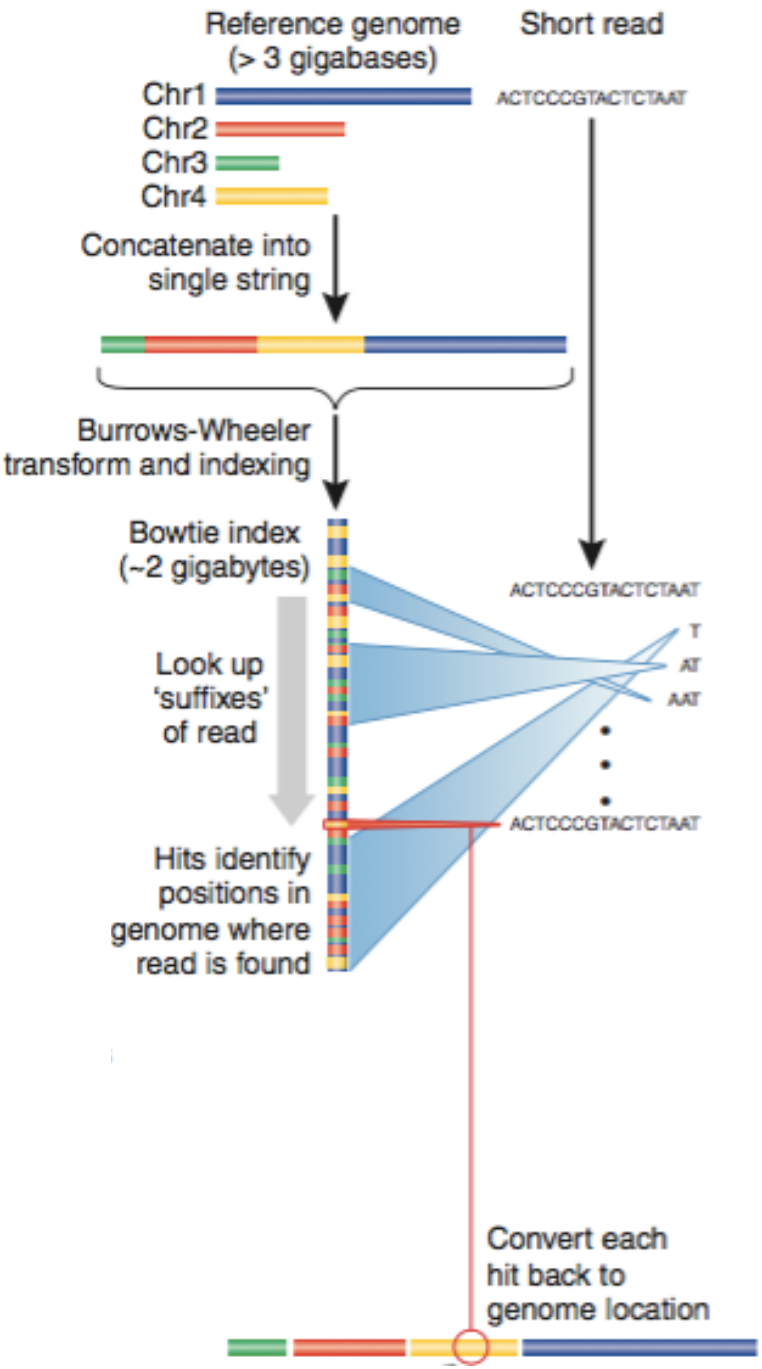
# Bowtie method

- Stores a memory-efficient representation of the reference genome

- Aligns a read one character at a time to the transformed genome

→ Each successively aligned new character allows Bowtie to winnow the list of positions to which the read might map

→ If Bowtie cannot find a location where a read aligned perfectly, the algorithm backtrack to the previous character, makes a substitution and resumes the search




*From Trapnell et al., Nature Biotechnology 2009; 27(5): 455-457*

# Bowtie features

- Input : DNA in Fasta/Fastq format (single-read or paired-end)
- Allows mismatches, indels, gaps (only bowtie2)
- Quality-aware
- Output : SAM, tsv
- When multiple alignments, reports either all, best, random or alignments with at least a user defined number of matches
- Main differences between bowtie1 and bowtie2
  - Bowtie2 indexes the genome with an index based on the Burrows-Wheeler transform
  - For reads longer than 50bp, bowtie2 is generally faster, more sensitive and uses less memory than bowtie1
    For shorter reads, bowtie1 is sometimes faster and/or more sensitive
  - Bowtie2 supports gapped alignment (in contrary to bowtie1)
  - There is no upper limit on read length in bowtie2 (upper limit in bowtie1 ~ 1kb)
  - Paired-end alignment more flexible in bowtie2 (for pairs that do not aligned in a paired fashion, bowtie2 attempts to find unpaired alignments for each mate)
  - Bowtie2 does not align colorspace reads (in contrary to bowtie1)

# 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

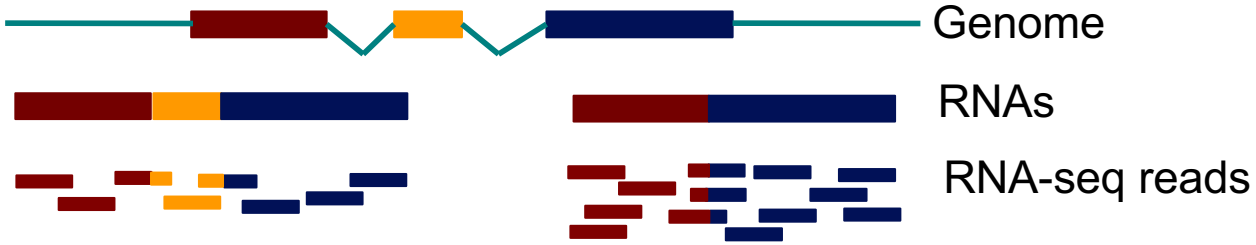
- Several comparative analyses
  - Very interesting to start with :

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

# NGS read mapping

- Introduction to NGS read mapping
- Short read mappers
- Specificity of RNA-seq read mapping
- *Exercise 1 : Mapping of RNA-seq data using Galaxy*
- Alignment and related file formats
- Alignment visualization
- *Exercise 1 : Interpretation of results*
- *Exercise 2 : Whole dataset alignment visualization*
- Quality control of RNA-seq data based on alignments

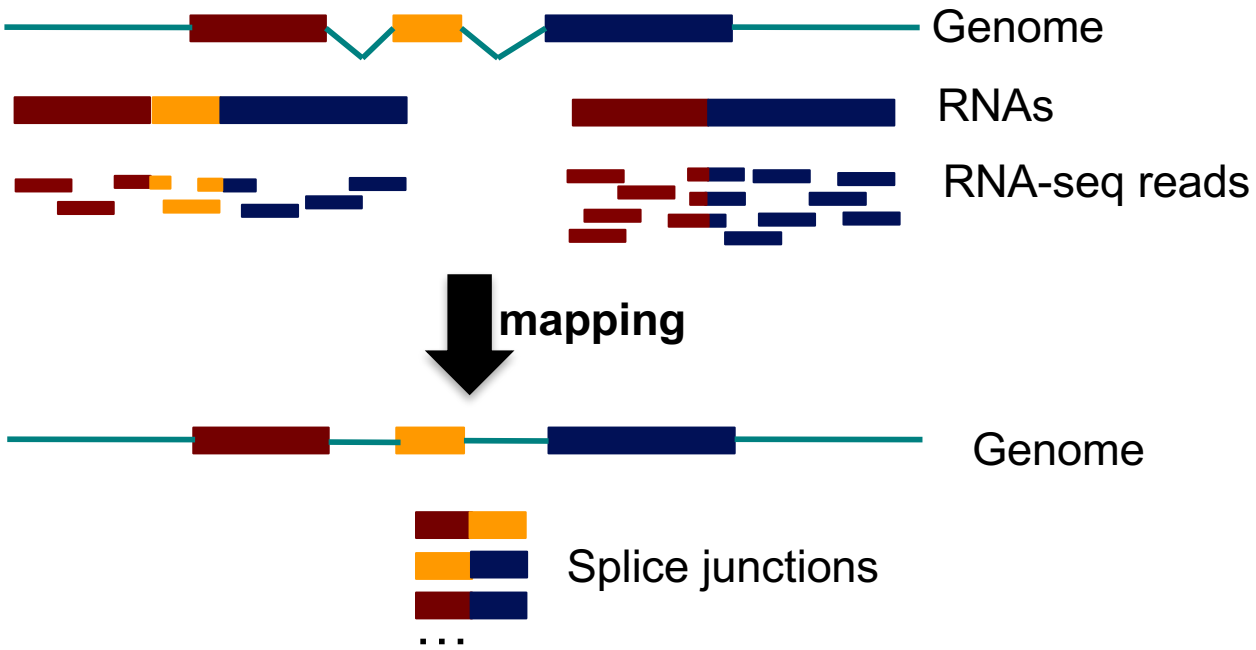
# Specificity of RNA-seq reads




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

# Map onto the genome and splice junctions ?

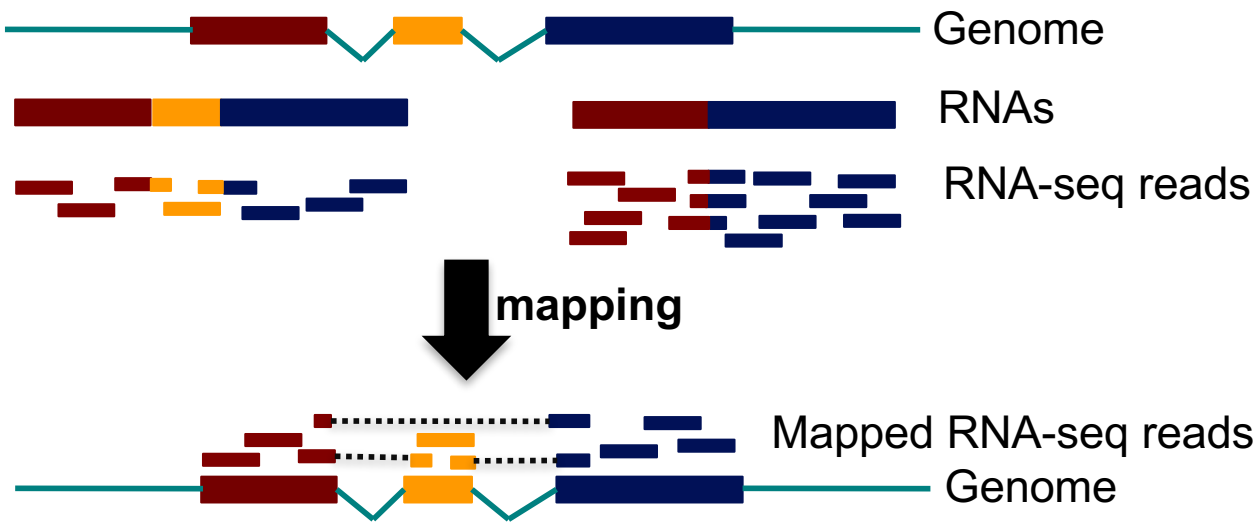
- ERANGE, RNA-Mate




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

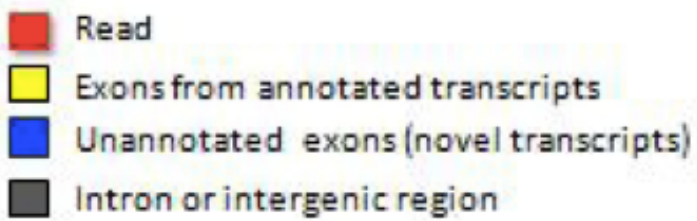
# Spliced mapping

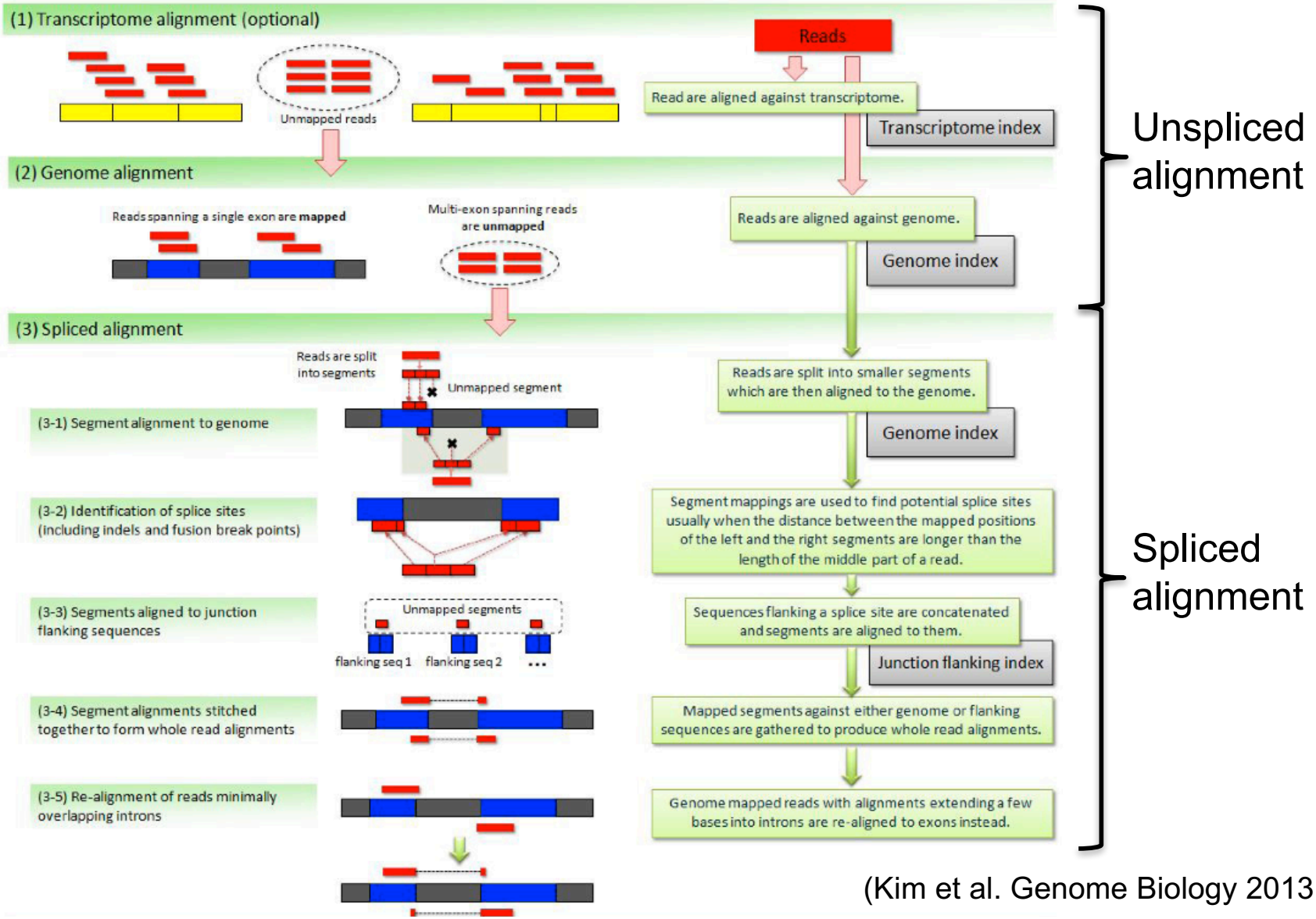
- Allows mapping of reads across splice junctions




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

# Spliced mapping : Tophat2 pipeline




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

# Genome annotations

- Generally provided in a GTF/GFF file
  - Tab-delimited text file format
  - Each line correspond to an annotation or feature
  - Each line has nine columns :

| Seqname | Source | Feature | Start | End | Score | Strand | Frame | Attribute |
|---|---|---|---|---|---|---|---|---|
| 2 | ensembl_havana | gene | 227813842 | 227817564 | . | + | . | |
| 2 | havana | transcript | 227813842 | 227817564 | . | + | . | |
| 2 | havana | exon | 227813842 | 227813987 | . | + | . | |
| 2 | havana | CDS | 227813912 | 227813987 | . | + | 0 | |
| 2 | havana | start_codon | 227813912 | 227813914 | . | + | 0 | |
| 2 | havana | exon | 227815457 | 227815568 | . | + | . | |
| 2 | havana | CDS | 227815457 | 227815568 | . | + | 2 | |

gene_id "ENSG00000115009"; gene_version "11"; transcript_id "ENST00000409189"; transcript_version "7"; exon_number "1"; gene_name "CCL20"; gene_source "ensembl_havana"; gene_biotype "protein_coding"; havana_gene "OTTHUMG00000133189"; havana_gene_version "3"; transcript_name "CCL20-001"; transcript_source "havana"; transcript_biotype "protein_coding"; …

# Genome annotations

- Ensembl project (www.ensembl.org)
  - Goal : automatically annotate the genome, integrate this annotation with other available biological data and make all this publicly available
  - Ensembl data is released on an approximately three-month cycle
- Ensembl genome annotations available on
  - ftp://ftp.ensembl.org/pub/
  - Important to use the same annotation version throughout a project (possible to access to old versions View in archive site )
  - Annotations for some species and Ensembl version already available on GalaxEast
- The main Ensembl site focuses on vertebrate genomes (87 species), other sites are dedicated to other metazoan genomes, plants, fungi, bacteria, … (http://www.ensembl.org/info/about/species.html)
- Other annotation sources
  - e.g., ordered from most to least complex : AceView, Ensembl, UCSC, Refseq Genes (Wu et al. BMC Bioinformatics 2013 ;14 Suppl 11:S8)

# NGS read mapping

- Introduction to NGS read mapping
- Short read mappers
- Specificity of RNA-seq read mapping
- *Exercise 1 : Mapping of RNA-seq data using Galaxy*
- Alignment and related file formats
- Alignment visualization
- *Exercise 1 : Interpretation of results*
- *Exercise 2 : Whole dataset alignment visualization*
- Quality control of RNA-seq data based on alignments

# Exercise 1
# Mapping of RNA-seq data using Galaxy

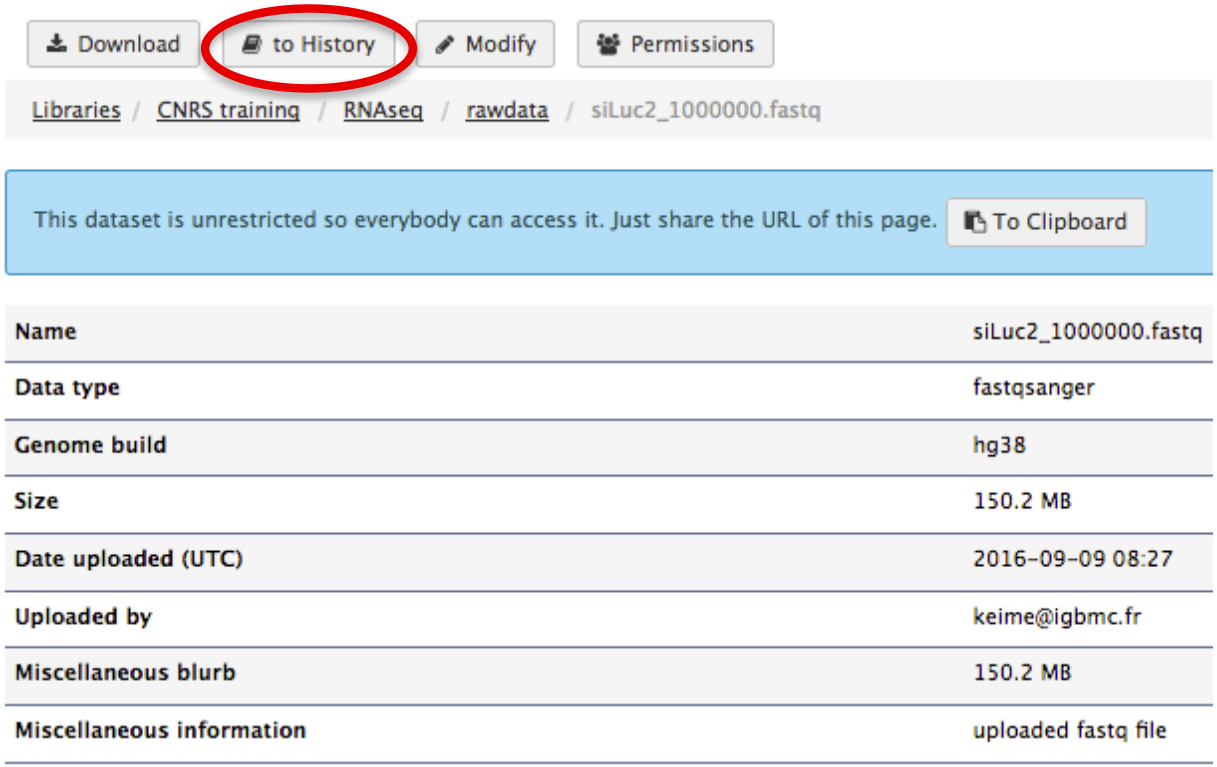
- Map **1 million** reads from siLuc2 mRNA-seq sample using Tophat2 and gene annotations from Ensembl release 85

1. Import the corresponding FASTQ file in your history
2. Import the corresponding gene annotations in your history
3. Launch Tophat2 on this FASTQ file using these annotations

# Exercise 1
# 1. Import FASTQ file in your history

- FASTQ file available in
  - Shared Data → Data Libraries → CNRS training
  - RNAseq → rawdata → **siLuc2_1000000.fastq**

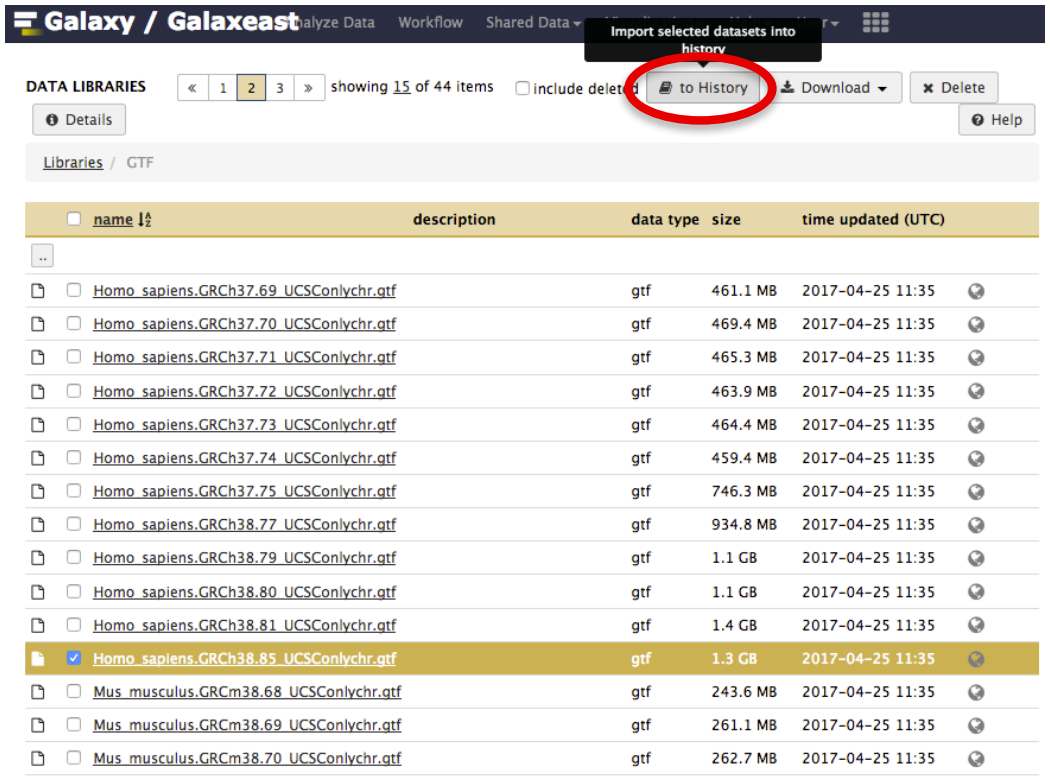
- Import this file in your current history



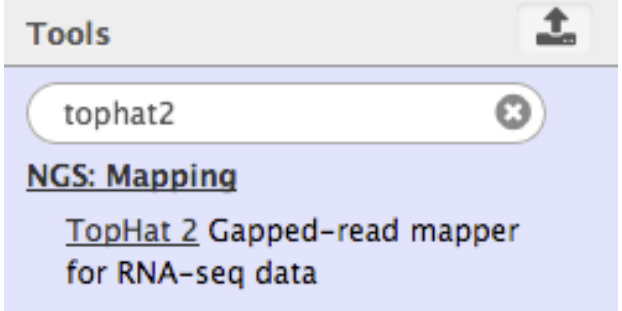
# Exercise 2
# 2. Import gene annotations in your history

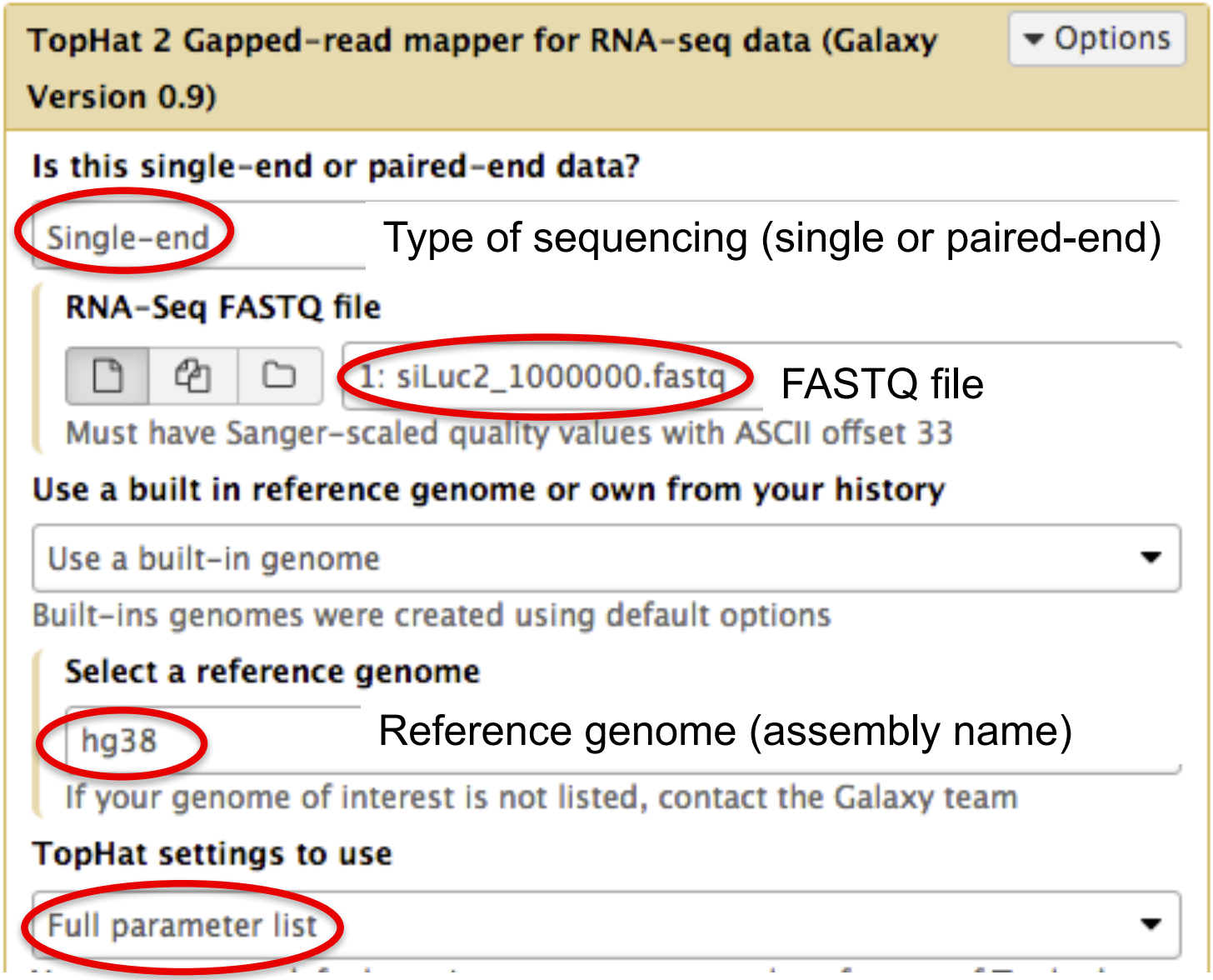
- Gene annotations available in
  - Shared Data → Data Libraries → GTF
  - Homo_sapiens.GRCh38.85_UCSConlychr.gtf

- Import this file in your current history



# Exercise 1
# 3. Launch Tophat2



# Exercise 1
# 3. Launch Tophat2

**TopHat settings to use**

Full parameter list

You can use the default settings or set custom values for any of Tophat's parameters.

**Max realign edit distance**

1000

--read-realign-edit-dist; Some of the reads spanning multiple exons may be mapped incorrectly as a contiguous alignment to the genome even though the correct alignment should be a spliced one - this can happen in the presence of processed pseudogenes that are rarely (if at all) transcribed or expressed. This option can direct TopHat to re-align reads for which the edit distance of an alignment obtained in a previous mapping step is above or equal to this option value. If you set this option to 0, TopHat will map every read in all the mapping steps (transcriptome if you provided gene annotations, genome, and finally splice variants detected by TopHat), reporting the best possible alignment found in any of these mapping steps. This may greatly increase the mapping accuracy at the expense of an increase in running time. The default value for this option is set such that TopHat will not try to realign reads already mapped in earlier steps.

**Max edit distance**

2

--read-edit-dist; Final re

**Library Type**

FR First Strand

--library-type; TopHat w
tag. Consider supplying l

Library preparation method :
Here the libraries have been prepared using a directional protocol where only the strand generated during first strand cDNA synthesis is sequencing
For a non directional protocol choose FR Unstranded

# Exercise 1
# 3. Launch Tophat2

**Do you want to supply your own junction data**

Yes

The options below allow you validate your own list of known transcripts or junctions with your RNA-Seq data. Note that the chromosome names in the files provided with the options below must match the names in the Bowtie index.

**Use Gene Annotation Model**

Yes

**Gene Model Annotations**

2: Homo_sapiens.GRCh38.85_UCSConlychr.gtf

-G/--GTF; TopHat with a set
formatted file. If this option i
to align reads to this virtual t
will then be mapped on the g
genomic mappings (spliced a
tophat output. Please note th
indicates the chromosome or
reference sequence in the Bowtie index you are using with TopHat.

Annotation file
→ Using this file, TopHat2 will first extract the transcript sequences and use Bowtie2 to align reads to this virtual transcriptome first.

**Use Raw Junctions**

No

**Only look for supplied junctions**

No

--no-novel-juncs; Only look for reads across junctions indicated in the supplied GFF or junctions file.

Execute

# NGS read mapping

- Introduction to NGS read mapping
- Short read mappers
- Specificity of RNA-seq read mapping
- *Exercise 1 : Mapping of RNA-seq data using Galaxy*
- Alignment and related file formats
- Alignment visualization
- *Exercise 1 : Interpretation of results*
- *Exercise 2 : Whole dataset alignment visualization*
- Quality control of RNA-seq data based on alignments

# 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   AACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAAC
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

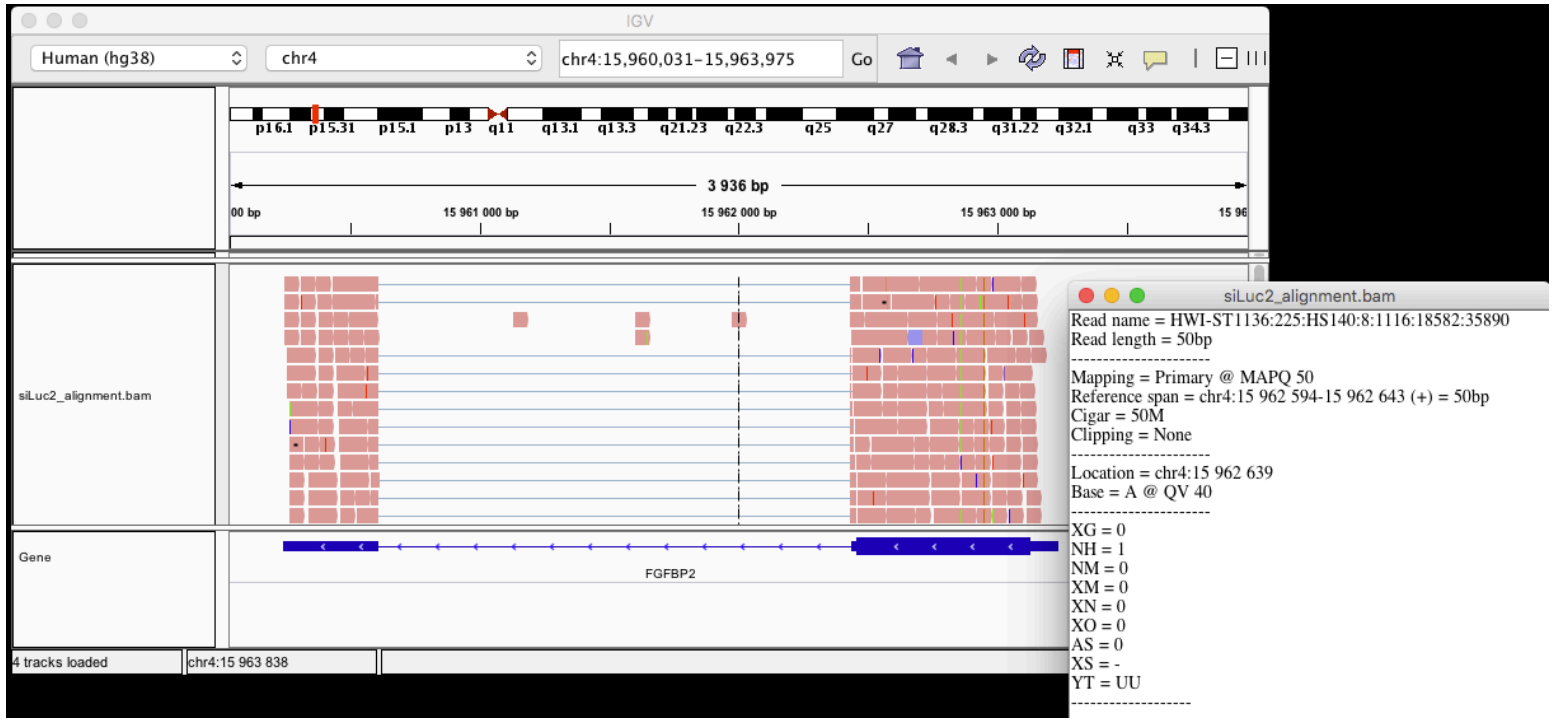
# Alignment file format : SAM

- Additional tags (format tag:type:value)

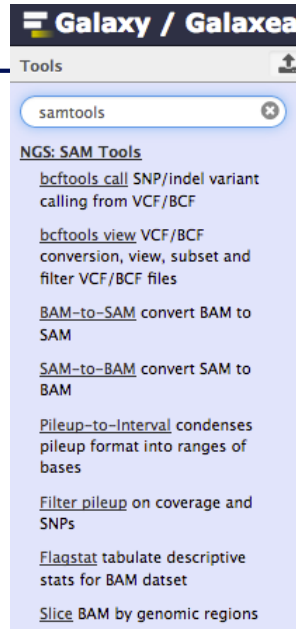
| Tag[4] | Type | Description |
|---|---|---|
| X? | ? | Reserved fields for end users (together with Y? and Z?) |
| AM | i | The smallest template-independent mapping quality of segments in the rest |
| AS | i | Alignment score generated by aligner |
| BC | Z | Barcode sequence, with any quality scores 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

- 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

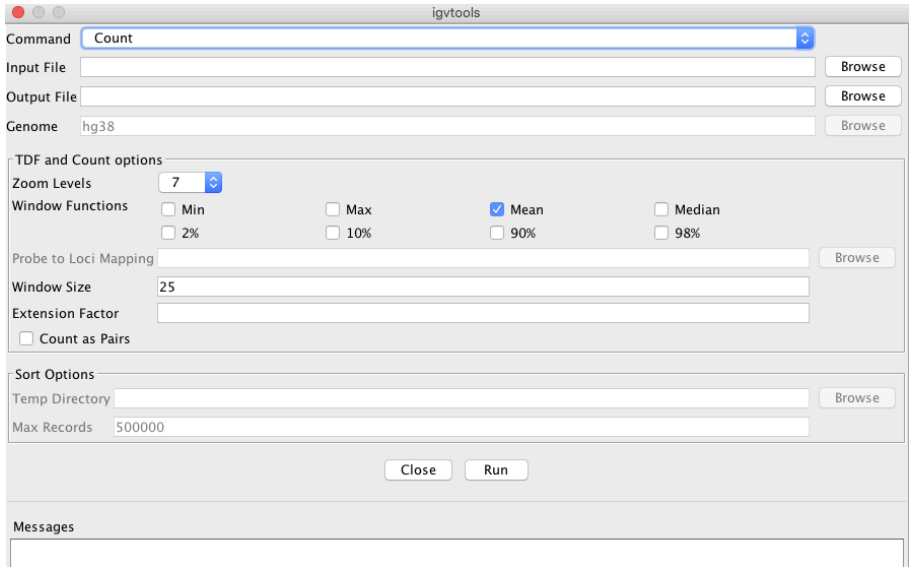


# Utilities to manipulate SAM/BAM files



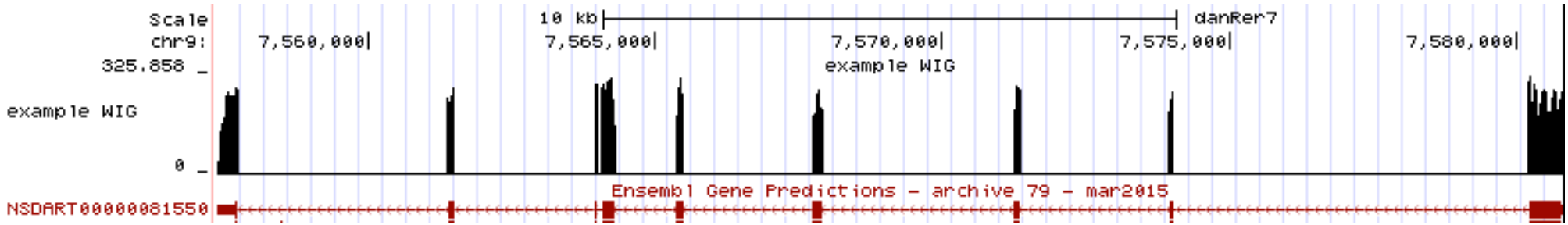

- Samtools (http://www.htslib.org/)
  - Various utilities for manipulating alignment in SAM format (SAM <> BAM conversion, calculating statistics on alignments, …)

- Igvtools (http://software.broadinstitute.org/software/igv/)
  - sort, index, …
  - Integrative Genomics Viewer
    - → Tools menu
    - → run igvtools



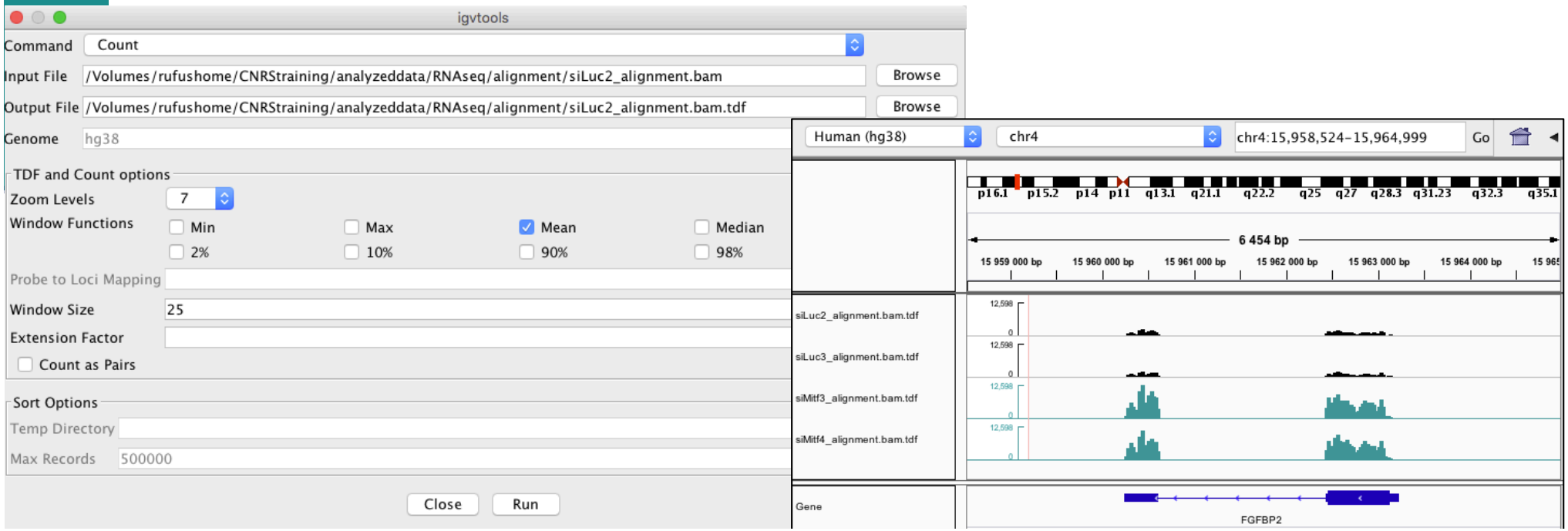
# Wiggle (WIG) file format

- Tab-delimited text file
- For dense continuous data
  - e.g. coverage : “summary” generated from an alignment
    → only density information
- 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



# TDF file format

- Tiled data file
- Binary file
- Read count density
  - Pre-processed data for faster display in IGV
- TDF file can be computed from a BAM file using igvtools
  - IGV Tools menu → run igvtools → Count

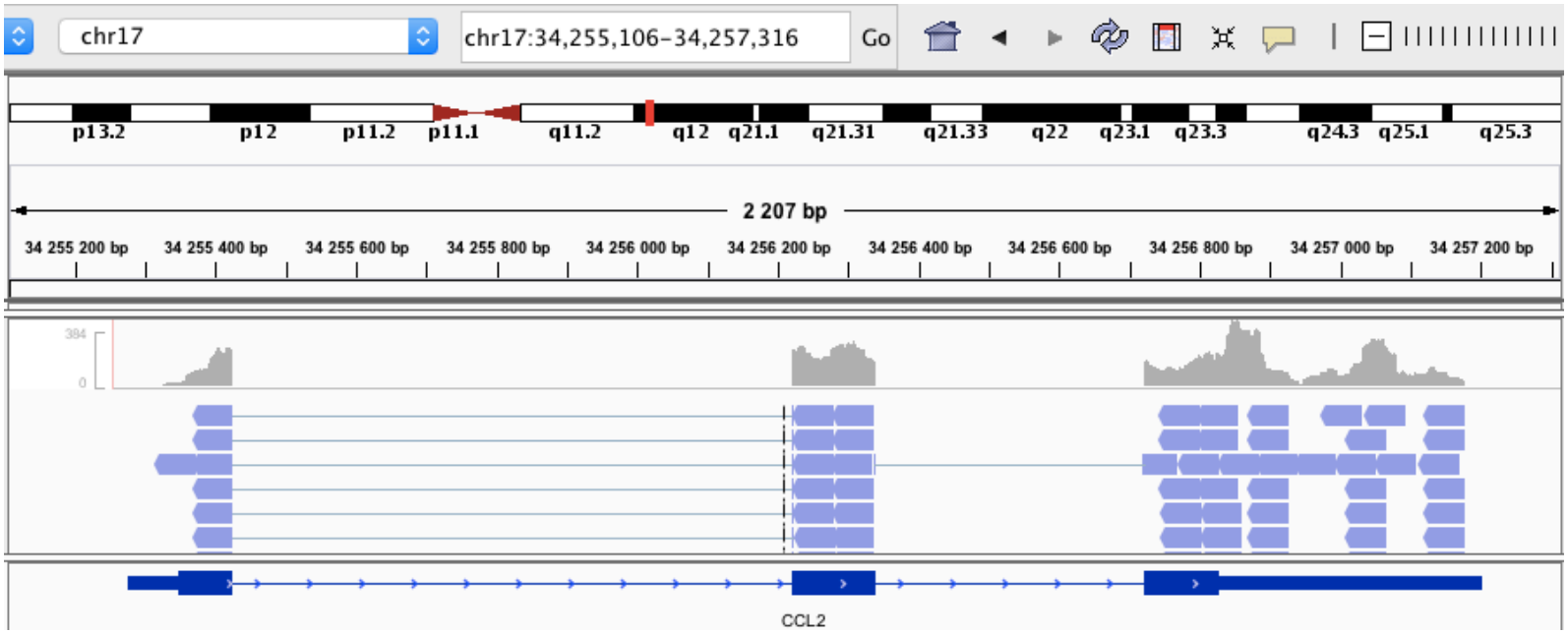


# Coverage vs alignment

Coverage

Alignment

Annotation

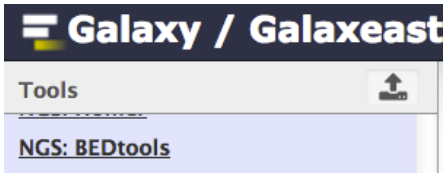


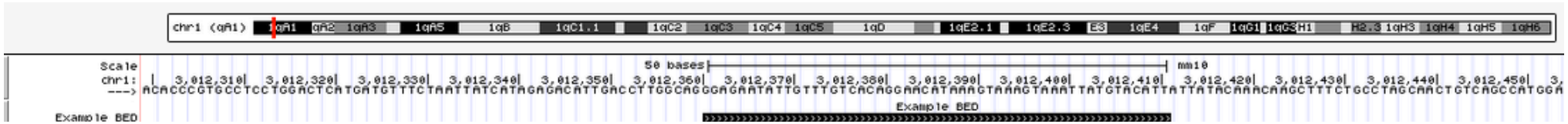
# Browser Extensible Data (BED) format

- 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://bedtools.readthedocs.io



# General Feature Format (GFF)

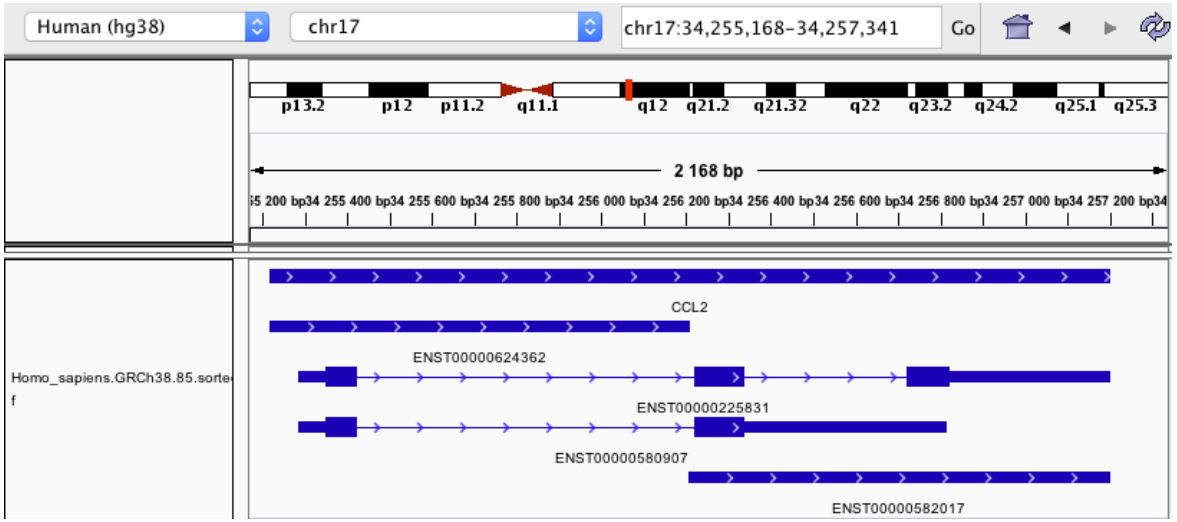
- Text file format to describe genes and other features associated to DNA, RNA and protein sequences
- Specifications
  - https://github.com/The-Sequence-Ontology/Specifications/blob/master/gff3.md
- e.g. human Ensembl 85 GFF file
  - ftp://ftp.ensembl.org/pub/release-85/gff3/homo_sapiens/Homo_sapiens.GRCh38.85.chr.gff3.gz

# General Feature Format (GFF)

- GFF files can be visualized using IGV
  - e.g. Ensembl 85 annotations
- Sort and index for faster display
  - Tools → Run igvtools → Sort

  → Homo_sapiens.GRCh38.85.sorted.gtf
  - Tools → Run igvtools → Index

  → Homo_sapiens.GRCh38.85.sorted.gtf.idx (in the same directory)
  - File → Load from file and choose Homo_sapiens.GRCh38.85.sorted.gtf



# Main NGS file formats : summary

- FASTQ — text
  - Raw data
- SAM / BAM — binary
  - alignment
- WIG / TDF
  - coverage
- BED
  - Genomic intervals
- GFF
  - annotations

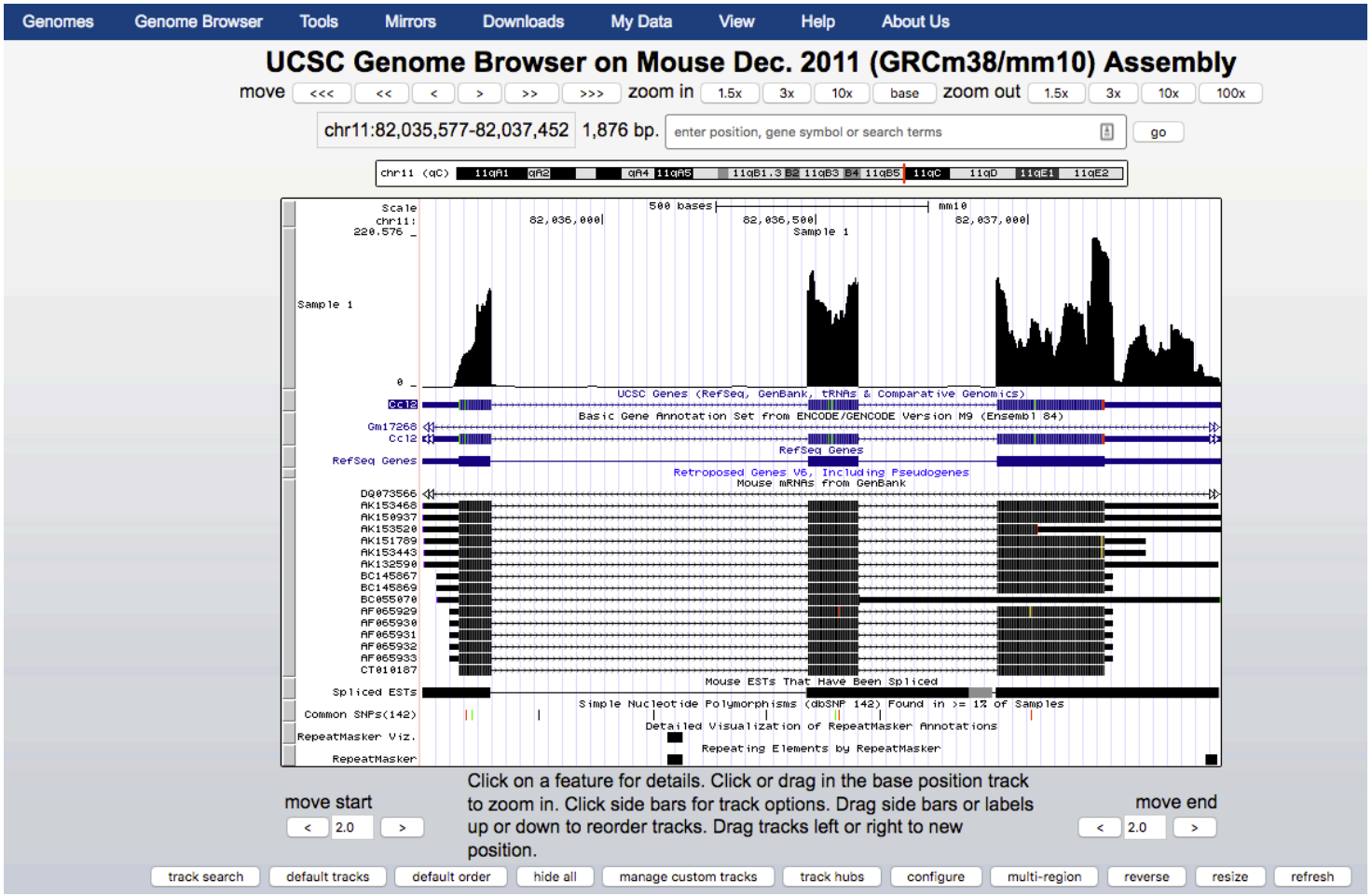
# NGS read mapping

- Introduction to NGS read mapping
- Short read mappers
- Specificity of RNA-seq read mapping
- *Exercise 1 : Mapping of RNA-seq data using Galaxy*
- Alignment and related file formats
- Alignment visualization
- *Exercise 1 : Interpretation of results*
- *Exercise 2 : Whole dataset alignment visualization*
- Quality control of RNA-seq data based on alignments

# Alignment visualization

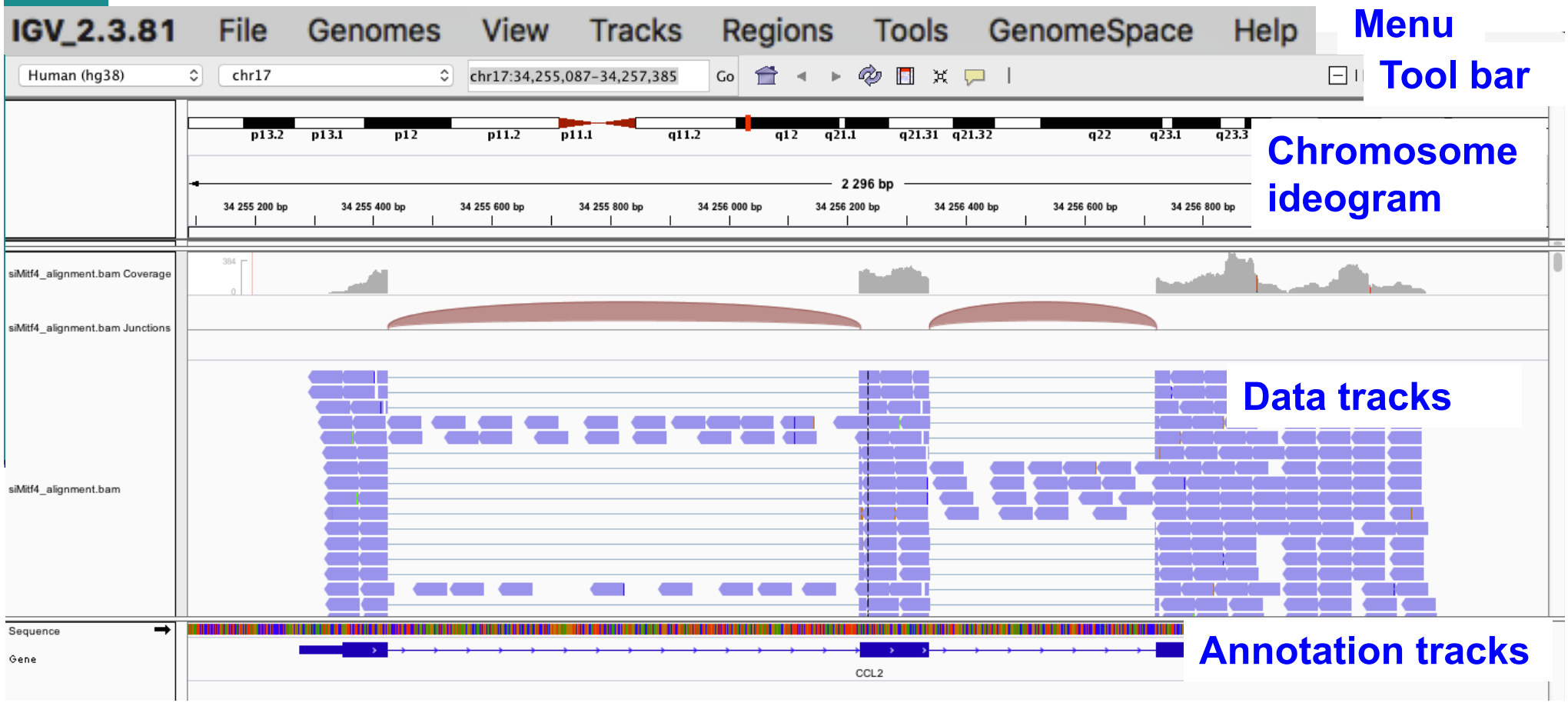
- Using a Genome Browser
  - A lot of available genome browsers
    - Ensembl, UCSC, GBrowse, JBrowse, IGB, IGV, …
  - During this training we will use
    - UCSC : http://genome.ucsc.edu
    - IGV : http://software.broadinstitute.org/software/igv/

# UCSC


Genomes
Genome Browser
Tools
Mirrors
Downloads
My Data
View
Help
About Us
UCSC Genome Browser on Mouse Dec. 2011 (GRCm38/mm10) Assembly
move
zoom in
zoom out
chr11:82,035,577-82,037,452
1,876 bp.
enter position, gene symbol or search terms
go
Sample 1
UCSC Genes (RefSeq, GenBank, tRNAs & Comparative Genomics)
Basic Gene Annotation Set from ENCODE/GENCODE Version M9 (Ensembl 84)
RefSeq Genes
Retroposed Genes V5, Including Pseudogenes
Mouse mRNAs from GenBank
Mouse ESTs That Have Been Spliced
Simple Nucleotide Polymorphisms (dbSNP 142) Found in >= 1% of Samples
Detailed Visualization of RepeatMasker Annotations
Repeating Elements by RepeatMasker
Click on a feature for details. Click or drag in the base position track to zoom in. Click side bars for track options. Drag side bars or labels up or down to reorder tracks. Drag tracks left or right to new position.
move start
move end
track search
default tracks
default order
hide all
manage custom tracks
track hubs
configure
multi-region
reverse
resize
refresh

# IGV (Integrative Genomics Viewer)

# IGV


IGV_2.3.81
File
Genomes
View
Tracks
Regions
Tools
GenomeSpace
Help
Menu
Human (hg38)
chr17
chr17:34,255,087-34,257,385
Go
Tool bar
p13.2
p13.1
p12
p11.2
p11.1
q11.2
q12
q21.1
q21.31
q21.32
q22
q23.1
q23.3
2 296 bp
34 255 200 bp
34 255 400 bp
34 255 600 bp
34 255 800 bp
34 256 000 bp
34 256 200 bp
34 256 400 bp
34 256 600 bp
34 256 800 bp
Chromosome ideogram
siMitf4_alignment.bam Coverage
siMitf4_alignment.bam Junctions
siMitf4_alignment.bam
Data tracks
Sequence
Gene
CCL2
Annotation tracks

# IGV menu : main features

- File
  - Load files into IGV
  - Manage sessions (e.g. save your current settings to a named session file)
  - Save an image
- Genome
  - Manage genomes available on IGV data server (http://software.broadinstitute.org/software/igv/Genomes)
  - Create new genomes (required : FASTA file, optional : annotation file, …)
- View
  - Preferences : customize the display
- Tools
  - Run igvtools : count (→ tdf), sort, index

# IGV tool bar : main features

Chromosome location being shown
To scroll to a different location,
enter the gene name or locus
e.g. CCL2 or chr17:34,255,087-34,257,385
and click Go

Genome assembly

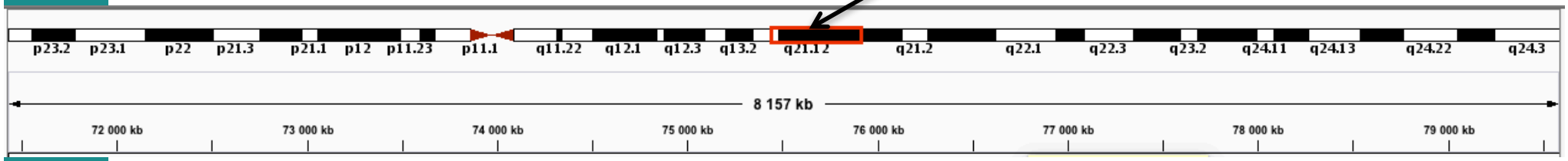
Chromosome


Moves backward and forward through views of the genome like the back and forward buttons in a web browser

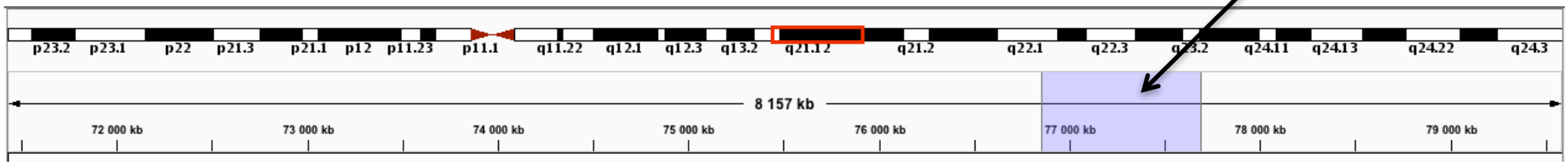
Zooms in and out on a chromosome

# IGV : chromosome ideogram

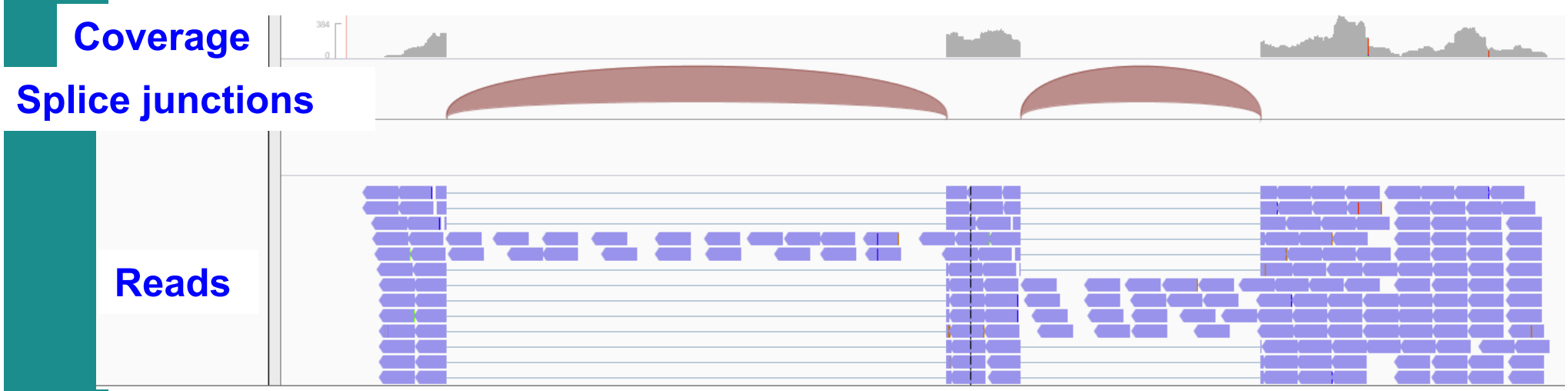
Chromosome location being shown




Click and drag to define
a new region to zoom in

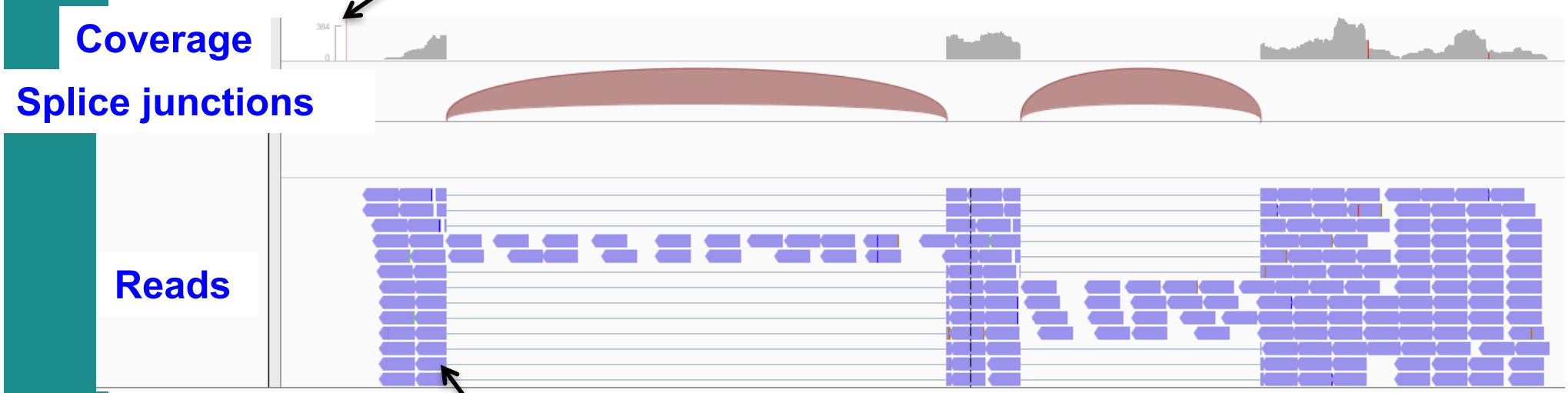


# IGV : Data track

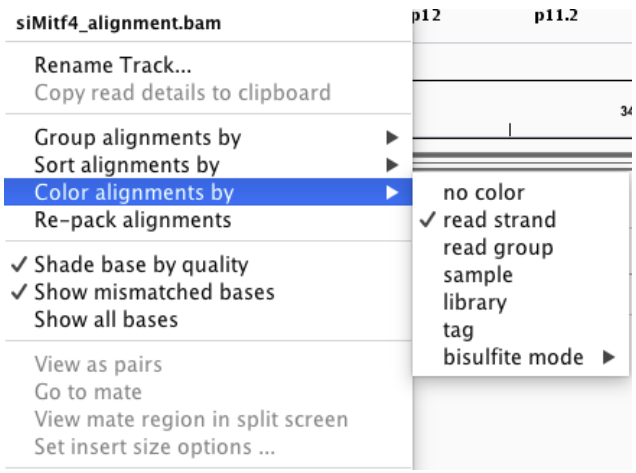


# IGV : Data track

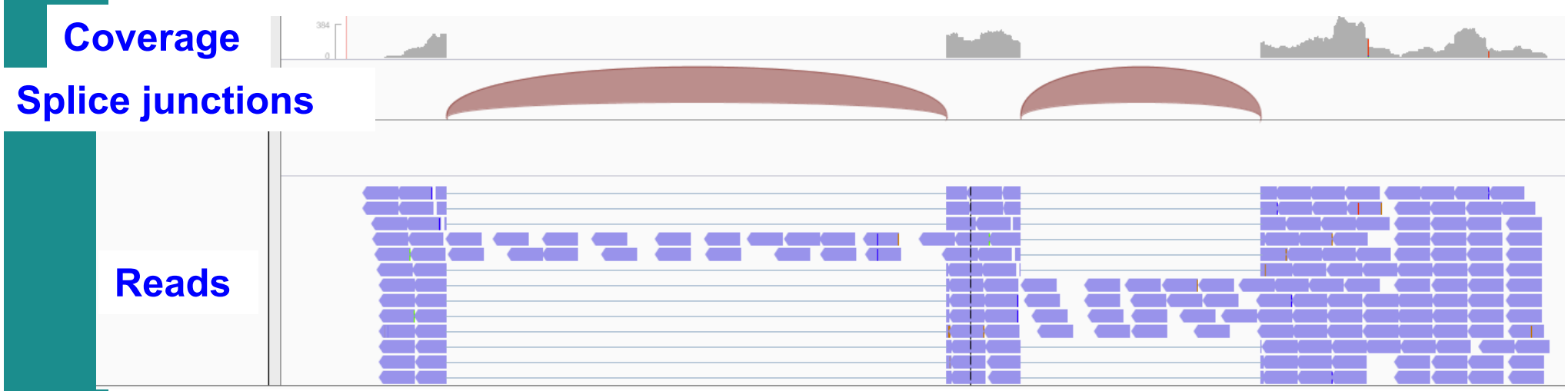
Data range (can be changed by right-clicking on track name)




Read color can be changed by right-clicking on track name

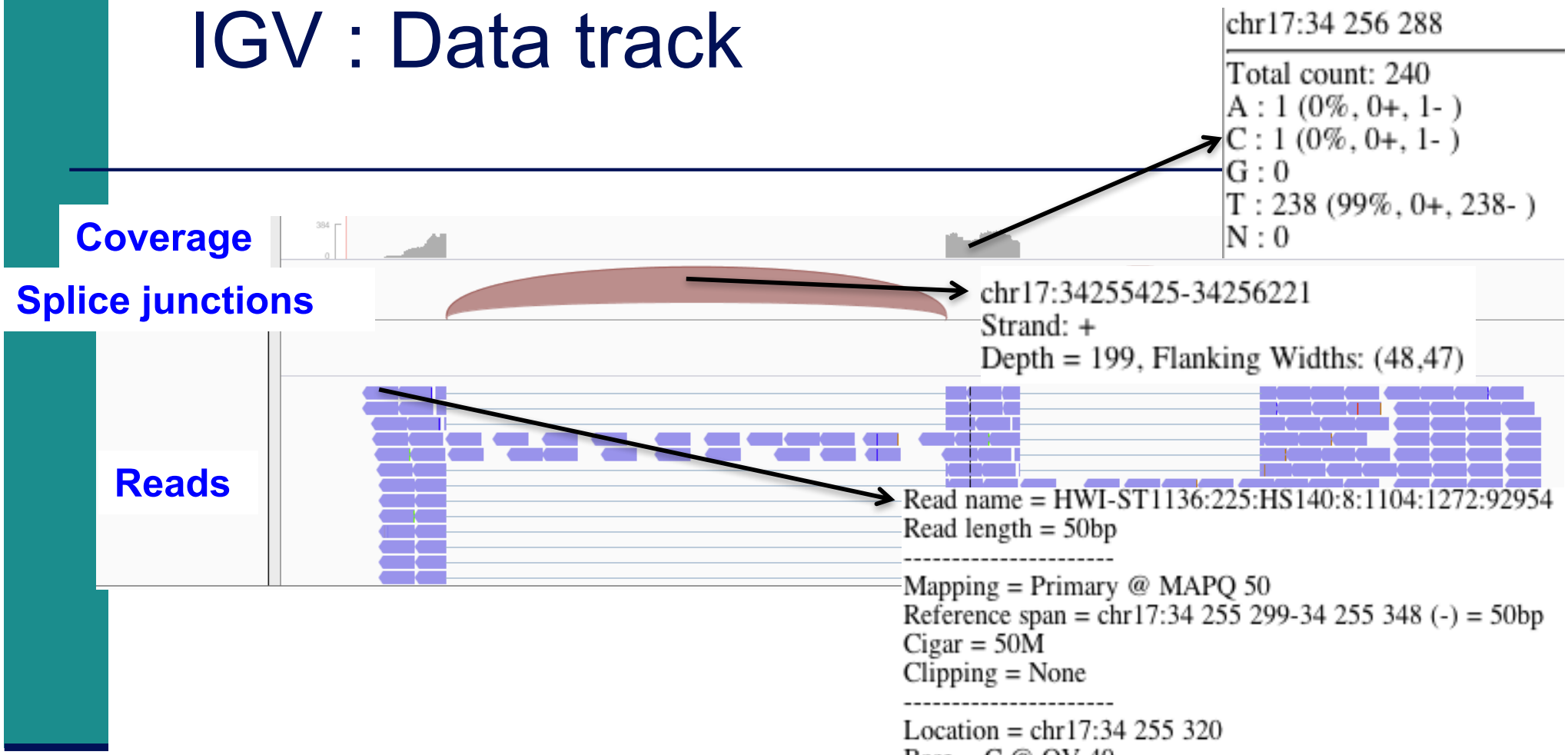


# IGV : Data track



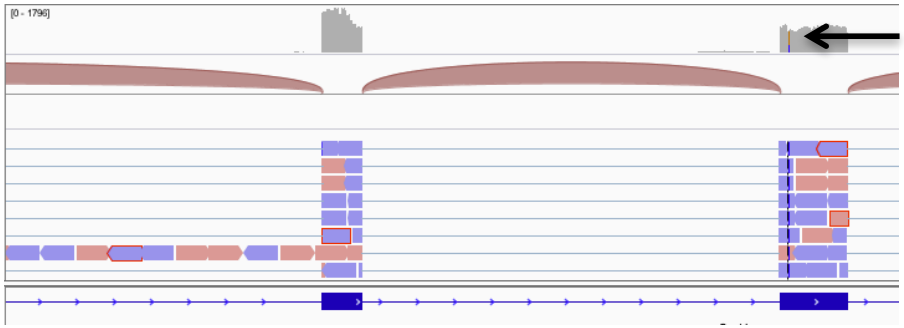

- Display of splice junctions
  - Strand
    - Blue junctions : + strand
    - Red junctions : - strand
  - Depth of coverage
    - The thickness of the arcs are proportional to the depth of coverage
    - All junctions with more than 50 reads have the same thickness

# IGV : Data track




→ When mouse hover on images, pop-up windows provide additional information

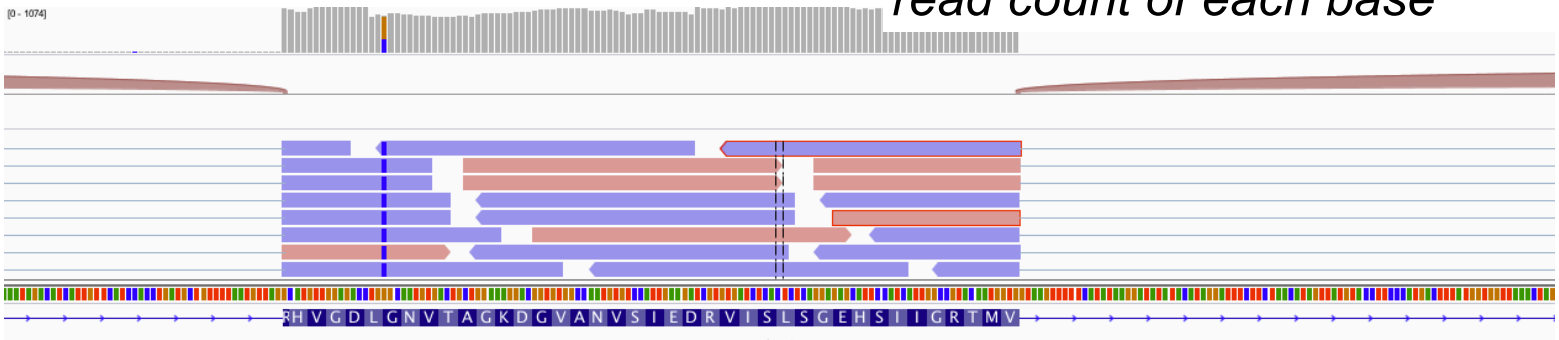
# IGV data track differences vs reference genome



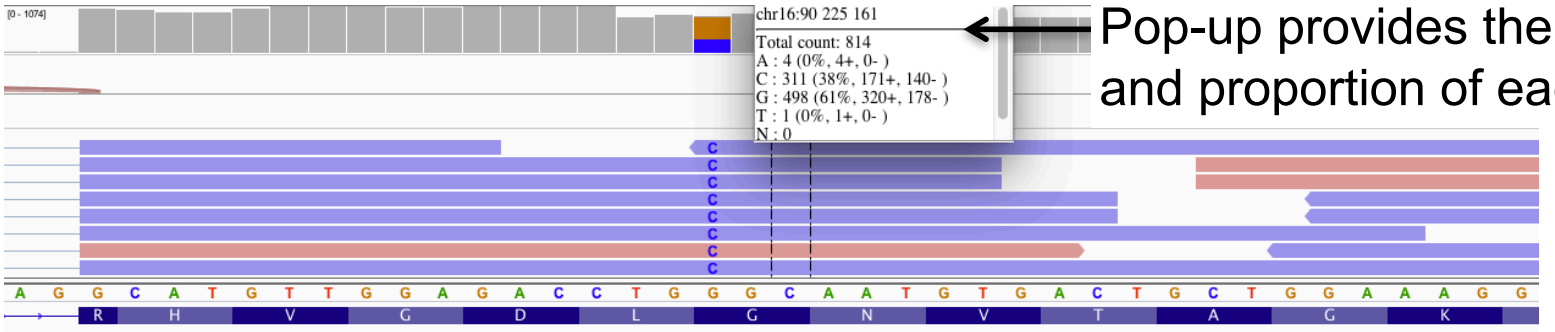
Nucleotide difference compared to the reference genome

*If a nucleotide differs from the reference sequence in greater than 20%* of reads, IGV colors the bar in proportion to the read count of each base*

Zoom in



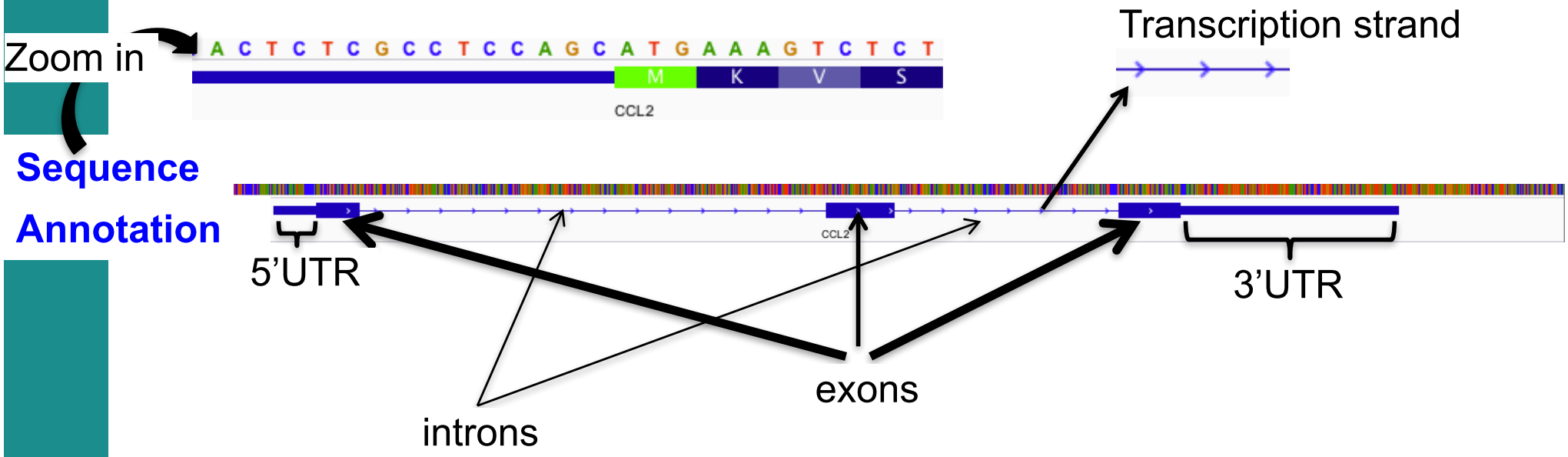
Zoom in




Pop-up provides the number and proportion of each base

* Default threshold, can be changed in
View → Preferences → Alignment → Coverage allele-fraction threshold

# IGV annotation track



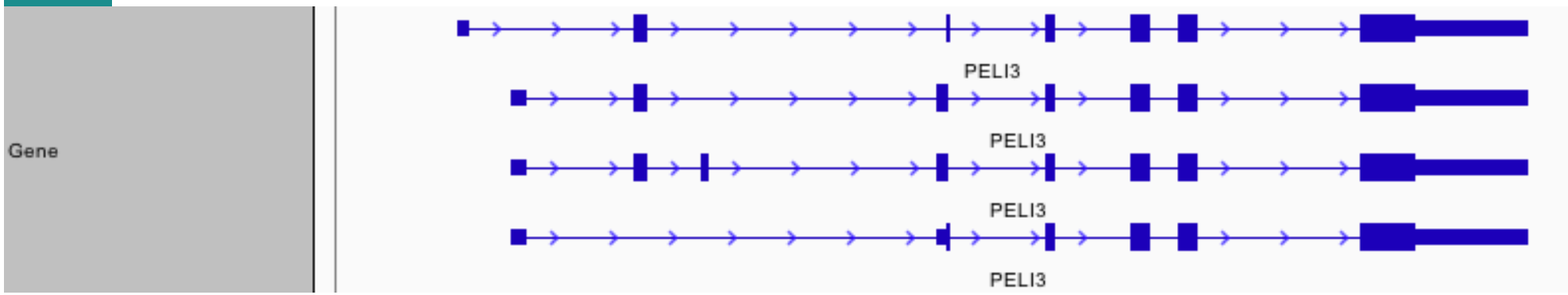

→ When mouse hover on images, pop-up windows provide additional information :

**CCL2**
chr17:34255277-34257201
id = NM_002982
---
Exon number: 2
Amino acid coding number: 51
chr17:34256222-34256339

# IGV annotation track

Default : collapsed


Right click on track name → Expanded
To see all isoforms



# NGS read mapping

- Introduction to NGS read mapping
- Short read mappers
- Specificity of RNA-seq read mapping
- *Exercise 1 : Mapping of RNA-seq data using Galaxy*
- Alignment and related file formats
- Alignment visualization
- *Exercise 1 : Interpretation of results*
- *Exercise 2 : Whole dataset alignment visualization*
- Quality control of RNA-seq data based on alignments

# Exercise 1 : results


Reads alignment

General information on alignment

# Exercise 1 : interpretation of results

1. Align summary
   1.1. How many reads have been mapped onto hg38 ?
   1.2. Among these reads, what is the proportion of multiple mapped reads ?
2. Splice junctions
   2.1. Which splice junctions file format is provided by Tophat2 ?
   2.2. Download this file and visualize these junctions using IGV
   2.3. Look at all splice junctions identified on *Park7* gene. How many reads span the junction between the two last exons of this gene ?
3. Alignment file (accepted_hits)
   3.1. Which alignment file format is provided by Tophat2 ?
   3.2. Download this file and visualize this alignment using IGV
   3.3. Visualize alignments of reads aligned on the junction between the 2 last exons of Park7 gene. Look at the CIGAR string of one of these reads.
   3.4. Verify the strand specificity of the reads, for example on *Pmel* and *Cdk2* genes (color alignments by strand)
   3.5. What do you observe at position chr16:2,771,988 ?

# NGS read mapping

- Introduction to NGS read mapping
- Short read mappers
- Specificity of RNA-seq read mapping
- *Exercise 1 : Mapping of RNA-seq data using Galaxy*
- Alignment and related file formats
- Alignment visualization
- *Exercise 1 : Interpretation of results*
- *Exercise 2 : Whole dataset alignment visualization*
- Quality control of RNA-seq data based on alignments

# Exercise 2 : whole dataset alignments (1/3)

- Tophat2 results for all samples from Mitf project are available on
  - Shared Data → Data Libraries → CNRS training
    - RNAseq → alignment
  - To save time the corresponding BAM, BAI and tdf files are already available on your computer

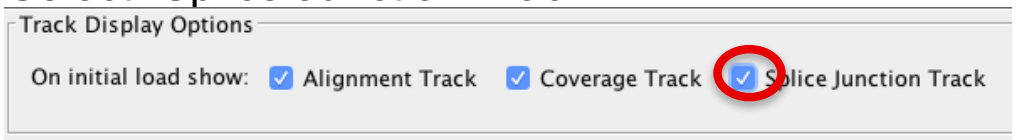
1. What is the proportion of mapped reads in all samples ?
2. Before visualizing these alignments using IGV :
   Use File → new session to start a new IGV session
   Verify in View → Preferences → Tracks tab that “Normalize coverage data” is selected
   Load the 4 tdf files on IGV

   A ChIP-seq peak has previously been identified near *Idh1* gene.
   Is this gene differentially expressed between siLuc and siMitf samples ?
3. Load the 4 BAM files on IGV.

   In the last exon of *Idh1* gene, do you identify a nucleotide difference in RNA-seq samples compared to the reference genome ? What is the position of this difference ?

# Exercise 2 : whole dataset alignments (2/3)

4. Look at splice junctions identified on *Acp5* gene
   - In View →Preferences → Alignments
     - Select “Splice Junction Track”




     - Filter to view only junctions with a minimal number of flanking bases and a minimum junction coverage


   - File → new session to and reload the 4 BAM files to apply these filters
   - To see all annotated isoforms, right click on annotation track and select Expanded
   - Are all these junctions annotated in Refseq ? and in Ensembl ?
     Ensembl release 85 annotations are available on your computer : RNAseq/annotations/Homo_sapiens.GRCh38.85.sorted.gtf
     → Load this file on IGV in order to visualize Ensembl annotations
   - 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 : Ensembl annotations → Select Alignment Tracks : all alignments

# Exercise 2 : whole dataset alignments (3/3)

5. The same RNA samples have been processed with a different RNA-seq protocol. The corresponding alignment file for siLuc2 sample is available on your computer : RNAseq/other_protocol/siLuc2_other_protocol_alignment.bam

   - What do you think about this protocol ? Look for example at *Idh1* and *Idh-as1* genes.

# NGS read mapping

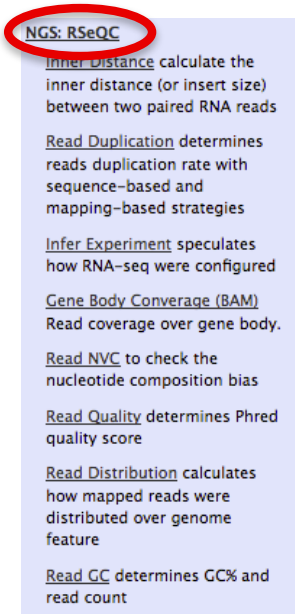
- Introduction to NGS read mapping
- Short read mappers
- Specificity of RNA-seq read mapping
- *Exercise 1 : Mapping of RNA-seq data using Galaxy*
- Alignment and related file formats
- Alignment visualization
- *Exercise 1 : Interpretation of results*
- *Exercise 2 : Whole dataset alignment visualization*
- Quality control of RNA-seq data based on alignments

# Quality control of RNA-seq data based on alignments

- Proportion of mapped, uniquely and multiple mapped reads in all samples within a project
- For paired-end sequencing : distance between reads
- For directional protocol : strand information
- Read coverage over genes
- Read distribution relative to known annotations

http://rseqc.sourceforge.net/

RSeQC

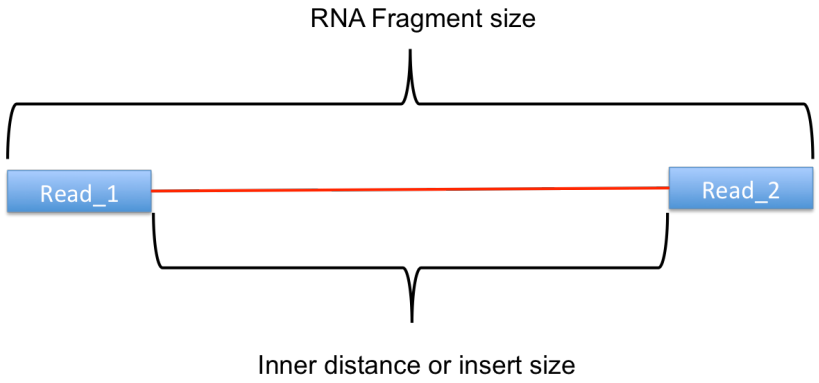



RSeQC available on GalaxEast

RSeQC input :
alignment (BAM/SAM) and annotation (BED) files

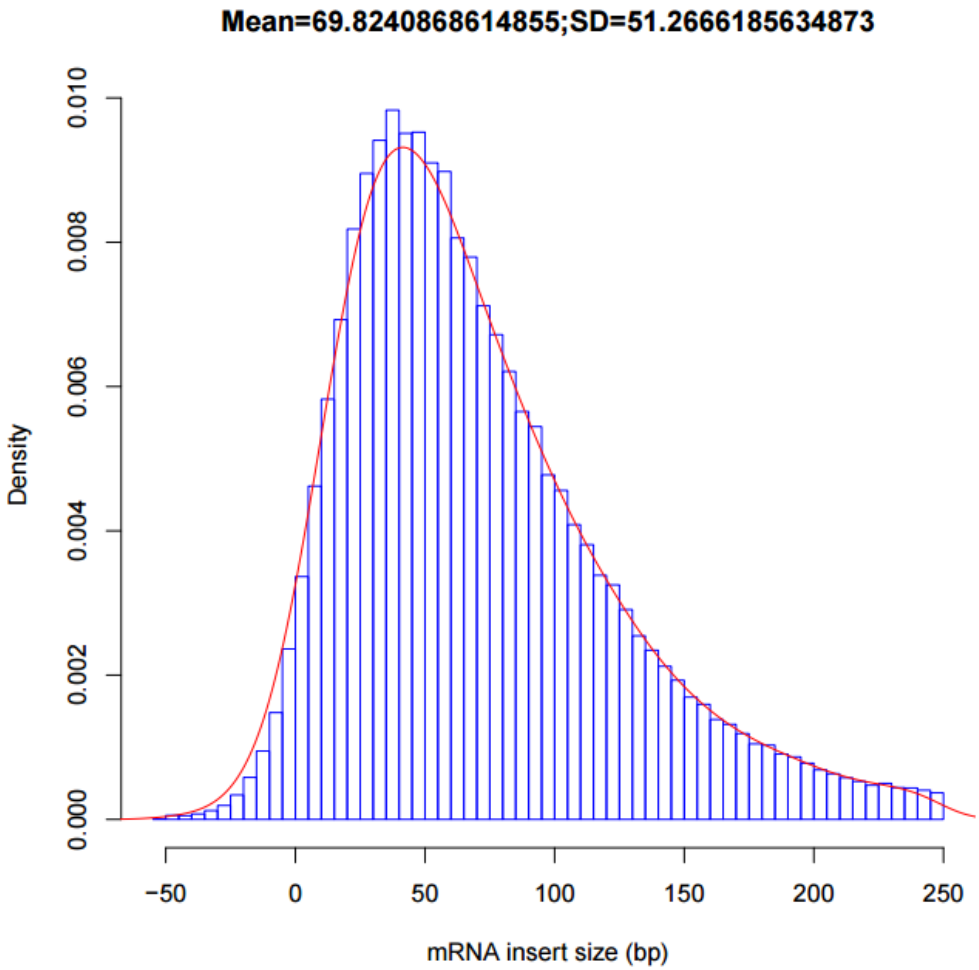
# Distance between reads (paired-end sequencing)

- To know inner distance (insert size) between paired reads
  - The distance is the mRNA length between two paired fragments




- RSeQC Inner Distance
  - Determines the genomic (DNA) size between two paired reads: D_size = read2_start - read1_end
    - if 2 paired reads map to the same exon or a non-exonic region
      - inner_distance = D_size
    - if 2 paired reads map to different exons
      - inner_distance = D_size - intron_size
  - The inner_distance might be a negative value if 2 fragments overlapped

# RSeQC inner distance : example of result



# Strand information (directional protocol)

- To infer how reads were stranded for strand-specific RNA-seq data
  - Compare the “strandness of reads” with the “standness of transcripts”
  - The “strandness of reads” is determined from alignment
  - The “standness of transcripts” is determined from annotation

- RSeQC infer experiment
  - Calculates the proportion of reads corresponding to :
    - ++,- -
    - +-,-+

| | Annotated gene on + strand | Annotated gene on - strand |
|---|---|---|
| Read mapped to + strand | ++ | +- |
| Read mapped to - strand | -+ | -- |

# RSeQC infer experiment : examples of result

Result on siLuc2 (directional protocol)

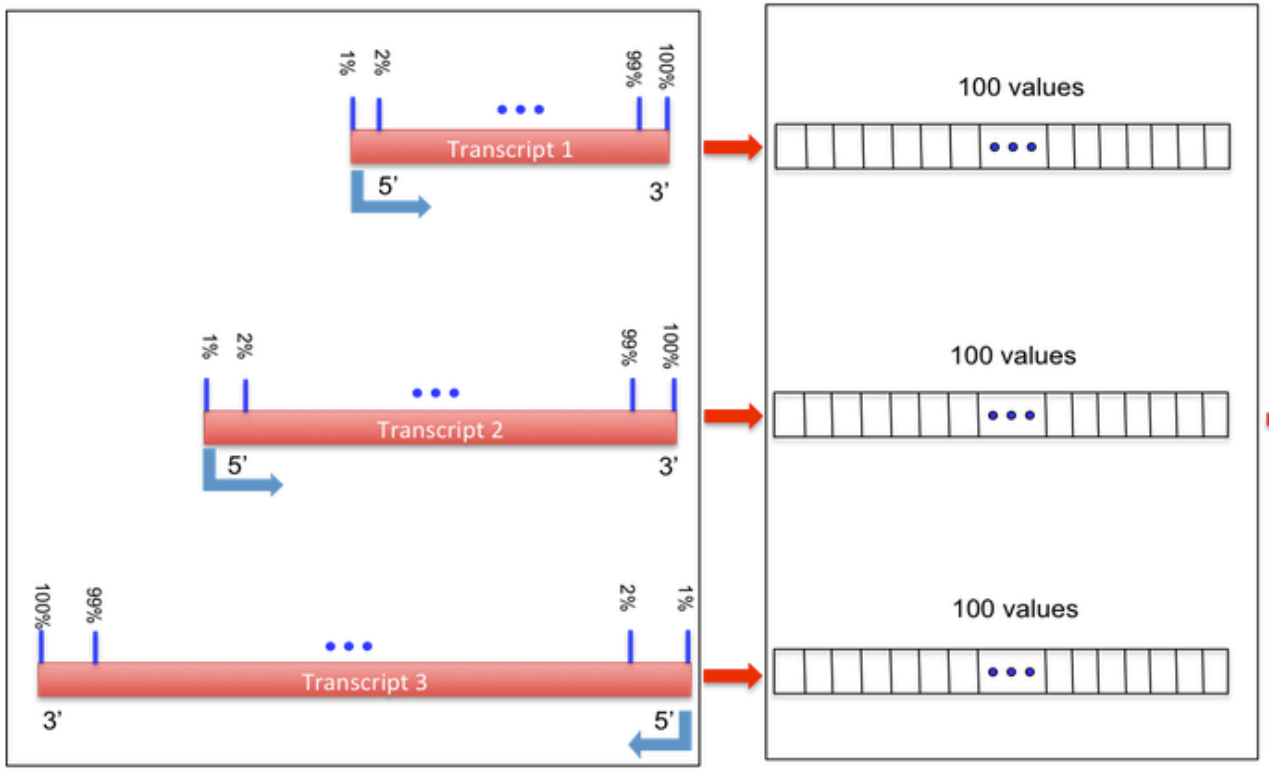
```
This is SingleEnd Data
Fraction of reads explained by "++,--": 0.0090
Fraction of reads explained by "+-,-+": 0.9910
Fraction of reads explained by other combinations: 0.0000
```

Result on siLuc2 (standard protocol)

```
This is SingleEnd Data
Fraction of reads explained by "++,--": 0.4984
Fraction of reads explained by "+-,-+": 0.5016
Fraction of reads explained by other combinations: 0.0000
```

# Read coverage over genes

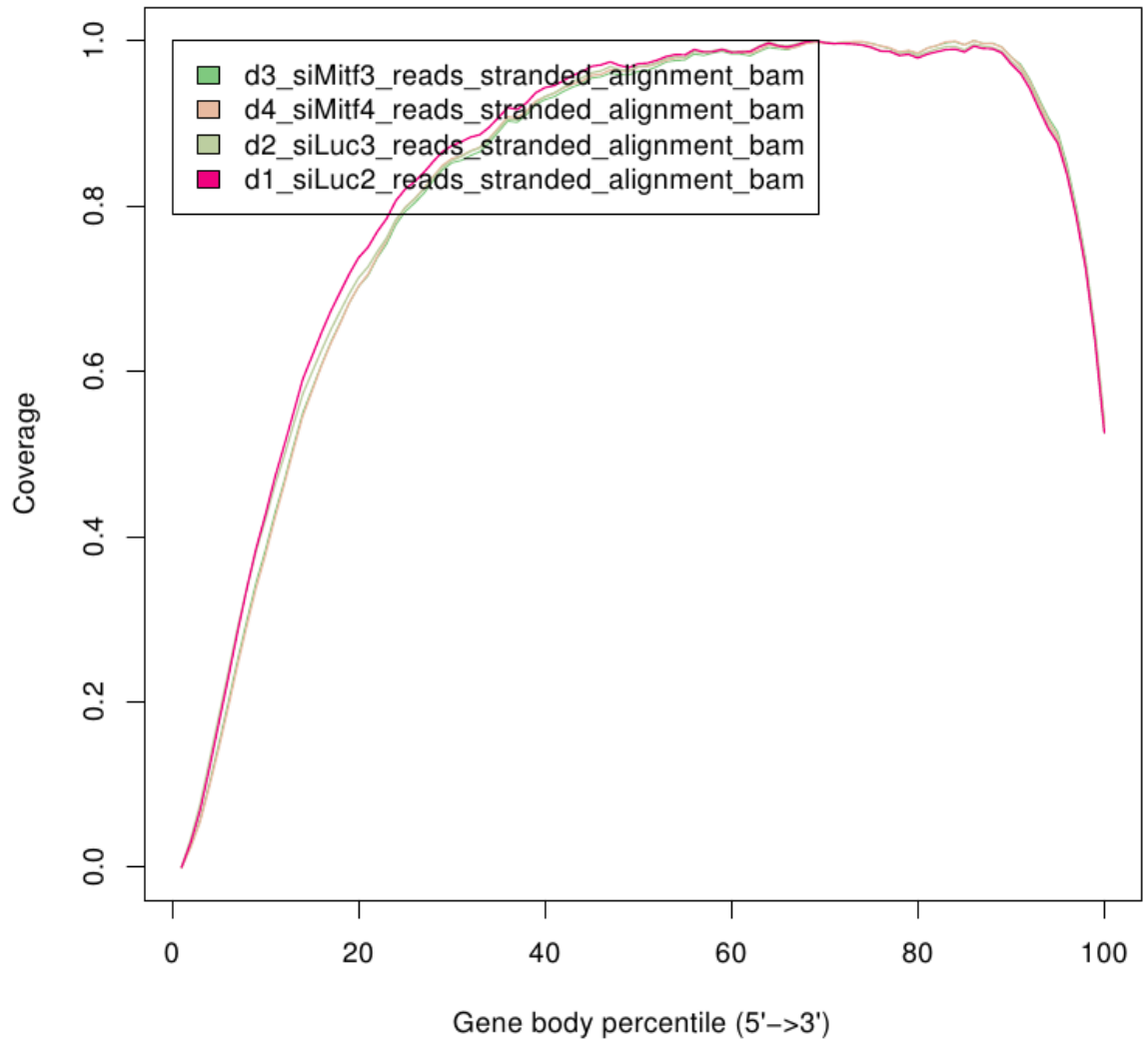
- To identify any bias in read coverage over genes
- RSeQC Gene Body Coverage



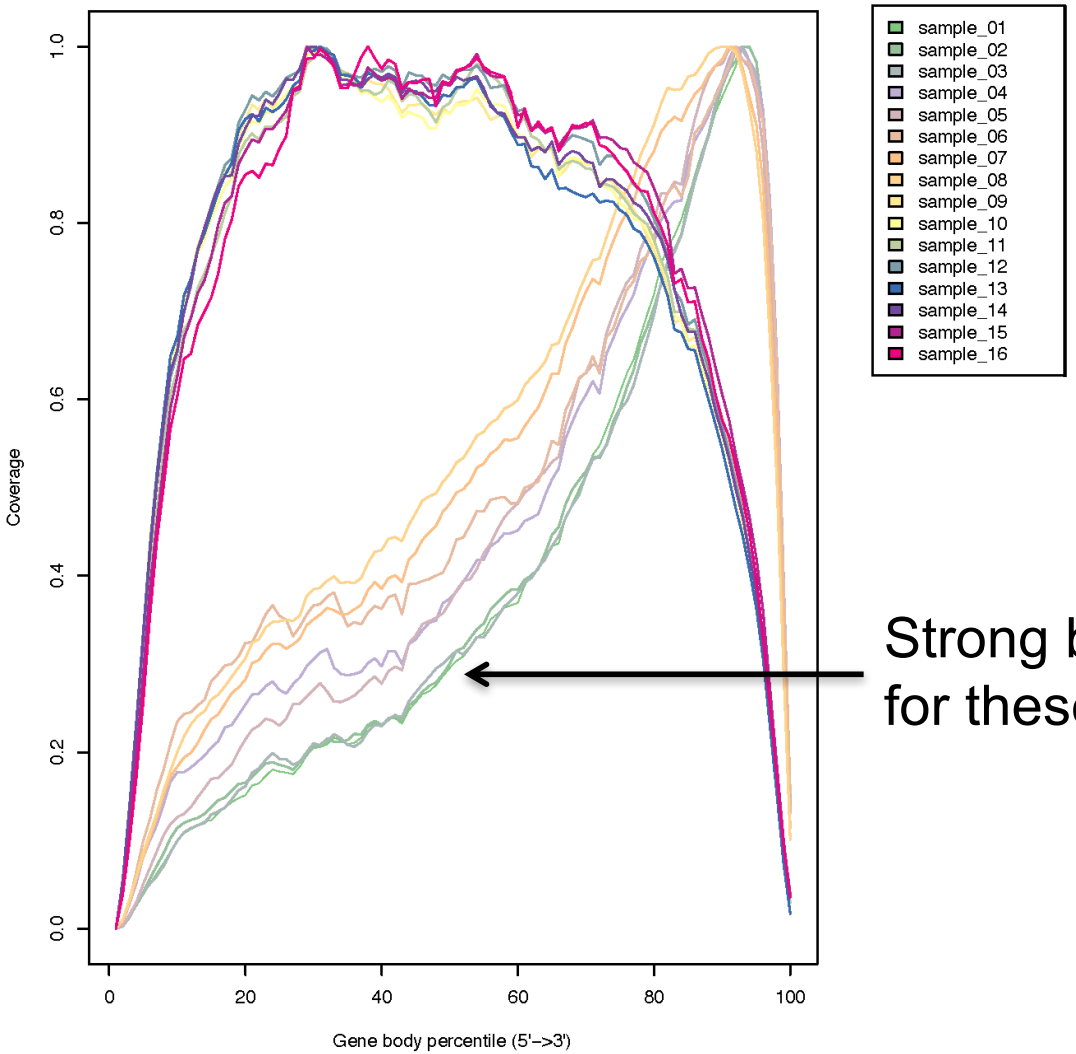

Take 100 quantiles from each transcripts in BED file

Extract coverage signals from BAM file

# Read coverage over genes : result



# Read coverage over genes : example with biased samples




Strong bias in read coverage for these samples

# Read distribution relative to known annotations

- How mapped reads are distributed over genomic features (CDS, UTR, intron, intergenic regions)

- RSeQC read distribution
  - Assigns mapped reads to a genomic feature
  - When genomic features overlap, they are prioritized as:
    - CDS > UTR > Introns > Intergenic regions
  - Does not assign reads located beyond TSS upstream 10Kb or TES downstream 10Kb

CDS : Coding DNA Sequence
UTR : UnTranslated Region
TSS : Transcription Start Site
TES : Transcription End Site

# Read distribution relative to known annotations : results on siLuc2

| | |
|---|---|
| Total Reads | 42797297 |
| Total Tags* | 48536773 |
| Total Assigned Tags° | 47567800 |

| Group | Total_bases | Tag_count | Tags/Kb |
|---|---|---|---|
| CDS_Exons | 92736826 | 36167119 | 390.00 |
| 5'UTR_Exons | 6812435 | 402686 | 59.11 |
| 3'UTR_Exons | 30815395 | 7355000 | 238.68 |
| Introns | 1469504677 | 3175039 | 2.16 |
| TSS_up_1kb | 29748818 | 42485 | 1.43 |
| TSS_up_5kb | 133216562 | 92407 | 0.69 |
| TSS_up_10kb | 238672534 | 132661 | 0.56 |
| TES_down_1kb | 31662314 | 173381 | 5.48 |
| TES_down_5kb | 137527800 | 279648 | 2.03 |
| TES_down_10kb | 242337608 | 335295 | 1.38 |

* reads spliced once are counted as 2 tags, reads spliced twice are counted as 3 tags, …

° number of tags that can be assigned to the 10 above groups

Tags assigned to “TSS_up_1kb” are also assigned to “TSS_up_5kb” and “TSS_up_10kb”

Tags assigned to “TSS_up_5kb” are also assigned to “TSS_up_10kb”