NGS read mapping

Céline Keime
keime@igbmc.fr

NGS read mapping

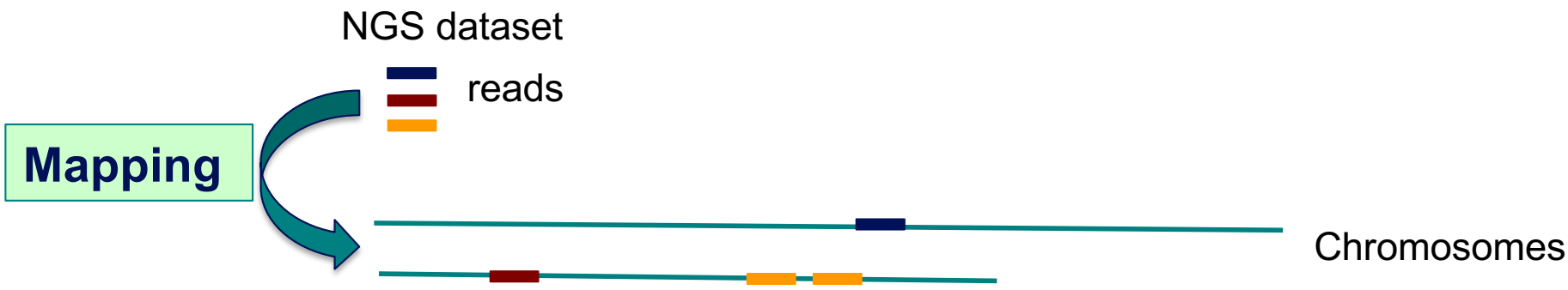
- Introduction to NGS read mapping
- Short read mappers
- Specificity of RNA-seq read mapping
- *Exercise 1 : Mapping of RNA-seq data using Galaxy*
- Alignment and related file formats
- Alignment visualization
- *Exercise 1 : Interpretation of results*
- *Exercise 2 : Whole dataset alignment visualization*
- Quality control of RNA-seq data based on alignments
  - *Exercise 3 : QC 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
  - *Exercise 3 : QC 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
  - *Exercise 3 : QC on alignments*

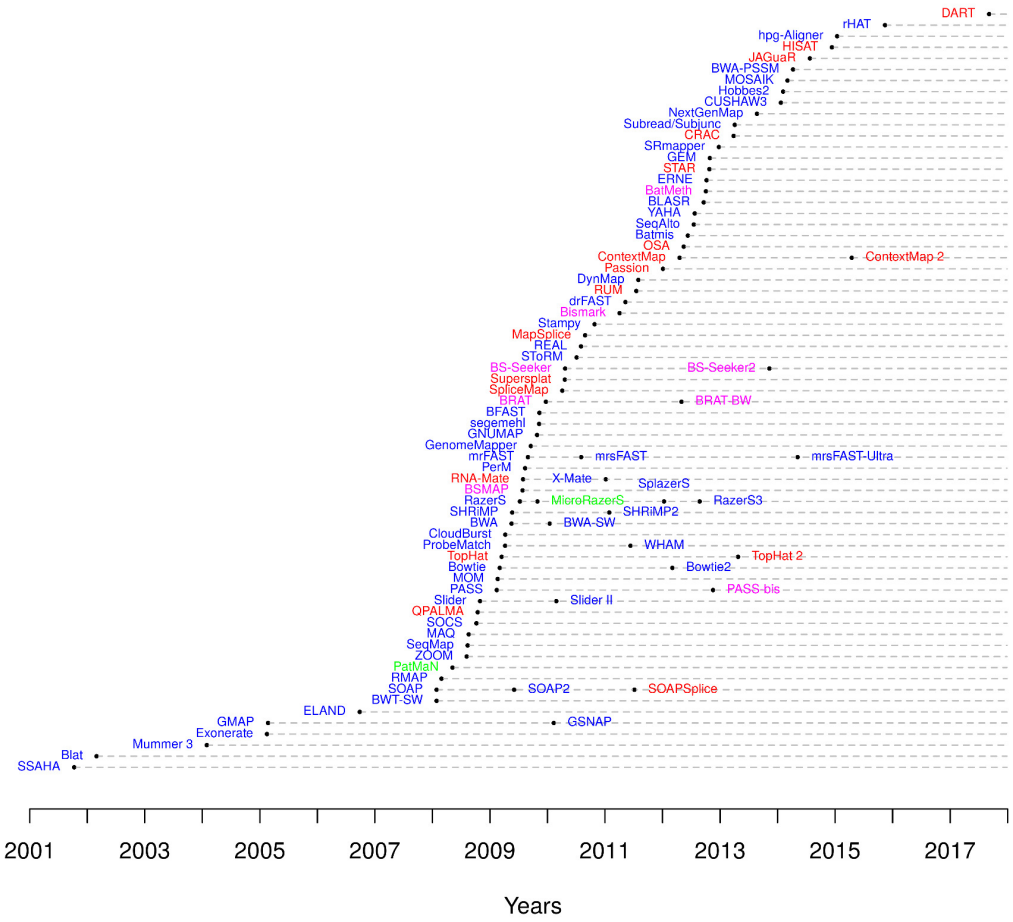
Computational strategies

- 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
- Transforming
  - Uses a technique originally developed for compressing large files called the Burrows-Wheeler transform (BWT)
    ➔ The transformed human genome fits into memory
- Example : Bowtie2 *(Langmead et al. Nature Methods 2012)*
  - To rapidly narrow the number of possible alignments that must be considered
    - Begins by extracting substrings ("seeds") from each read and its reverse complement
    - Aligning them in an ungapped fashion using an index

    → Trade-off between speed and sensitivity can be adjusted by setting the seed length, the interval between extracted seeds and the number of mismatches in seed
  - Extend seeds to full reads alignment (allowing gaps)

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/

How to choose a mapper ?

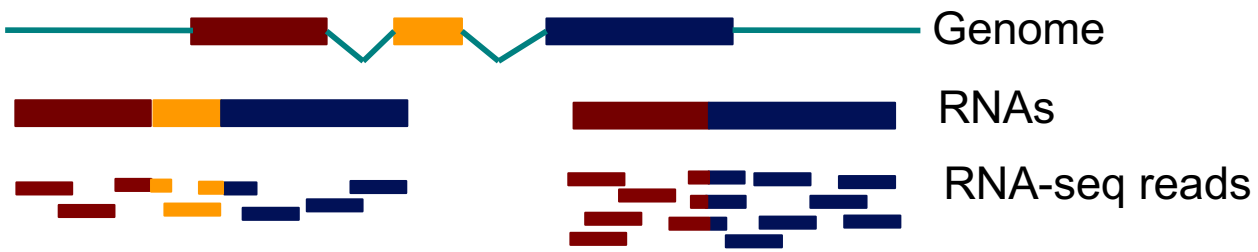
- Main criteria to take into account
  - Sensitivity
    - Ability to align a large fraction of reads with errors and variants
  - Accuracy
    - If an aligner aligns a large fraction of reads, but most alignments are wrong, this is useless !
  - Type of data (DNA, RNA), support of paired-end
  - Read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

- Feature comparison
  - 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
  - *Exercise 3 : QC on alignments*

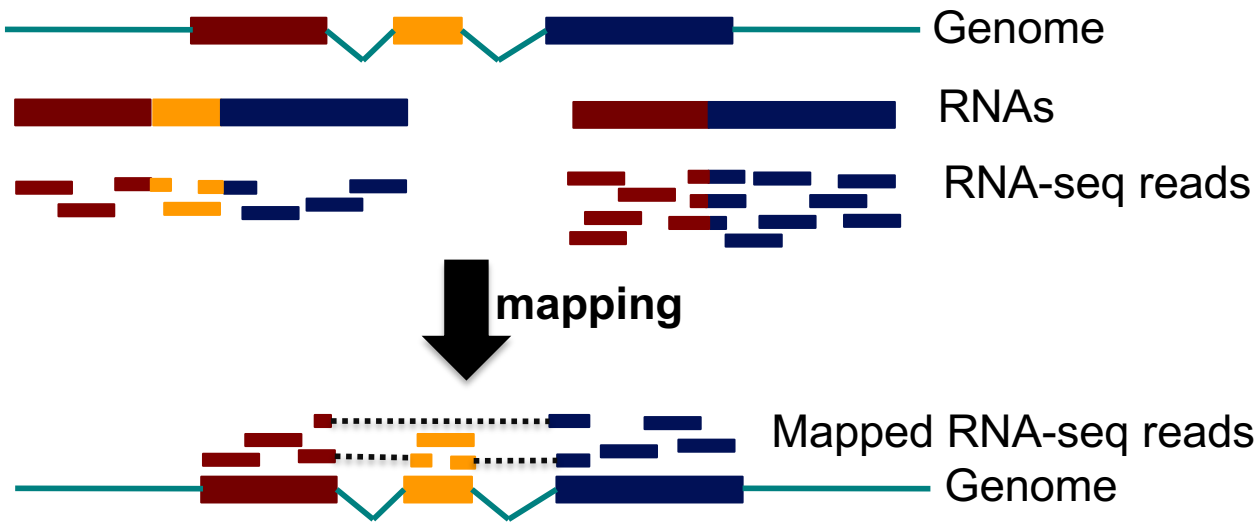
Specificity of RNA-seq reads



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

Spliced mapping

- Allows mapping of reads across splice junctions



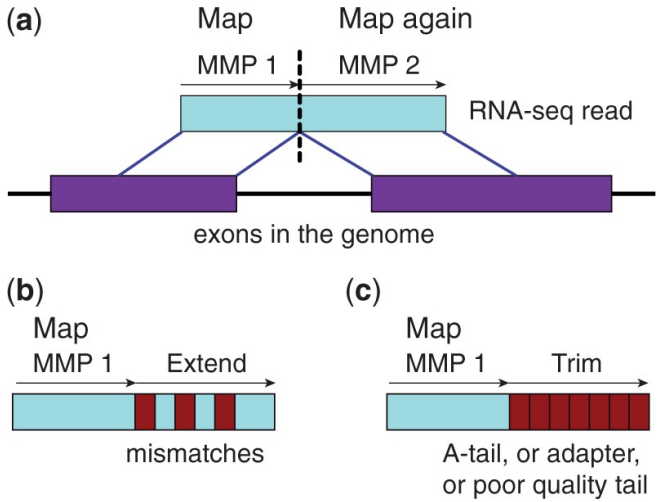
- Spliced alignment programs comparison
  - Engström et al. Nature Methods 2013
  - Baruzzo et al. Nature methods 2017

STAR
Spliced Transcripts Alignment to a Reference

1. Searching for seeds

- For every read : searches for the longest sequence that exactly matches one or more locations on the reference genome : Maximal Mappable Prefix (MMP) → MMP1 (seed 1)
- Searches for only the unmapped portion of the read to find the next longest sequence that exactly matches the reference genome → MMP2
- MMP search enables finding mismatches or tails :

  If MMP search does not reach the end of a read
  → MMPs serve as anchors in the genome that can be extended
  → If the extended alignment is not good : tail is soft-clipped



2. Stitching all seeds

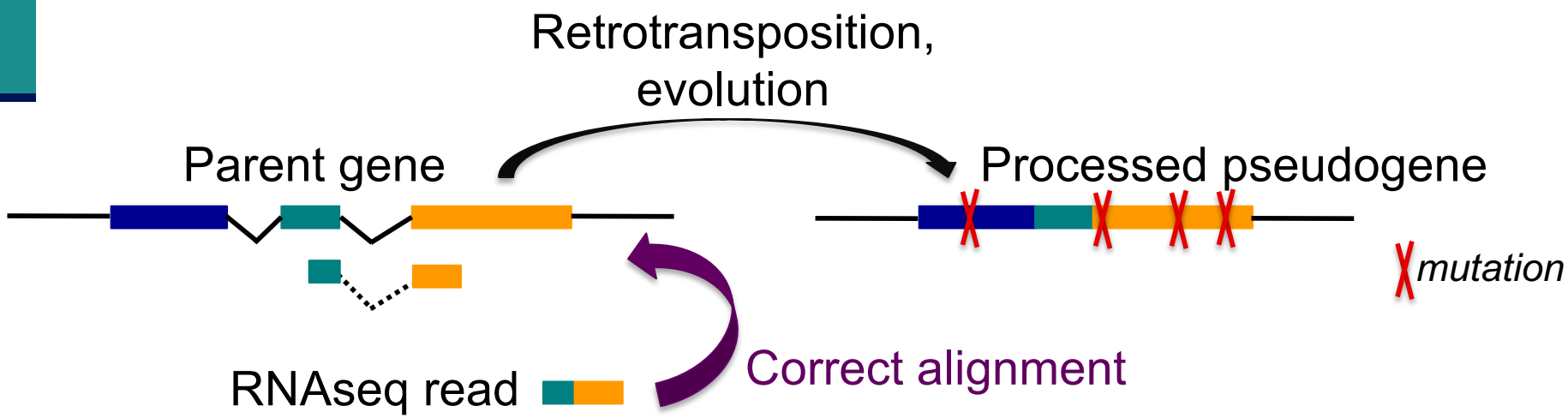
→ alignment of the entire read sequence

Dobin et al. Bioinformatics 2013

Main problems of RNA-seq aligners

- Difficult to accurately detect splicing events involving short sequence overhangs on the donor or acceptor side of a junction

- Alignments biased toward processed pseudogenes



Use of annotations in spliced mapping

- Use splice junctions annotations to mitigate this problem
- STAR
  - Option to provide annotations
  - Incorporates annotated junction sequences into the suffix array
  - Searches the seeds that cross the junctions simultaneously with the seeds that map contiguously to the genome

Genome annotations

- Ensembl project (www.ensembl.org)
  - Goal : automatically annotate genomes, integrate this annotation with other available biological data and make all this publicly available
  - Includes manual curation (by HAVANA) for some species : human, mouse, zebrafish, rat
  - 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, access to old versions via View in archive site
- The main Ensembl site focuses on vertebrate genomes and some other representative species (http://www.ensembl.org/info/about/species.html), other sites are dedicated to plants, fungi, bacteria (cf "Our sister sites" links at the bottom of www.ensembl.org)
- 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)

Genome annotations

- Generally provided in a GTF (Gene Transfert Format) / GFF (General Feature Format) file
- GTF file :
  - Tab-delimited text file format
  - Each line correspond to an annotation or feature
  - Specifications :
    - http://www.ensembl.org/info/website/upload/gff.html
  - e.g. human Ensembl 105 GTF file
    - ftp.ensembl.org/pub/release-105/gtf/homo_sapiens/Homo_sapiens.GRCh38.105.chr.gtf.gz
  - Caution : use annotations corresponding to the version of genome assembly you are working on
    - **GRCh38 (1 - 22, X, Y, MT)** / hg38 (chr1 - chr22, chrX, chrY, chrM)

Genome annotations

- Generally provided in a GTF (Gene Transfert Format) file
  - Nine columns :

Seqid	Source	Type	Start	End	Score	Strand	Phase	Attributes
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"; …

NGS read mapping

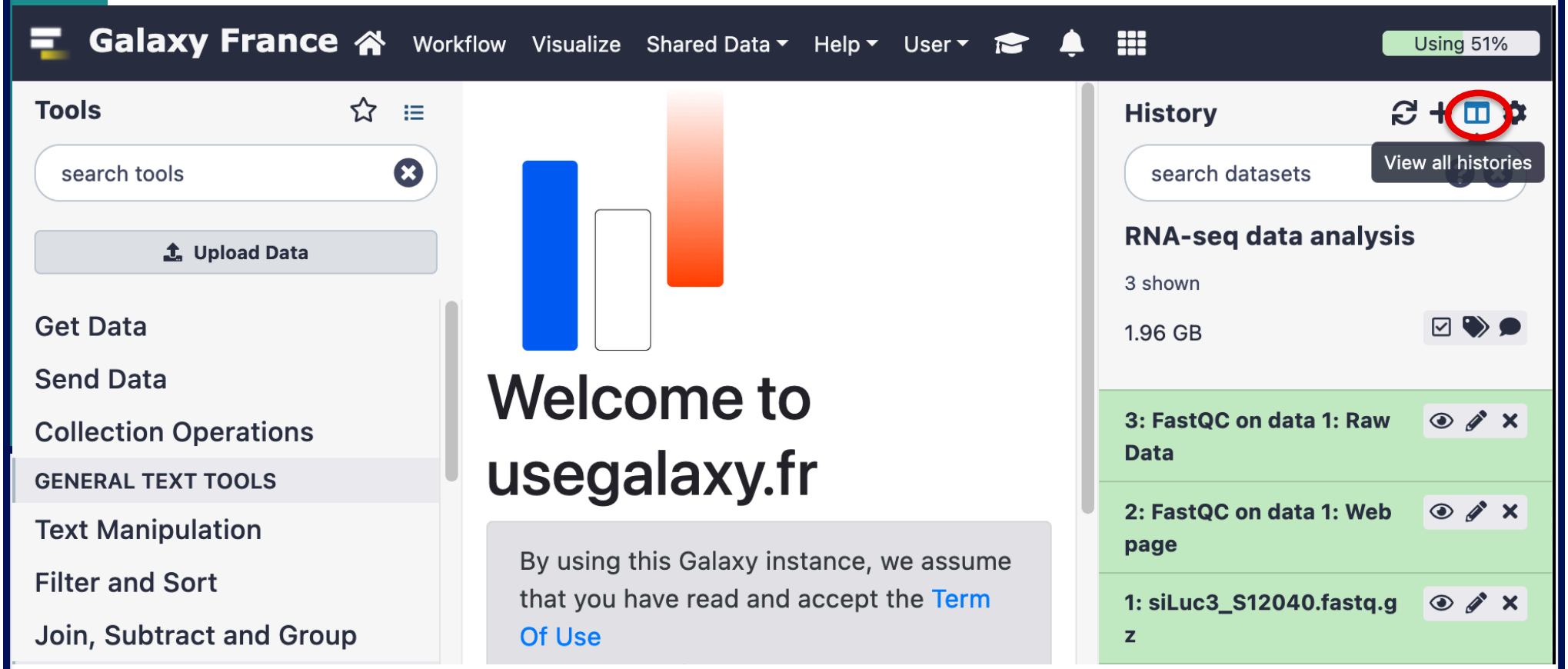
- Introduction to NGS read mapping
- Short read mappers
- Specificity of RNA-seq read mapping
- *Exercise 1 : Mapping of RNA-seq data using Galaxy*
- Alignment and related file formats
- Alignment visualization
- *Exercise 1 : Interpretation of results*
- *Exercise 2 : Whole dataset alignment visualization*
- Quality control of RNA-seq data based on alignments
  - *Exercise 3 : QC on alignments*

Exercise 1
Mapping of RNA-seq data using Galaxy

- Map **1 million** reads from siLuc2 mRNA-seq sample using STAR
  1. Copy to your history
     - The corresponding FASTQ file :
       2: siLuc2_1000000.fastq.gz
     - The GTF annotation file :
       3: Homo_sapiens.GRCh38.105.chr.gtf.gz
  2. Launch STAR on this FASTQ file using
     - GRCh38 reference genome
     - Homo_sapiens.GRCh38.105.chr.gtf.gz GTF annotation file

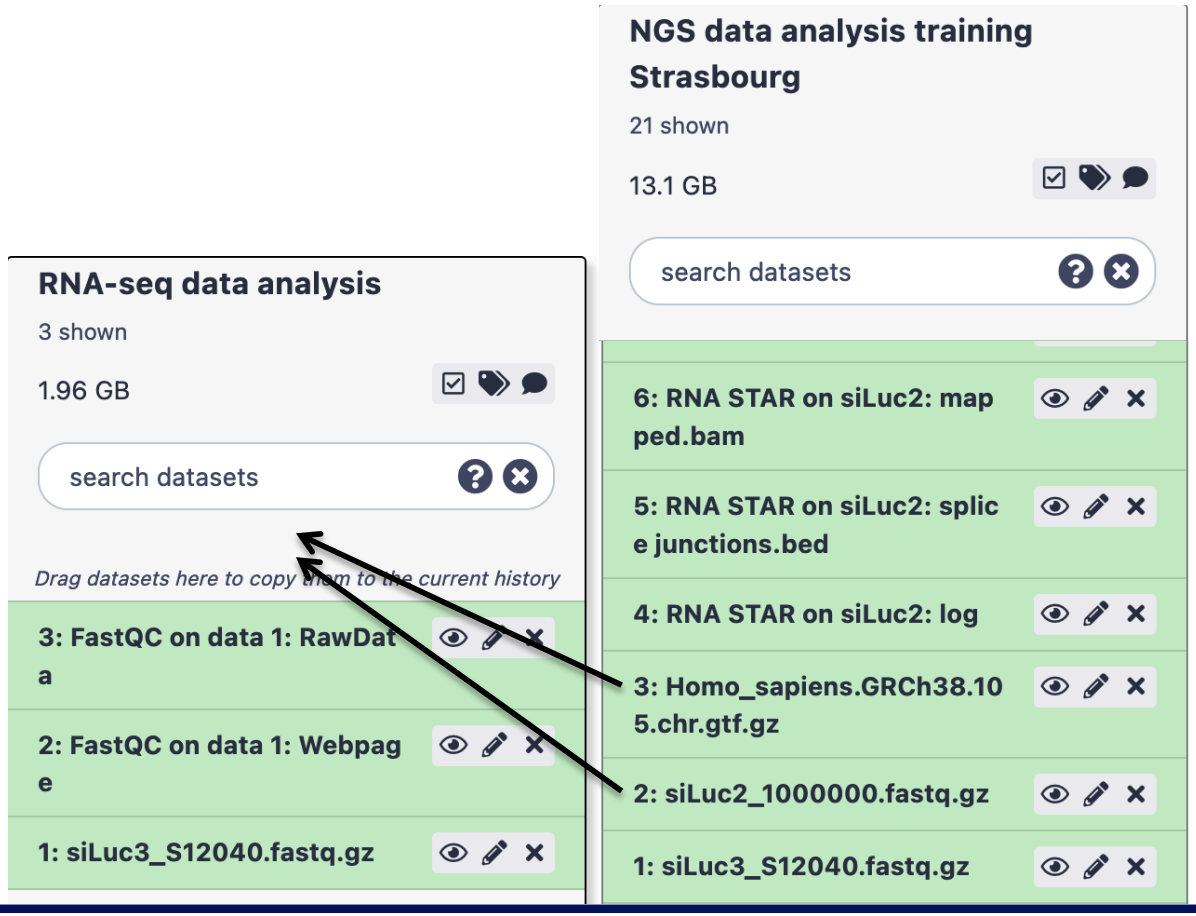
Exercise 1
1. Copy files to your history

- Click on “View all histories”



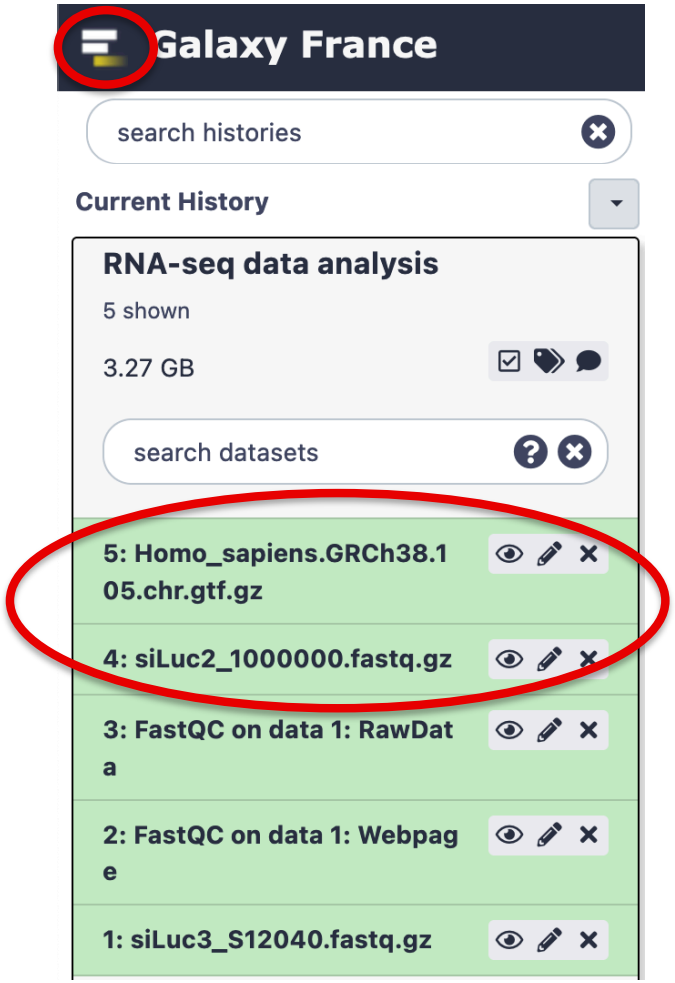
Exercise 1
1. Copy files to your history

- Drag datasets 2 and 3
  - from “NGS data analysis training Strasbourg” history
  - to “RNA-seq data analysis” history

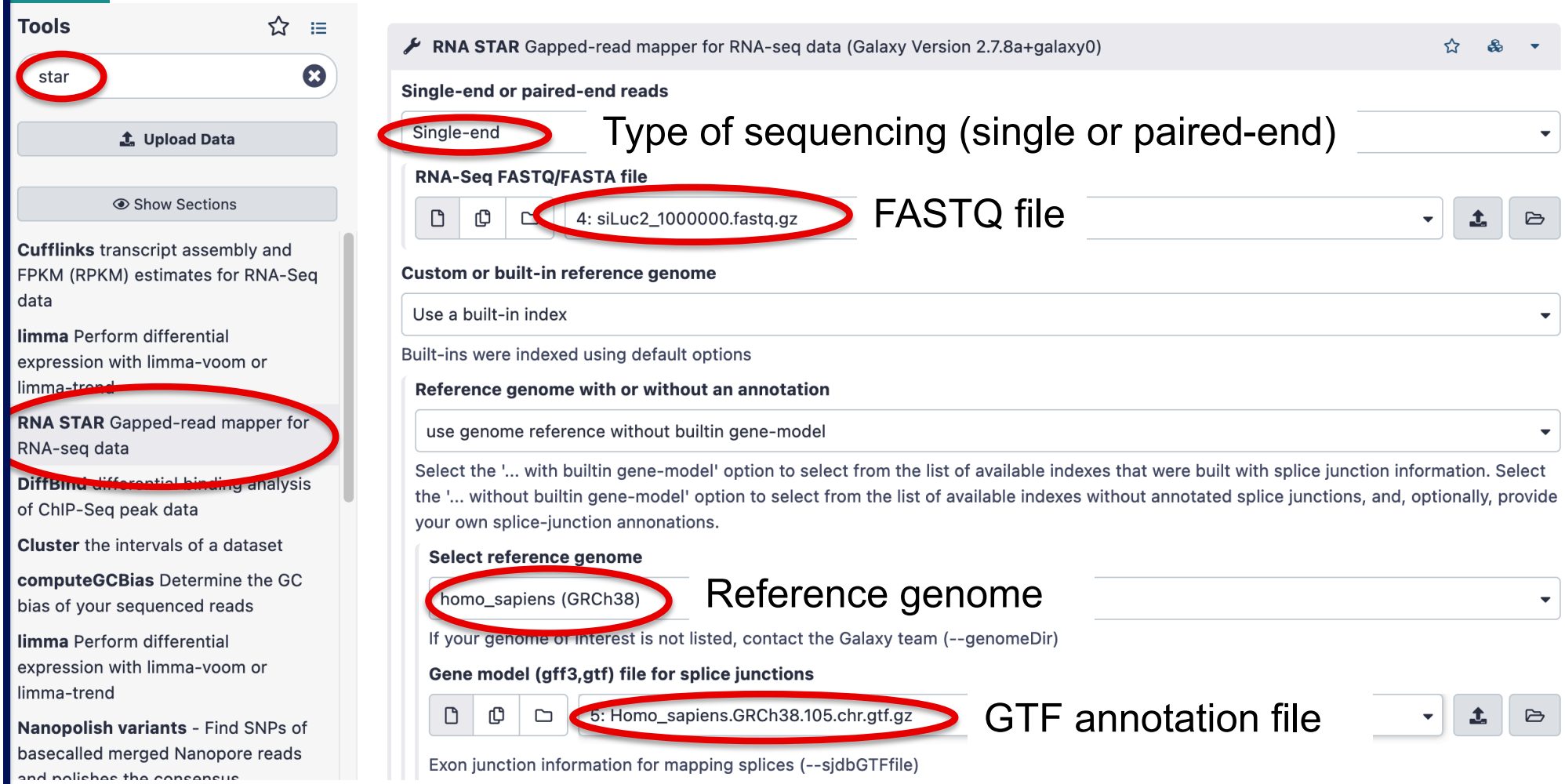


Exercise 1
1. Copy files to your history

- You have now in your history all files needed to launch STAR :



Exercise 1
2. Launch STAR



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 3 : QC 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:1.4 SO:coordinate
@SQ SN:1 LN:248956422
@SQ SN:10 LN:133797422
@SQ SN:11 LN:135086622
@SQ SN:12 LN:133275309
…
```

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-ST1136:225:HS140:8:1313:4696:85038  256 1   14406   0   1S49M   *   0   0   GTCTGCTCAGTTCTTTATTGATTGGTGTGCCGTTTTCTCTGGAAGCCTCT
CCCFFFFFHHHHHJJJJJJJIJJJJHIHIIJJIJJJIJJJJJIJJJJJJJ  NH:i:8 HI:i:5 AS:i:48 nM:i:0
HWI-ST1136:225:HS140:8:2308:6999:65454  0   1   14453   1   50M *   0   0   CTTAAGAACACAGTGGCGCAGGCTGGGTGGAGCCGTCCCCCCATGGAGCA
@@@DDBFFGHHDFADFABBGGAHGDBB??00?DDB?FHHIIE=BD;;@>?  NH:i:3 HI:i:1 AS:i:49 nM:i:0
HWI-ST1136:225:HS140:8:2109:9064:22156  0   1   14467   0   50M *   0   0   GGCGCAGGCTGGGTGGAGCCGTCCCCCCATGGAGCACAGGCAGACAGAAG
```

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)

  Indicates whether the read is correctly mapped to this location in the reference genome

  - STAR mapping quality
    - 60 by default on Galaxy for uniquely mapped reads
    - $\text{int}(-10*\log_{10}(1-1/N_{map}))$ for multi-mapping reads
      N_{map} : the number of loci a read maps to

N_{map}	MAPQ
2	3
3-4	1
≥ 5	0

Alignment file format : SAM

- CIGAR (string)
  - M : alignment (can be a sequence match or mismatch)
  - I : insertion to the reference
  - D : deletion from the reference
  - N : skipped region from the reference
  - S : soft clipping (clipped sequences present in SEQ)
    - Bases of the read that are not aligned
  - H : hard clipping (clipped sequences not present in SEQ)
    - Bases of the read that are not aligned and that have been removed from the read sequence in the SAM file

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

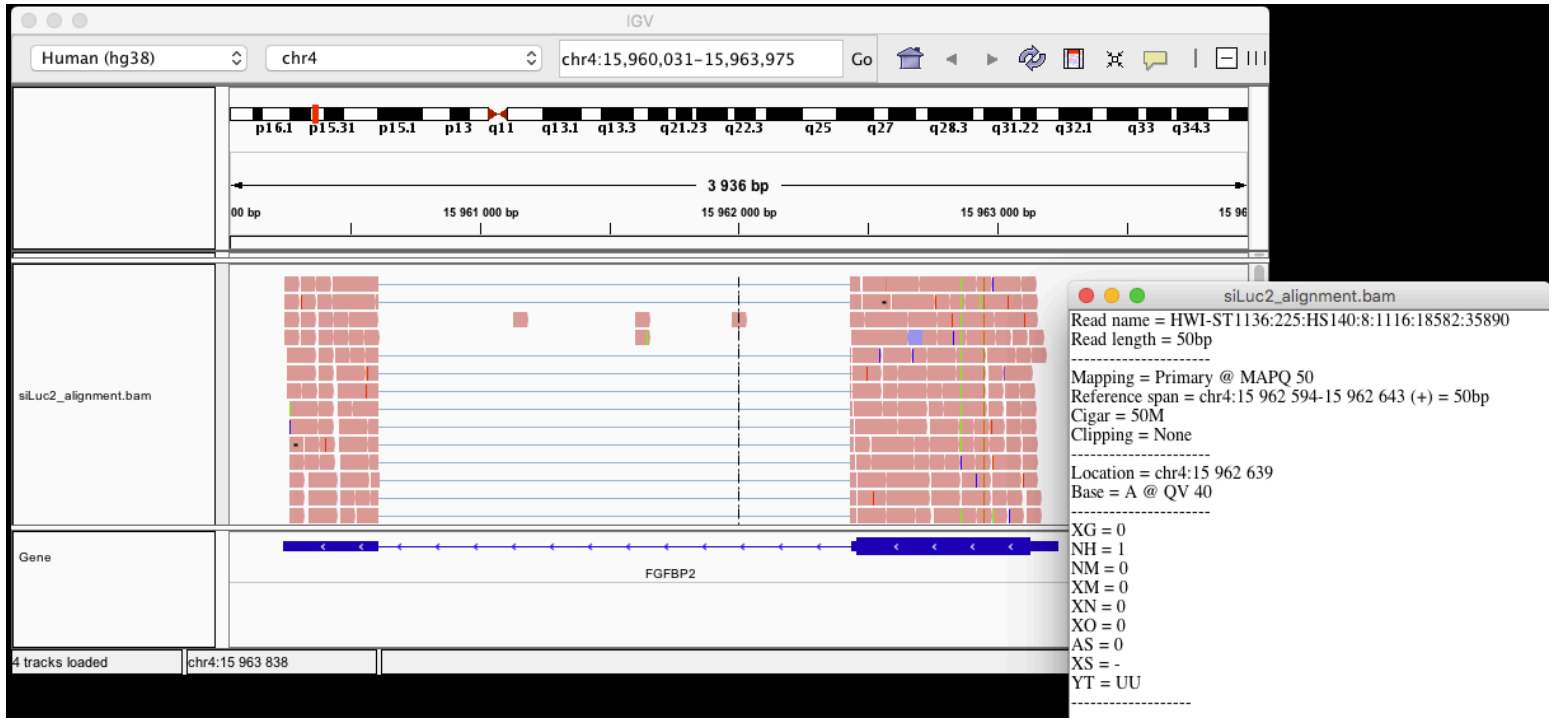
Alignment file format : SAM

- 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



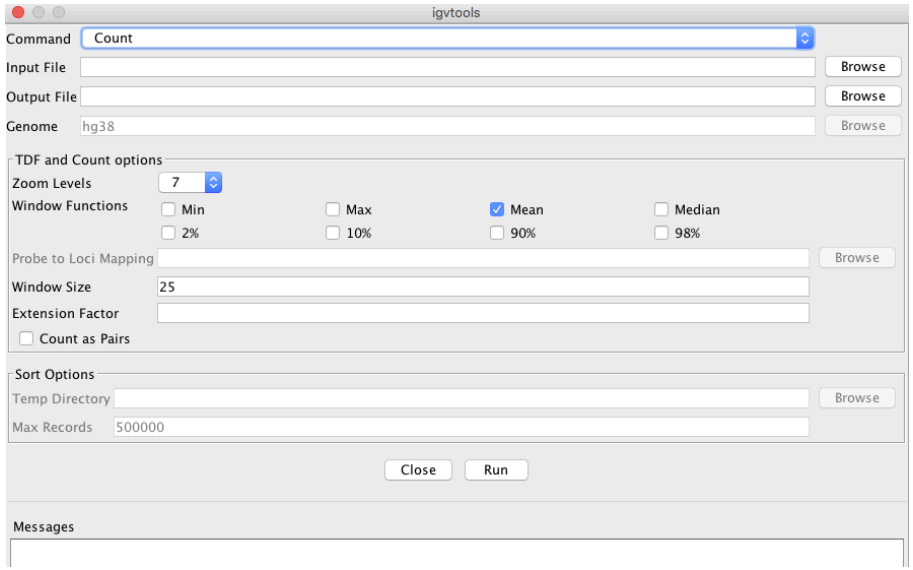
Utilities to manipulate SAM/BAM files

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

- Igvtools (http://software.broadinstitute.org/software/igv/)
  - sort, index, …
  - Integrative Genomics Viewer

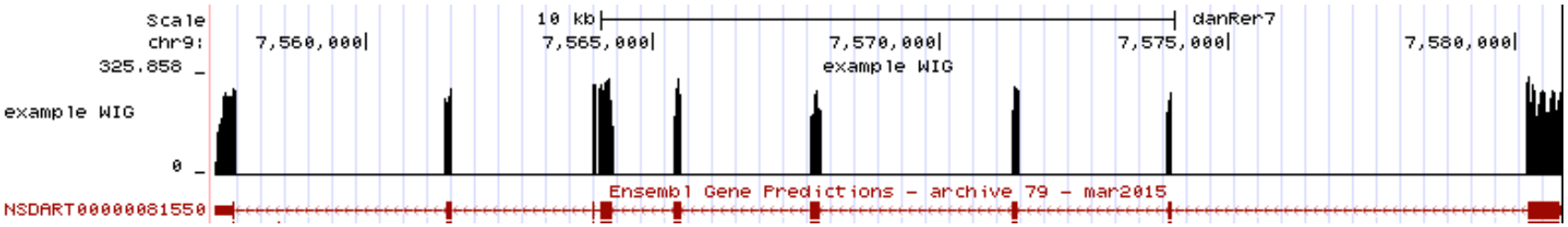
    → Tools menu

    → run igvtools



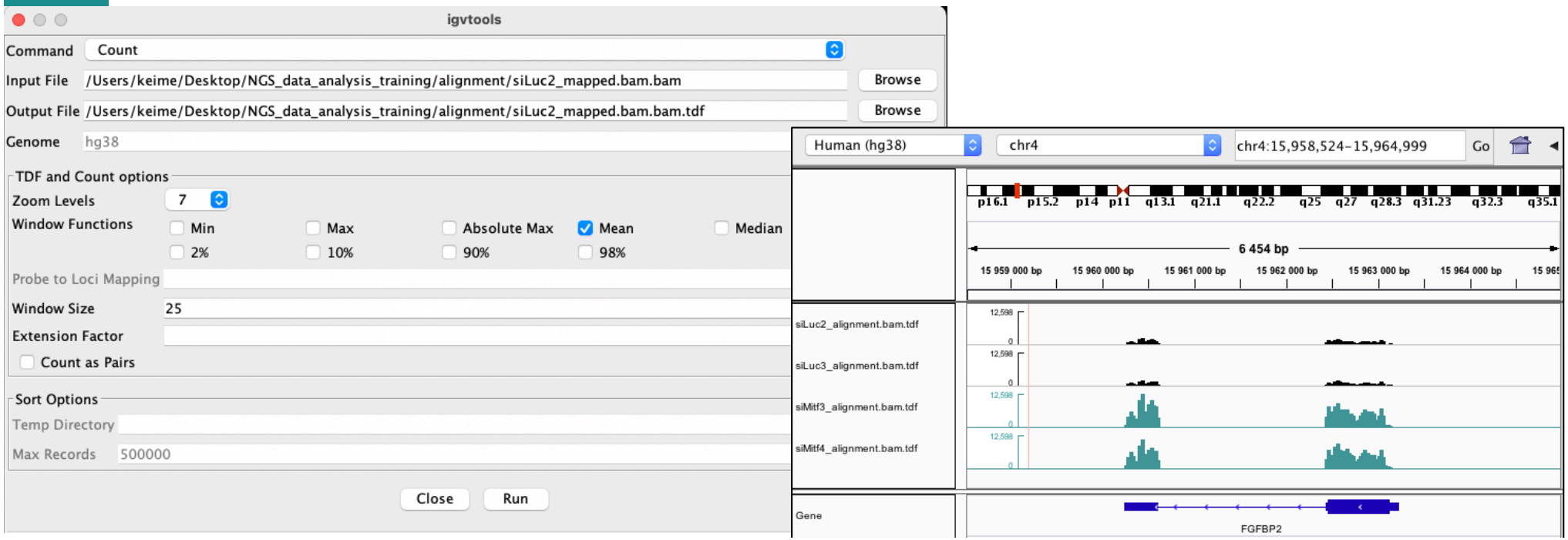
Wiggle (WIG) file format

- 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
- Compressed binary indexed file derived from a WIG file : bigWig



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

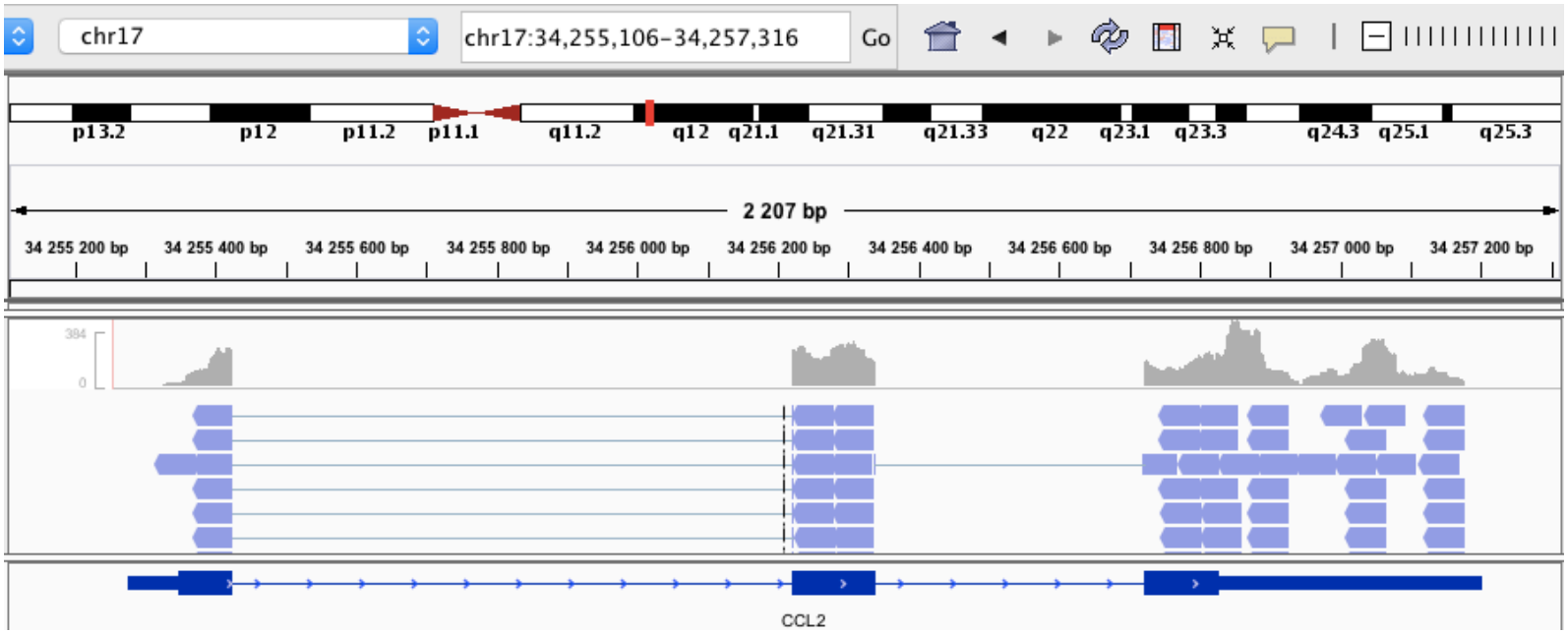


Coverage vs alignment

Coverage

Alignment

Annotation

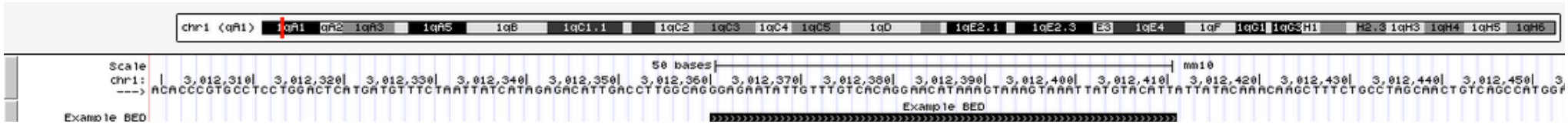


Browser Extensible Data (BED) format

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

  Chromosome, Start, End: required

  Chromosome, Start, End, Name, Score, Strand: Most common : 6 columns
- More precise definition and examples
  - http://genome.ucsc.edu/FAQ/FAQformat.html#format1
- Manipulation of BED files
  - BEDTools (different tools available on Galaxy) : https://bedtools.readthedocs.io



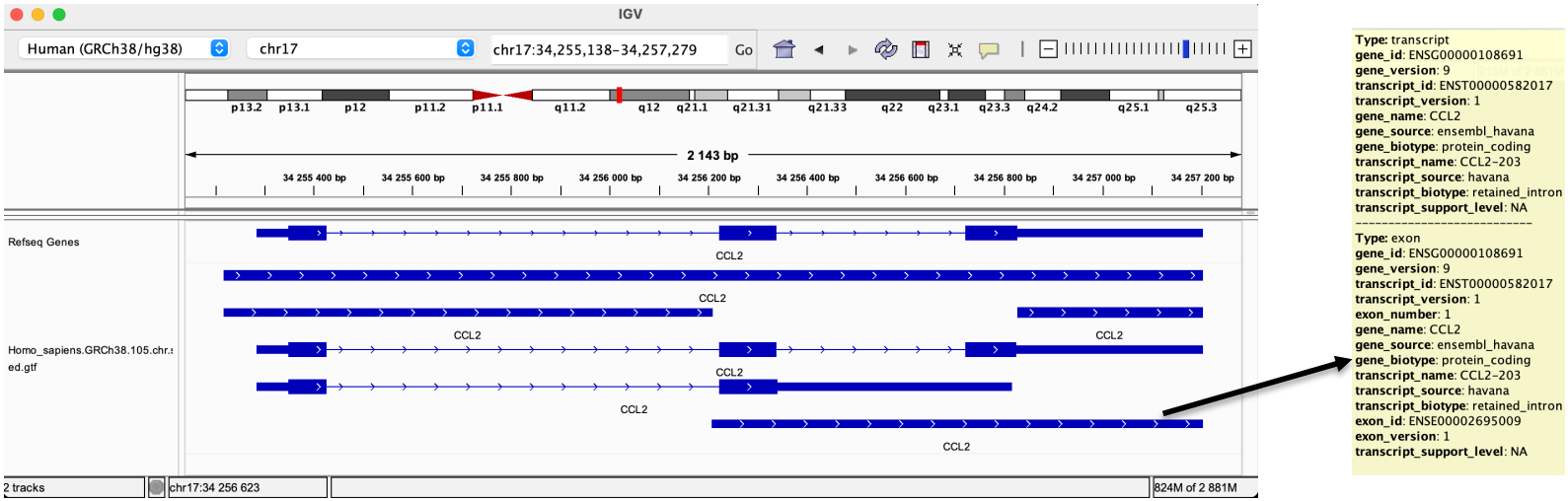
Gene Transfert Format (GTF)

- GTF files can be visualized using IGV
  - e.g. Ensembl 105 annotations

  downloaded from
  http://ftp.ensembl.org/pub/release-105/gtf/homo_sapiens/Homo_sapiens.GRCh38.105.chr.gtf.gz
- Sort (by start position) and index for faster display
  - Tools → Run igvtools → Sort

  → Homo_sapiens.GRCh38.105.chr.sorted.gtf
  - Tools → Run igvtools → Index

  → Homo_sapiens.GRCh38.105.chr.sorted.gtf.idx (in the same directory)
  - File → Load from file and choose Homo_sapiens.GRCh38.105.chr.sorted.gtf



Main NGS file formats : summary

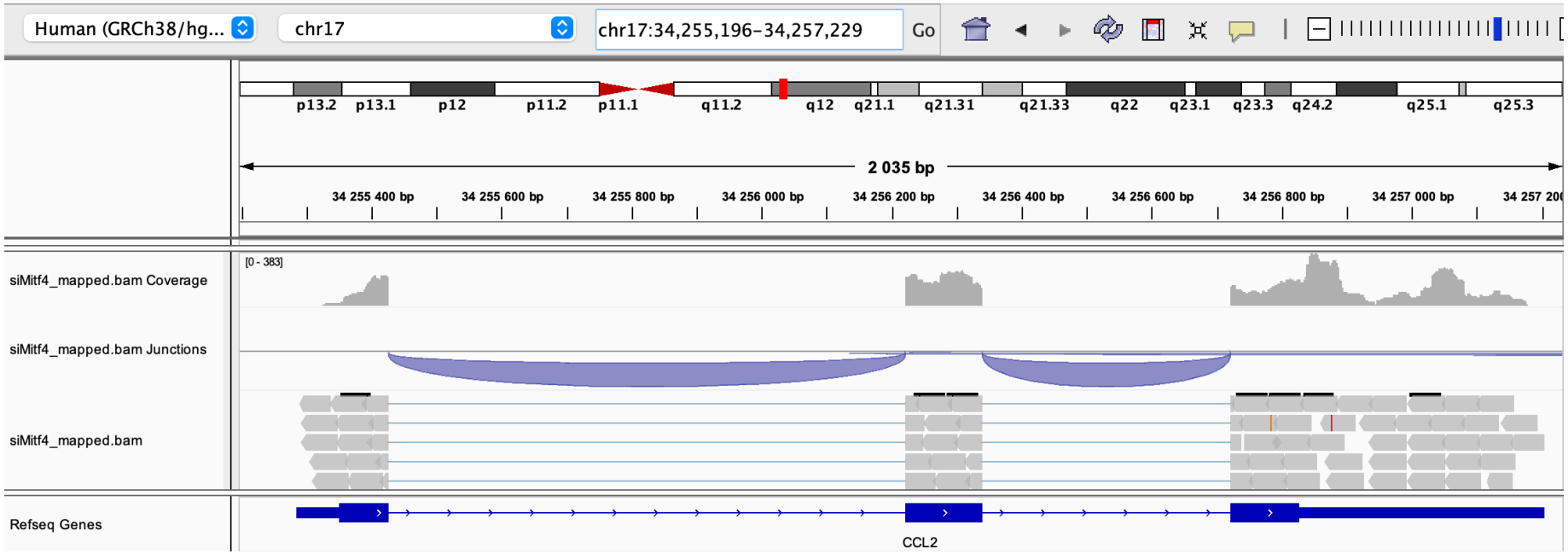
- FASTQ — text
  - Raw data
- SAM / BAM — binary
  - alignment
- WIG / bigWig / TDF
  - coverage
- BED
  - Genomic intervals
- GTF
  - annotations

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
  - *Exercise 3 : QC on alignments*

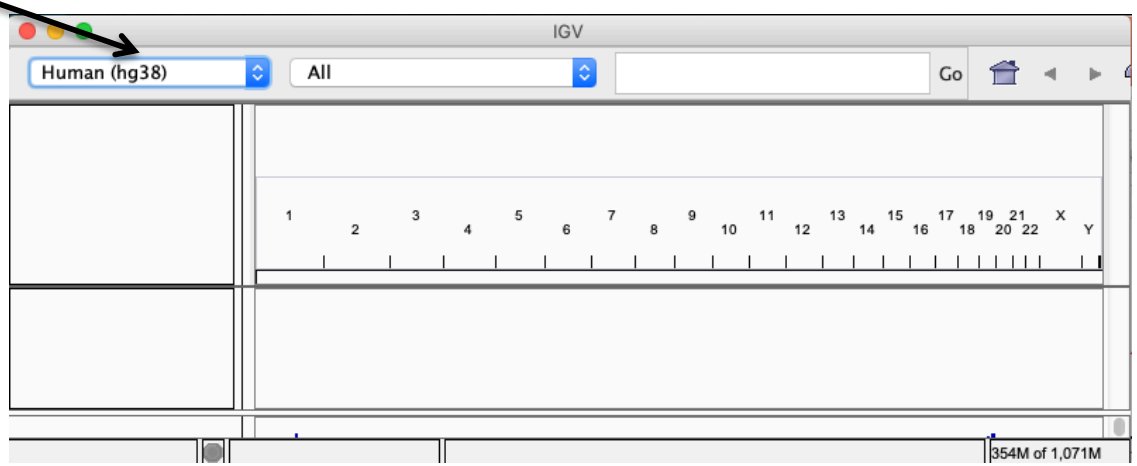
Alignment visualization

- Using a Genome Browser
  - A lot of available genome browsers
    - Ensembl, UCSC, Jbrowse, IGB, IGV, …
  - During this training we will use Integrative Genomics Viewer
    - http://www.broadinstitute.org/igv/



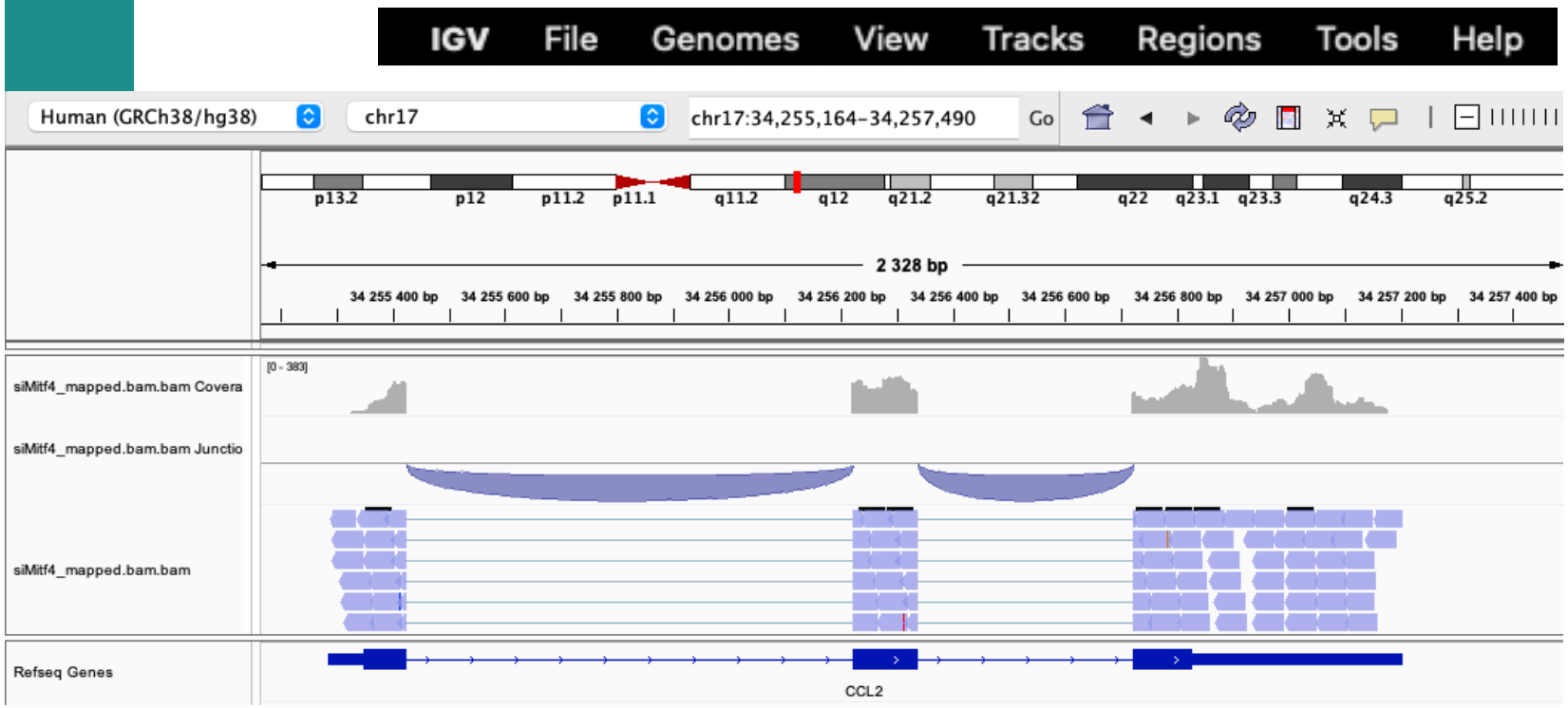
Using IGV : basic steps

- Select a reference genome



- Load data
  - File → load from file
  - File → load from server
  - Many tracks from different formats can be visualized on the same window (but they must correspond to the same assembly !)
- Navigate through the data

IGV



Menu

Tool bar

Chromosome ideogram

Data tracks

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

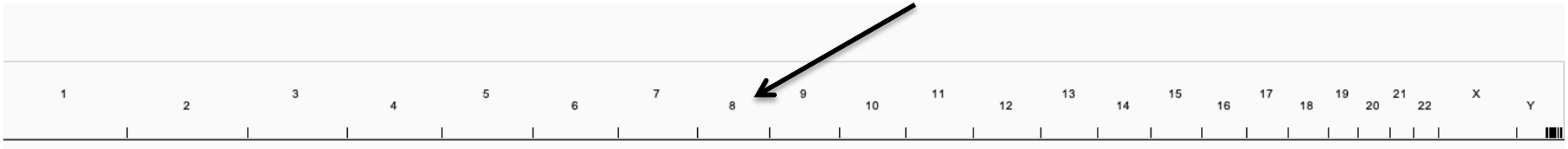
Move backward and forward through views of the genome like the back and forward buttons in a web browser

Zoom in and out on a chromosome

Modify popup text behaviour in data panels
- Show details on hover
- Show details on click

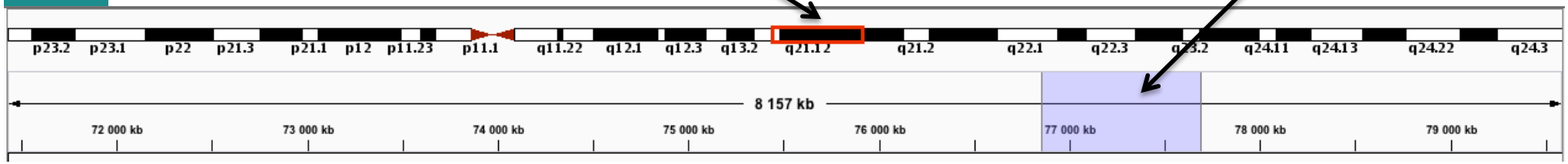
IGV : chromosome ideogram

Click on a chromosome number to jump to this chromosome



Chromosome location being shown

Click and drag to define a region to zoom in

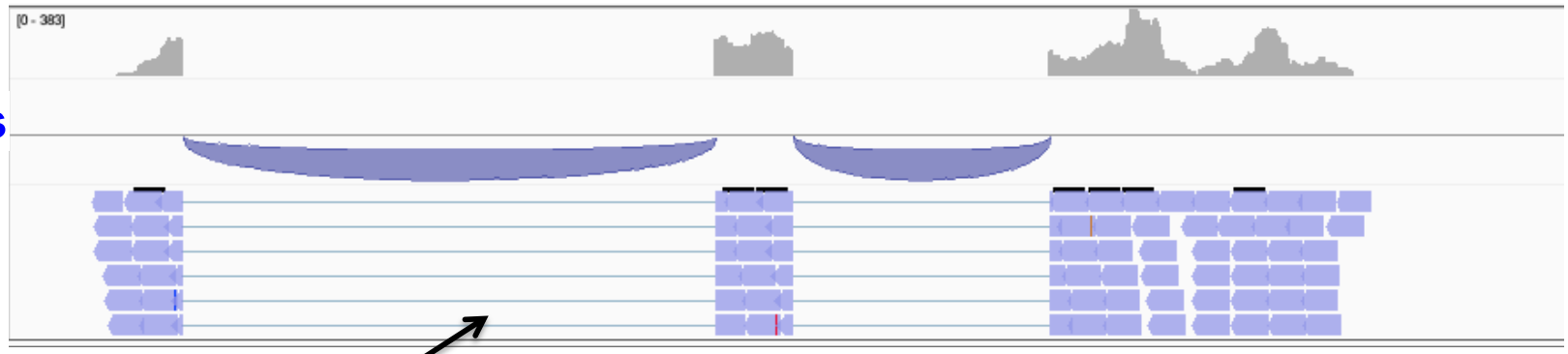


IGV : Data track

Coverage

Splice junctions

Reads



By default a sample of the alignments, to use less memory
(can be changed in View → Preferences → Alignments)

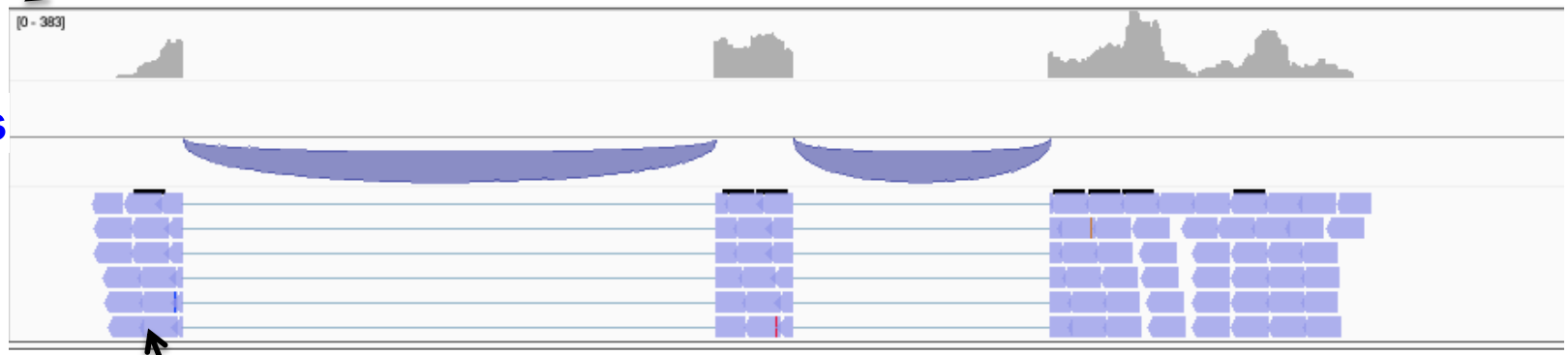
IGV : Data track

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

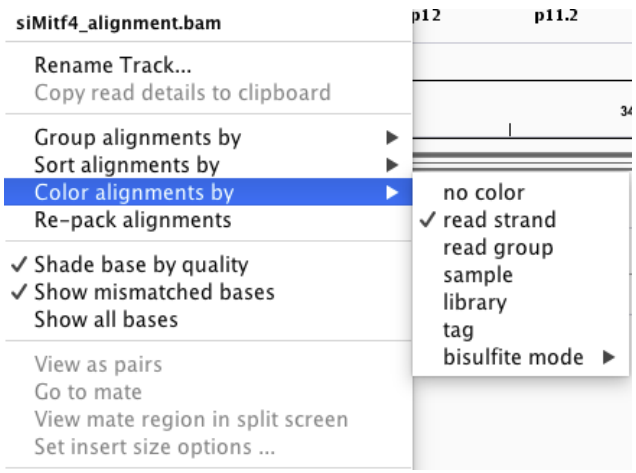
Coverage

Splice junctions

Reads



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

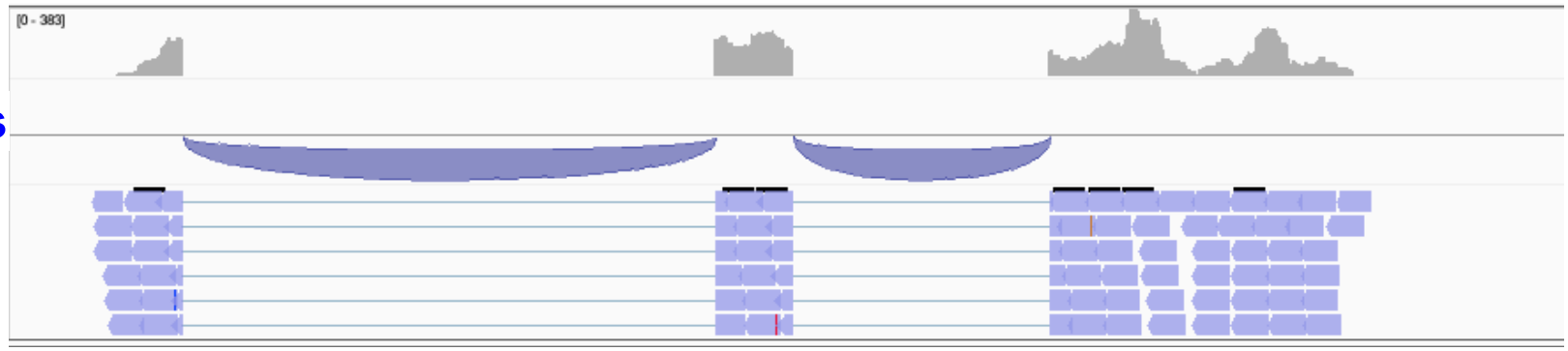


IGV : Data track

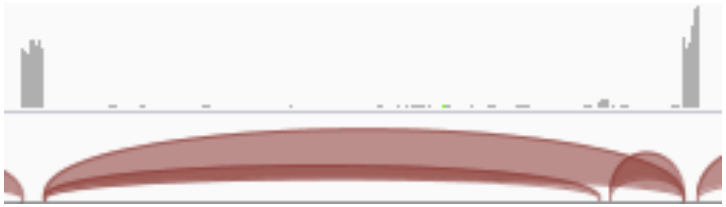
Coverage

Splice junctions

Reads



- Display of splice junctions
  - Color → strand
  - Thickness → depth of coverage
    - All junctions with more than 50 reads have the same thickness

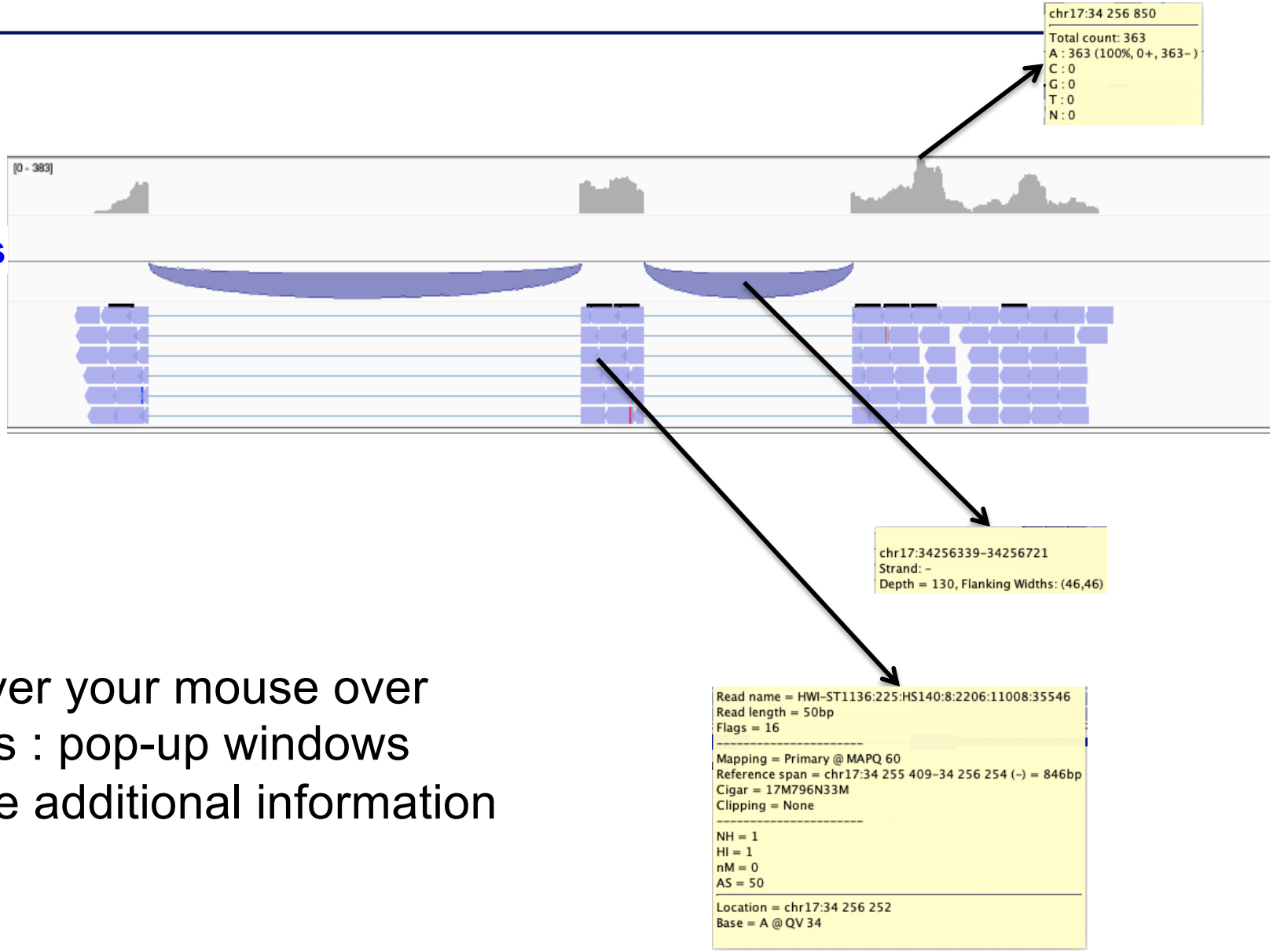


IGV : Data track

Coverage

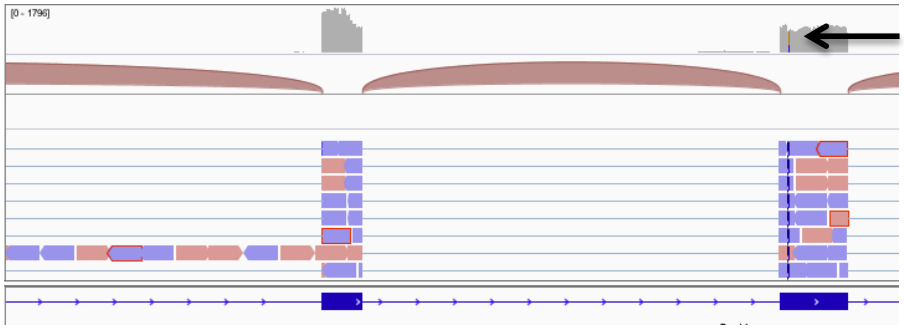
Splice junctions

Reads



→ Hover your mouse over images : pop-up windows provide additional information

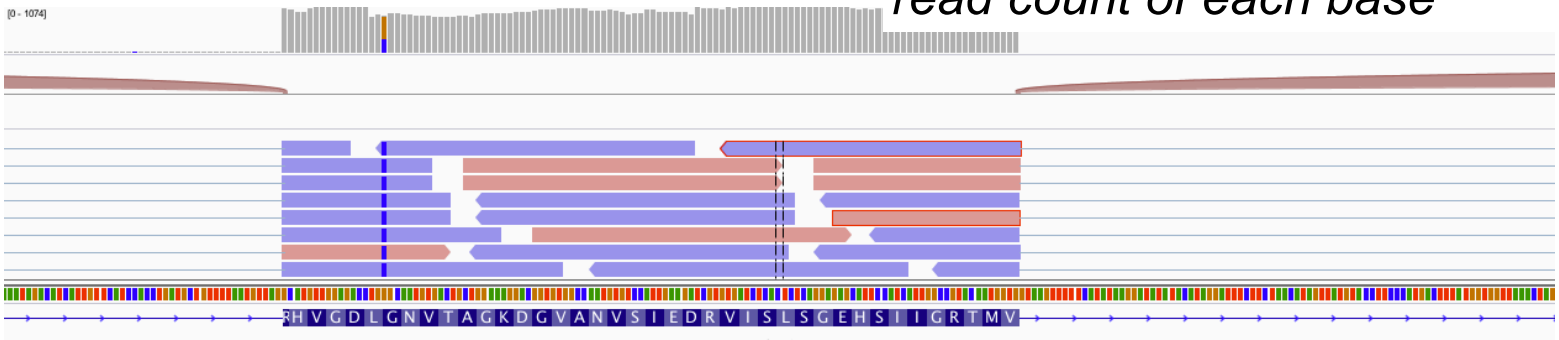
IGV data track differences vs reference genome



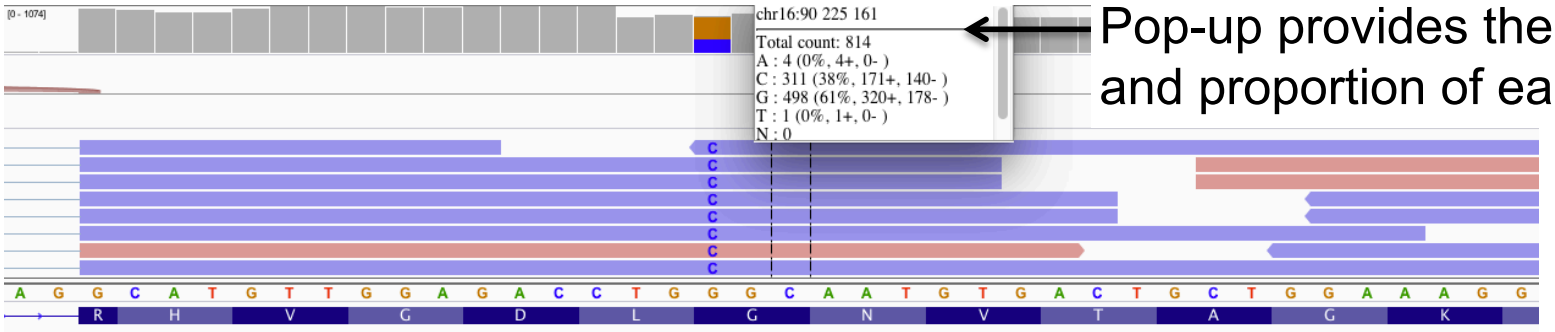
Nucleotide difference compared to the reference genome

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



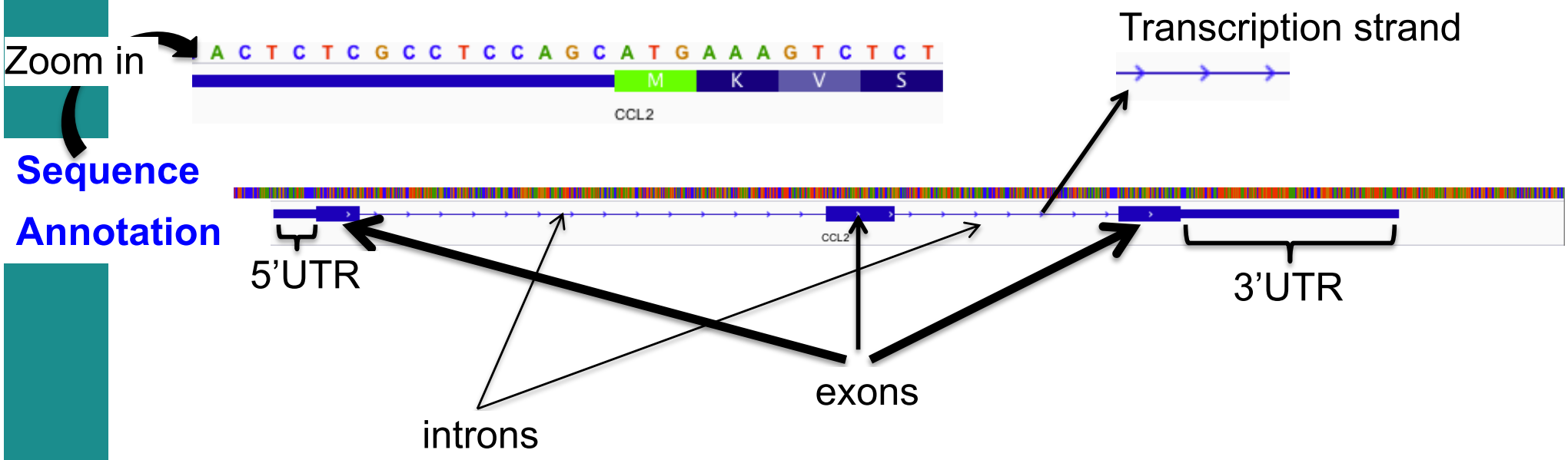
Zoom in



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



→ Hover your mouse over images, pop-up windows provide additional information :

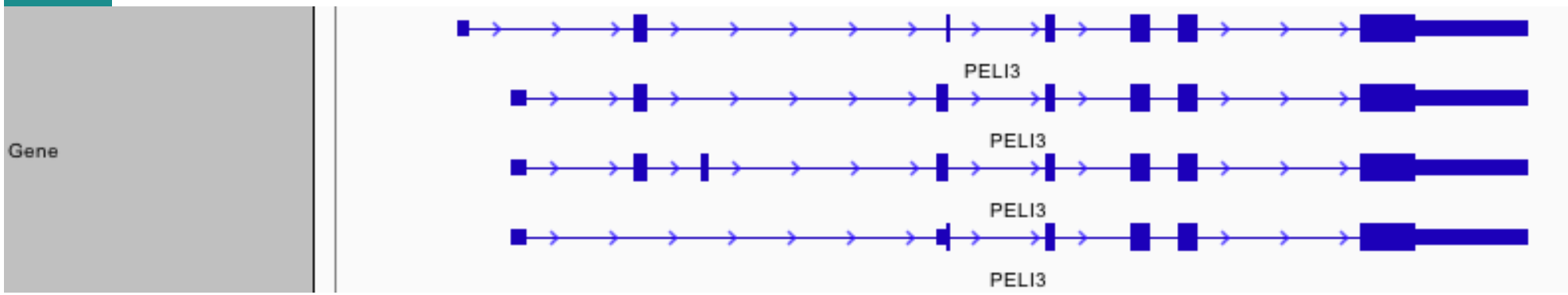
CCL2
chr17:34255277-34257201
id = NM_002982

Exon number: 2
Amino acid coding number: 51
chr17:34256222-34256339

IGV annotation track

Default : collapsed

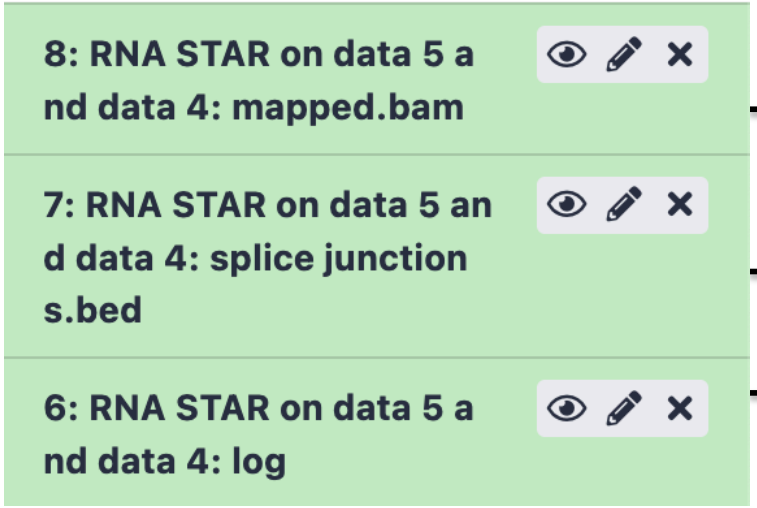
Right click on track name → Expanded
To see all isoforms



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 3 : QC on alignments*

Exercise 1 : results



→ Reads alignment

→ Splice junction

→ General information on alignment

Exercise 1 : interpretation of results

1. Log file
   - What is the proportion of uniquely mapped reads ?

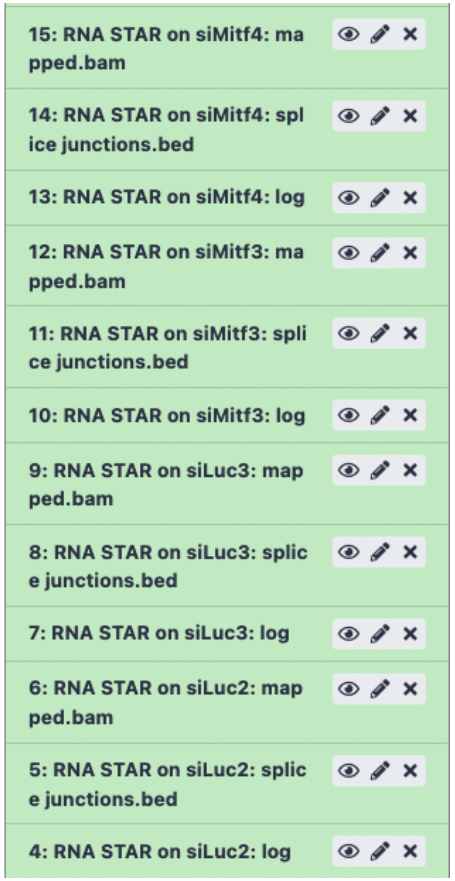
2. Alignment file
   - Which alignment file format is provided by STAR ?
   - Download this file and the index, visualize this alignment using IGV
   - Look at reads mapped on the junction between the 2 last exons of *Park7* gene. How many reads span this junction ? Look at the CIGAR string of one of these reads
   - Visualize the strand specificity of the reads, for example on *Park7* and *Chmp2a* genes (color alignments by strand)
   - Look at reads aligned on *Actb* gene (color alignments by number of reported alignments : tag=NH). What do you observe ?

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 3 : QC on alignments*

Exercise 2 : whole dataset alignments

- STAR results for all samples from Mitf project are available on Galaxy
  - Datasets 4 to 15:



Exercise 2 : whole dataset alignments

1. What is the proportion of uniquely mapped reads in all samples ?

- To save time, the corresponding BAM, BAI and tdf files are already available on your computer (RNAseq/alignment folder)
- Start a new IGV session (File → new session)
- In View → Preferences → Tracks tab, select “Normalize coverage data”
- Load the 4 tdf files on IGV
- Right-click on all track names and choose “Group Autoscale”

2. We are interested in *Idh1* gene.
   Is this gene differentially expressed between siLuc and siMitf samples ?

Exercise 2 : whole dataset alignments

In IGV preferences (View → Preferences) “Alignments” tab

- In “Track Display” section, check “Show junction track”
- In “Splice Junction Track” section, choose “Minimum junction coverage”: 10

Open a new session (File → New session), then load the 4 BAM files

3. What do you observe in exons 11 and 13 of *Eef2* gene ?
4. What do you observe at position chr4:6707961 ?
5. Which transcript isoforms do you observe in region chr20:44,935,294-44,939,521 ?

Notes :

- To see all annotated isoforms right click on an annotation track and select Expanded
- You can perform a Sashimi-plot for a better visualization of isoforms : Right-click on a BAM track → Sashimi plot → Select Alignment Tracks : all alignments

Exercise 2 : whole dataset alignments

6. 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_mapped.bam

   What do you think about this protocol ?

   Look for example at *Park7* gene

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 3 : QC 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
- Read distribution relative to known annotations
- Read coverage over genes
- Strand information (directional protocol)
- For paired-end sequencing : distance between reads

http://rseqc.sourceforge.net/

RSeQC tools available on Galaxy

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

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

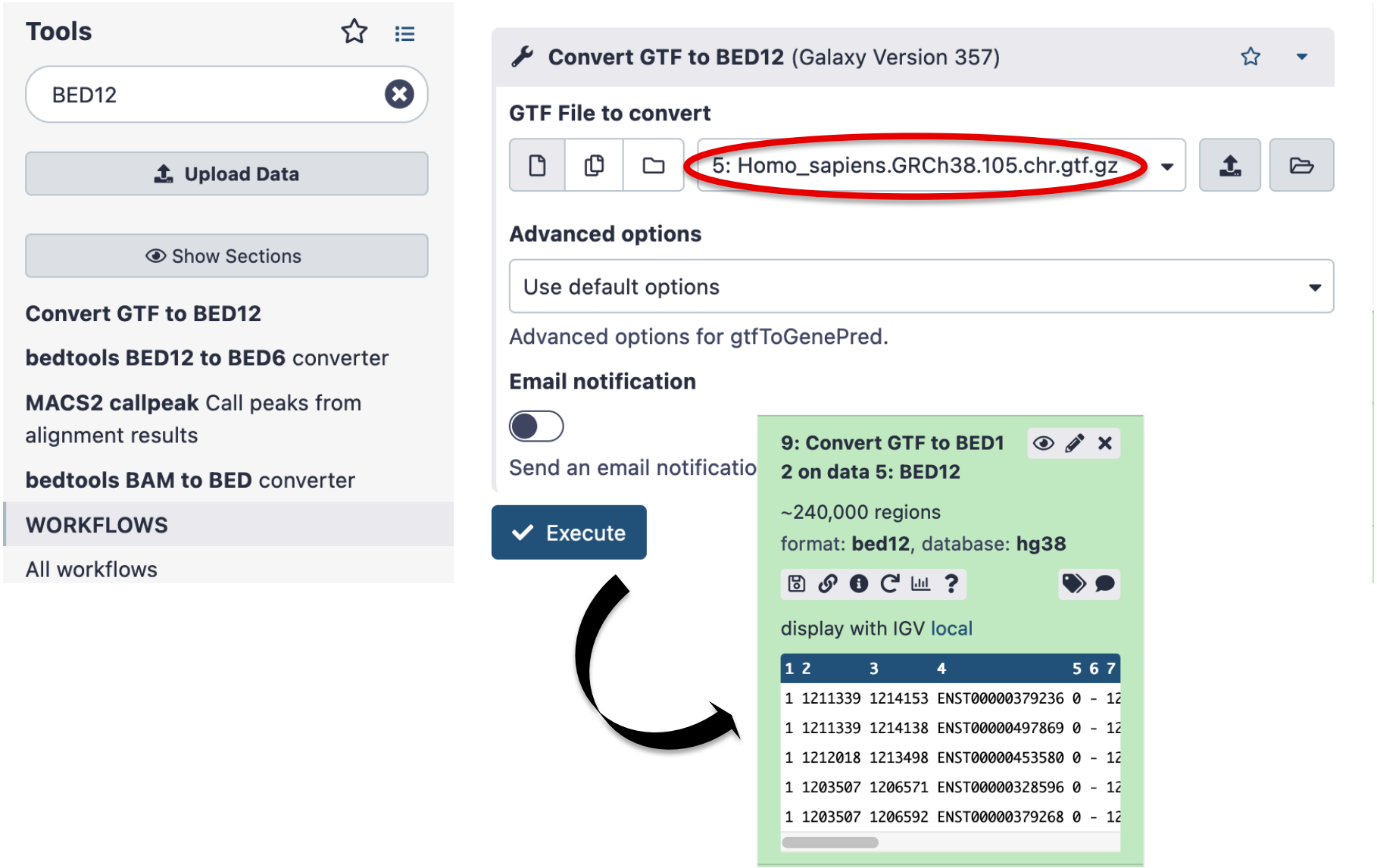
Exercise 3 – Question 1

1. Convert GTF annotation file to BED file using **Convert GTF to BED12** tool
   - Annotation file to use (already imported)

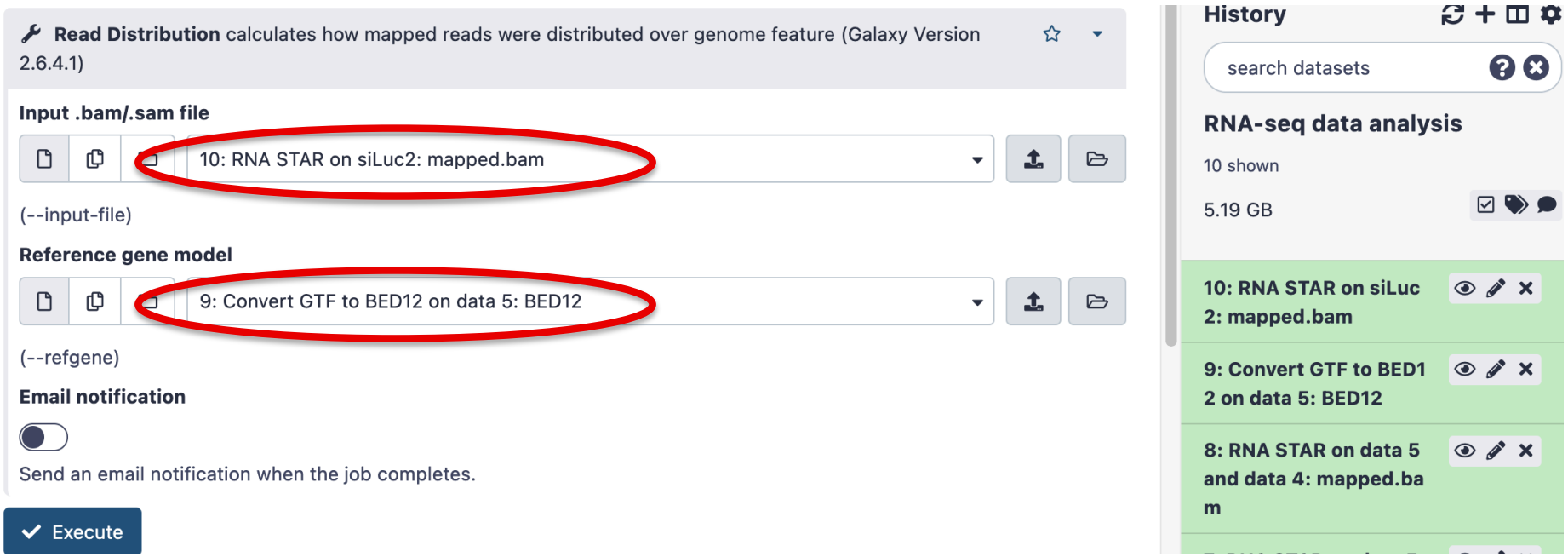
     Homo_sapiens.GRCh38.105.chr.gtf.gz

2. Launch **Read distribution** on the mapping results from siLuc2 sample
   - Alignment file to import
     - 6: RNA STAR on siLuc2: mapped.bam
   - Annotations
     - Bed file obtained during step 1

Exercise 3 – Question 1



Exercise 3 – Question 1



Exercise 3 – Question 1

- Read distribution

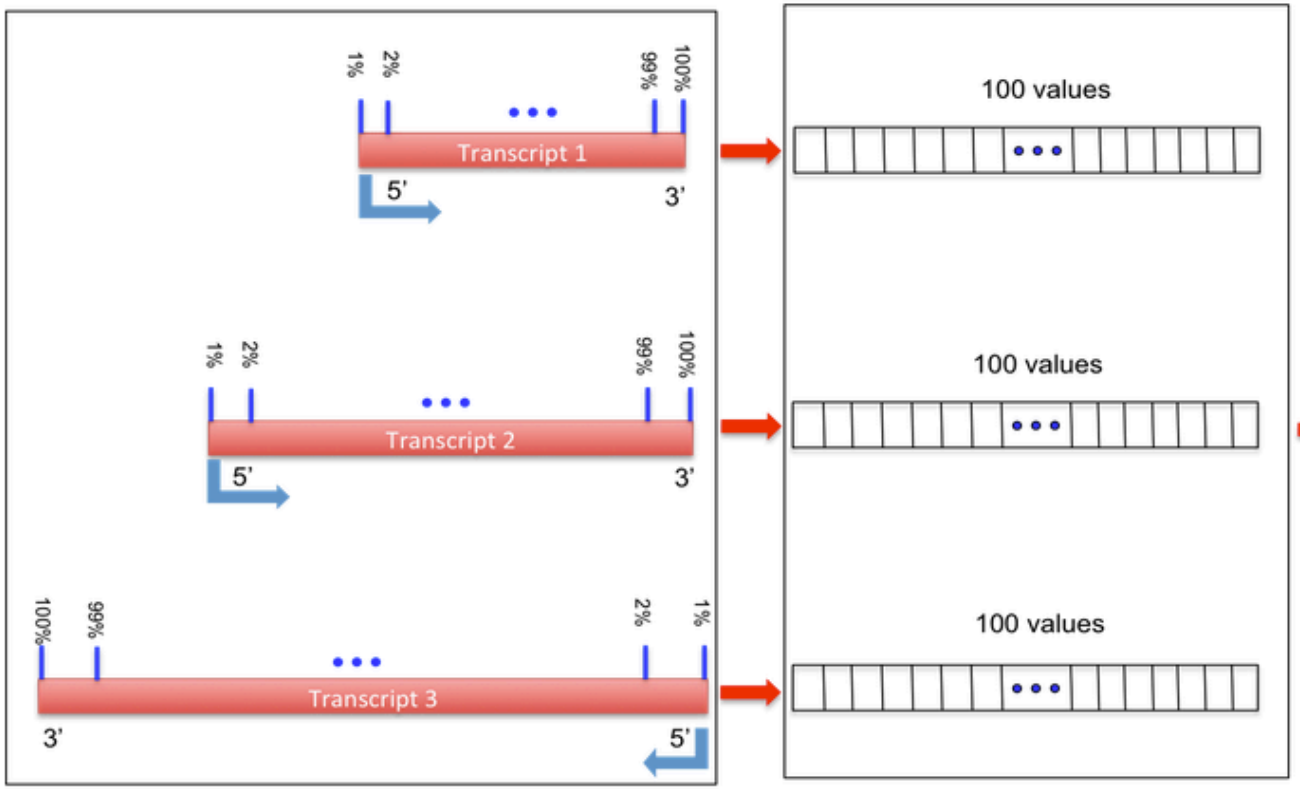
```
Total Reads                    43080660
Total Tags*                    49982200
Total Assigned Tags°           46821353
=================================================================
Group                Total_bases          Tag_count            Tags/Kb
CDS_Exons            35875018             30475010             849.48
5'UTR_Exons          48312525             2562470              53.04
3'UTR_Exons          75444264             10416230             138.07
Introns              1613277549           3000938              1.86
TSS_up_1kb           28388776             36020                1.27
TSS_up_5kb           126596357            66649                0.53
TSS_up_10kb          225614841            90707                0.40
TES_down_1kb         30986381             124903               4.03
TES_down_5kb         133535951            201514               1.51
TES_down_10kb        233464669            275998               1.18
=================================================================
```

* 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 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”

Read coverage over genes

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



From http://rseqc.sourceforge.net/

Read coverage over genes : Galaxy

Don’t perform this analysis today

Gene Body Coverage (BAM) Read coverage over gene body (Galaxy Version 2.6.4.3)

Run each sample separately, or combine mutiple samples into one plot

Combine multiple samples into a single plot

Input .bam file(s)

4: RNA STAR on siMitf4: mapped.bam
3: RNA STAR on siMitf3: mapped.bam
2: RNA STAR on siLuc3: mapped.bam
1: RNA STAR on siLuc2: mapped.bam

(--input-file)

Reference gene model

5: Convert GTF to BED12 on data 5: BED12

(--refgene)

Minimum mRNA length (default: 100)

100

Minimum mRNA length in bp, mRNA that are shorter than this value will be skipped (--minimum_length).

Output R-Script

No

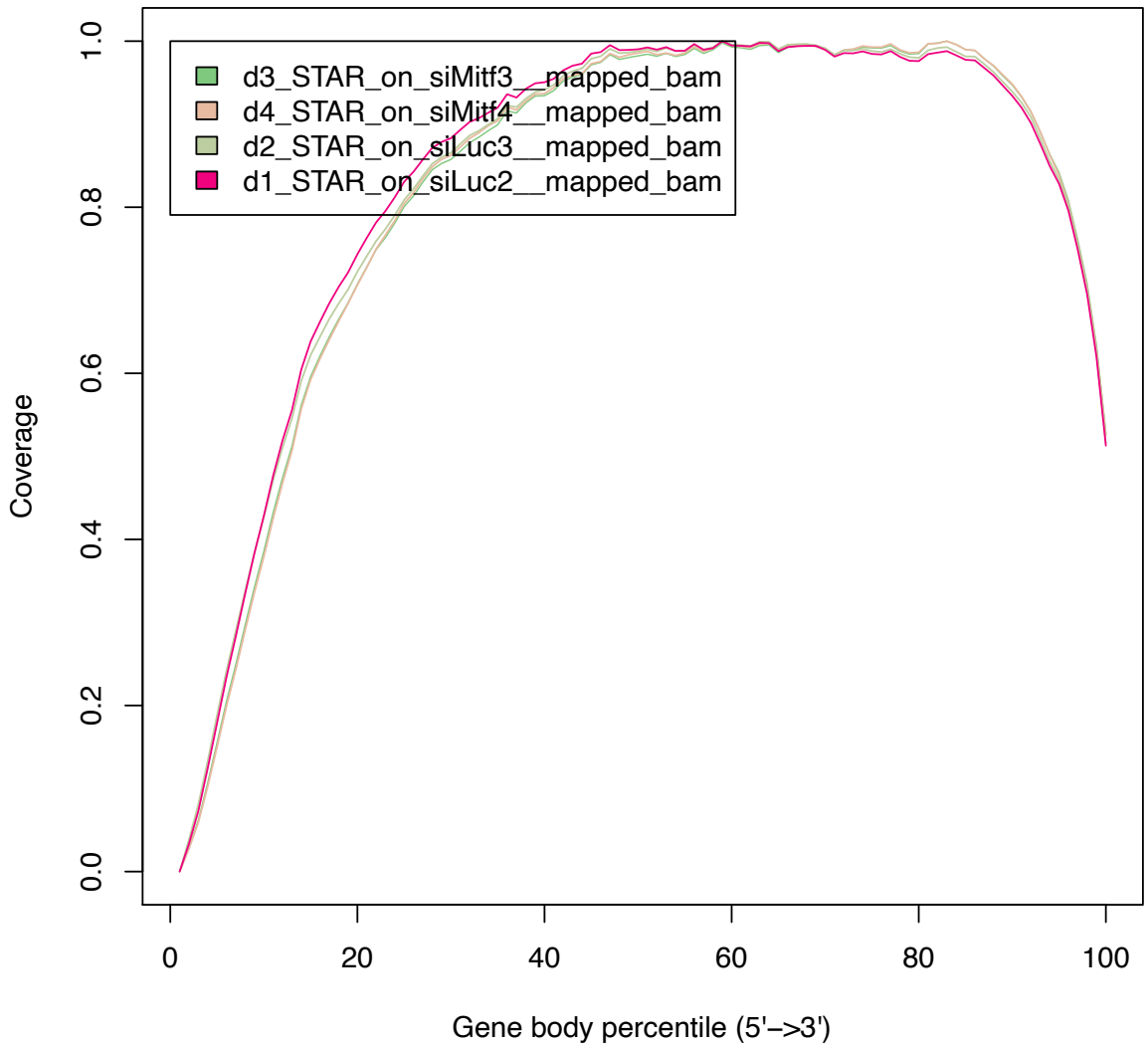
Output the R-Script used to generate the plots

Email notification

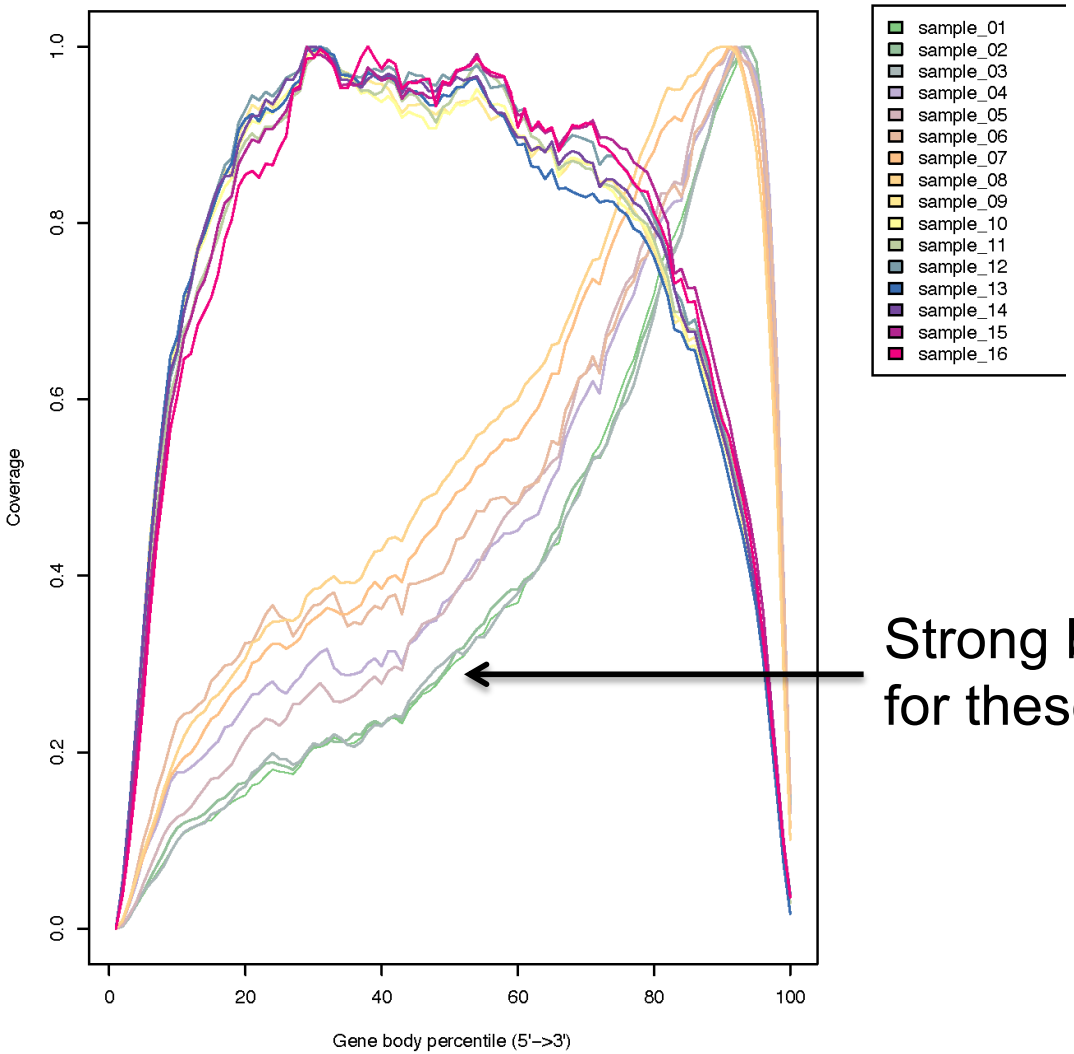
Send an email notification when the job completes.

Execute

Read coverage over genes : result



Read coverage over genes : example with biased samples



Strong bias in read coverage for these samples

Strand information (directional protocol)

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

Exercise 3 – Question 2

- Launch **Infer experiment** on the mapping results obtained on siLuc2 data from the two different protocols and compare the two results
  - Alignment files
    - RNA STAR on siLuc2: mapped.bam (already imported)
    - 16: RNA STAR on siLuc2_other_protocol: mapped.bam (to import)
  - Annotations
    - Bed file obtained during the previous exercise

Exercise 3 – Question 2

- Infer experiment on siLuc2 mapping results :

Infer Experiment speculates how RNA-seq were configured (Galaxy Version 2.6.4.1)

Input .bam file

10: RNA STAR on siLuc2: mapped.bam

(--input-file)

Reference gene model

9: Convert GTF to BED12 on data 5: BED12

(--refgene)

Number of reads sampled from SAM/BAM file (default = 200000)

200000

(--sample-size)

Minimum mapping quality

30

Minimum mapping quality for an alignment to be considered as "uniquely mapped" (--mapq)

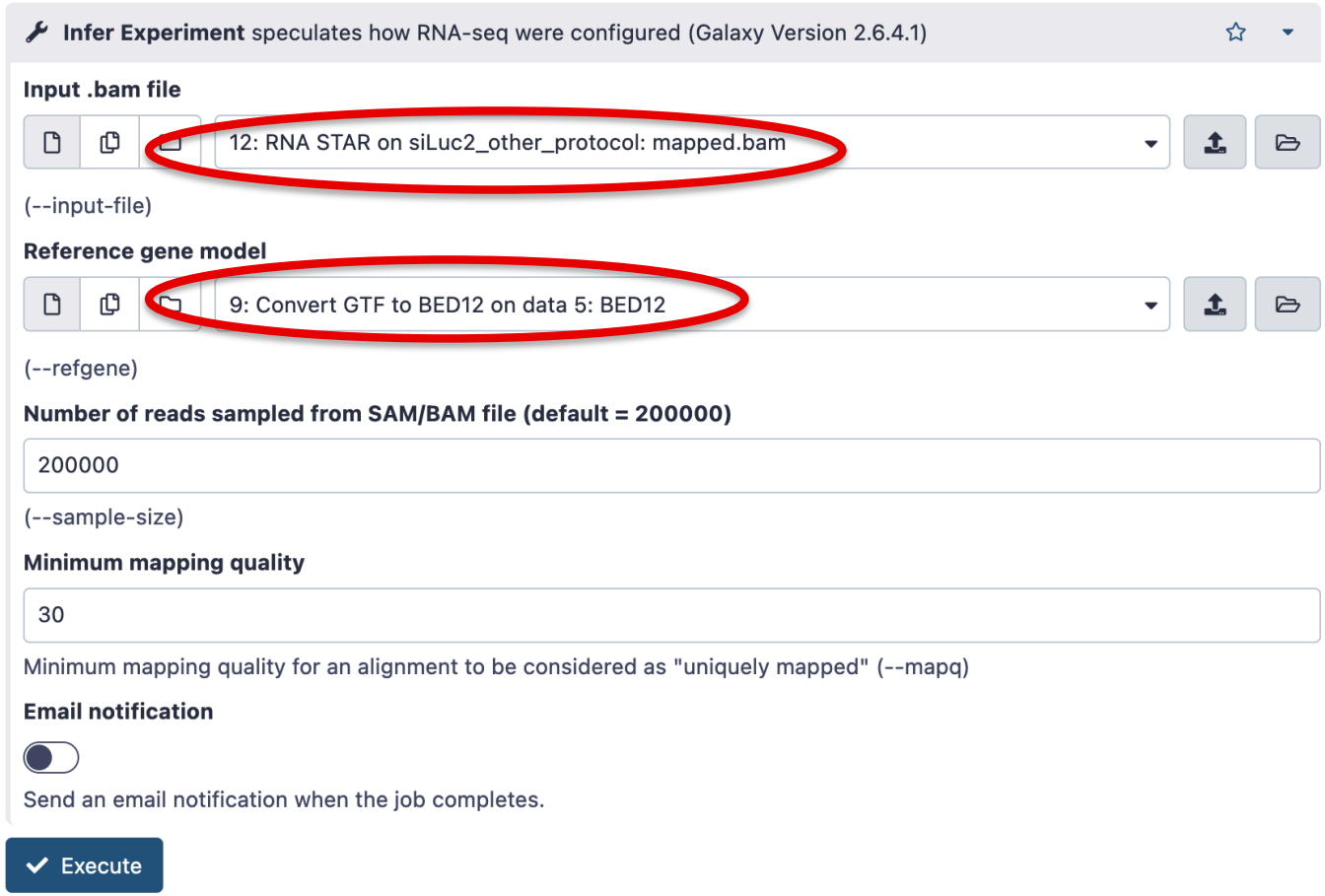
Email notification

Send an email notification when the job completes.

Execute

Exercise 3 – Question 2

- Infer experiment on siLuc2 mapping results from the library prepared with another protocol :



Exercise 3 – Question 2

- Infer experiment
  - on siLuc2 library prepared with a directional protocol :

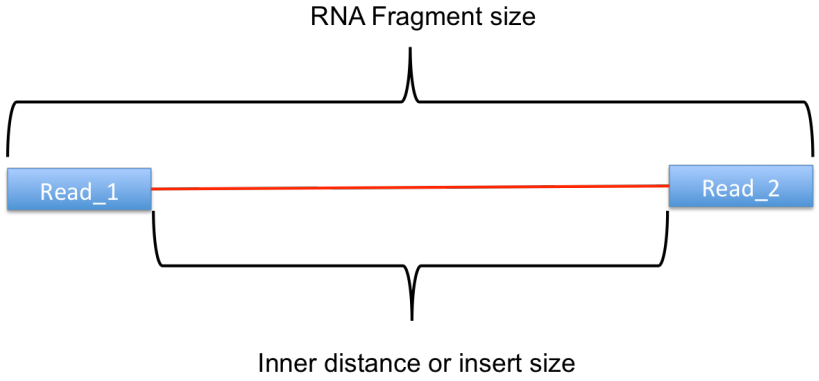
```
This is SingleEnd Data
Fraction of reads failed to determine: 0.1034
Fraction of reads explained by "++,--": 0.0078
Fraction of reads explained by "+-,-+": 0.8887
```

  - On siLuc2 library prepared with a non directional protocol :

```
This is SingleEnd Data
Fraction of reads failed to determine: 0.1446
Fraction of reads explained by "++,--": 0.4278
Fraction of reads explained by "+-,-+": 0.4277
```

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