



NGS read mapping

Céline Keime
keime@igbmc.fr

NGS read mapping

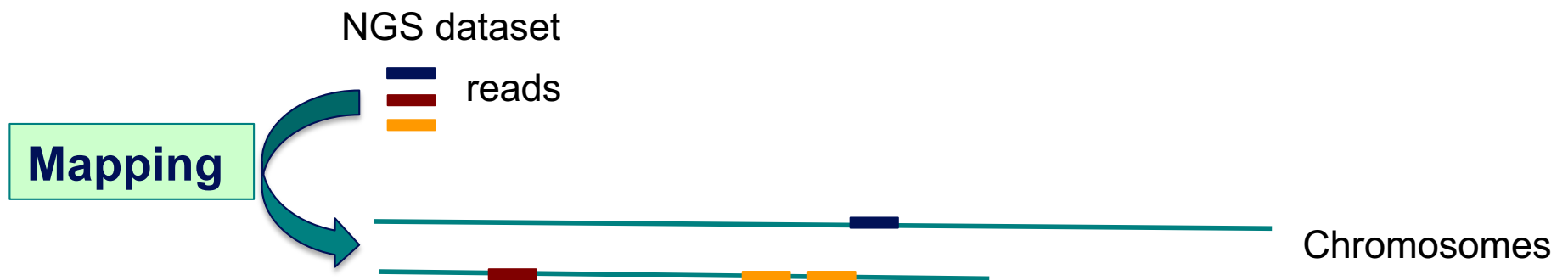
- Introduction to NGS read mapping
- Short read mappers
- Specificity of RNA-seq read mapping
- *Exercise 1 : Mapping of RNA-seq data using Galaxy*
- Alignment and related file formats
- Alignment visualization
- *Exercise 1 : Interpretation of results*
- *Exercise 2 : Whole dataset alignment visualization*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

CACGTACC
CACGT**T**CC

mismatch

CACGTA_CC
CACGTA**T**CC

indels (insertion/deletion)

CACGTACC
CACGT_**_**CC

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

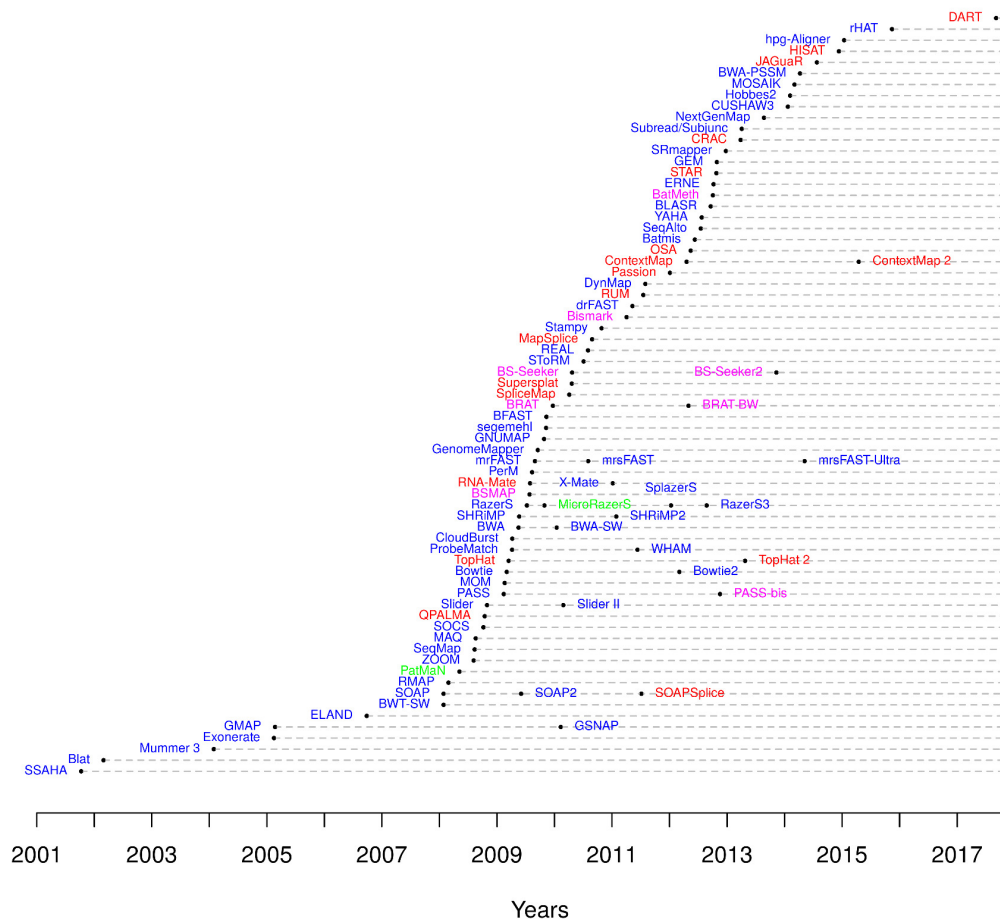
- Uses a technique originally developed for compressing large files called the Burrows-Wheeler transform (BWT)
 - The transformed human genome fits into memory

■ 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



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, bisulfite), support of paired-end
- Read length limits
- Quality aware
- Multi-mapping reporting
- Speed
- Memory requirements

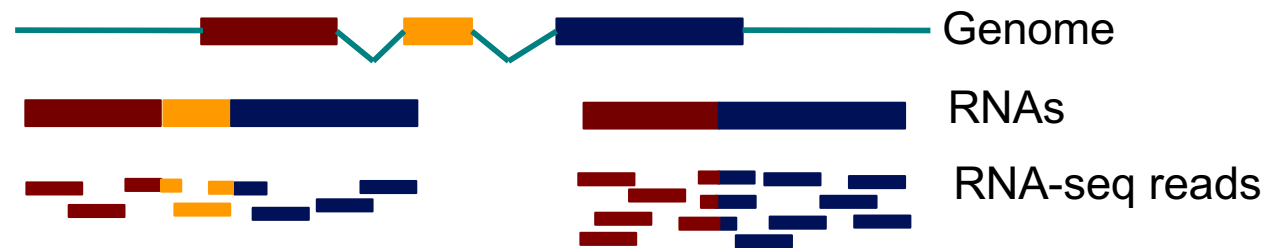
■ Feature comparison

- 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*

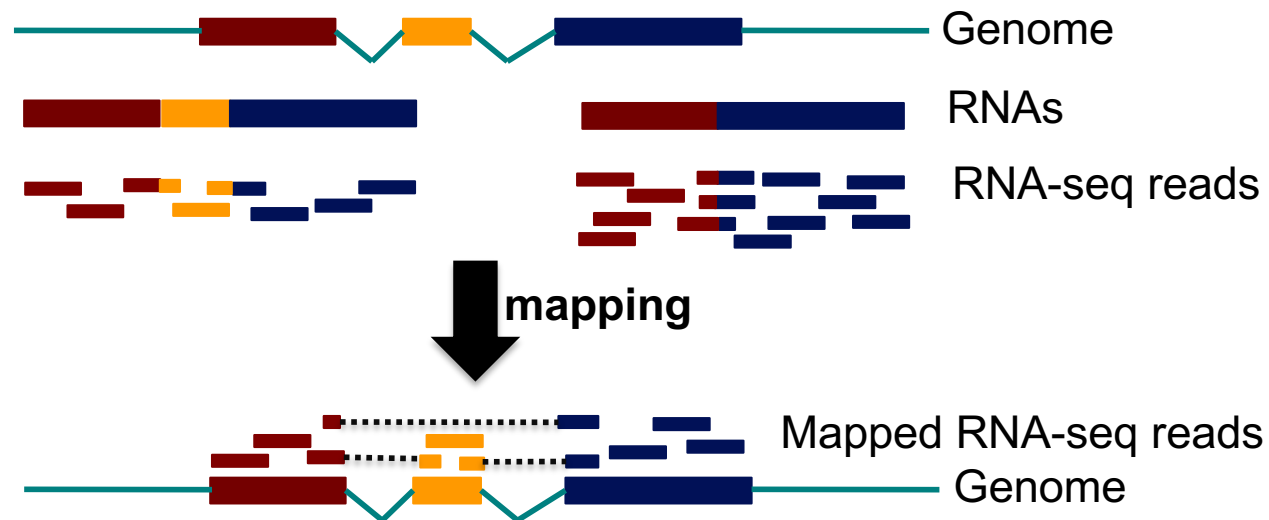
Specificity of RNA-seq reads



→ 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

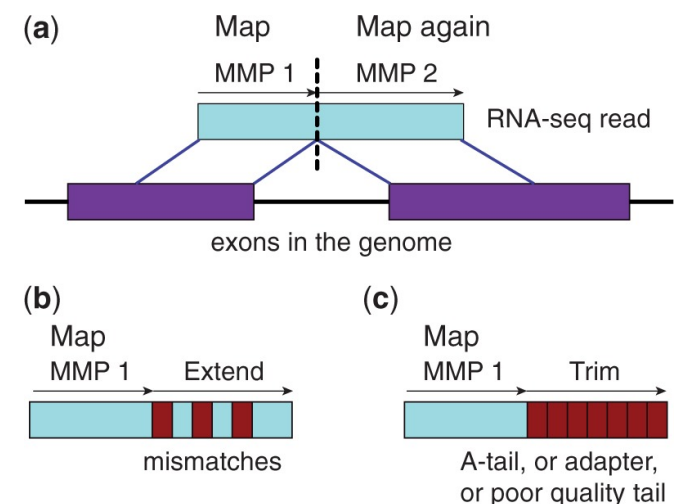
1. Searching for seeds

- 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
- STAR uses a suffix array to efficiently search for the MMPs → allows for quick searching against large reference genomes
- 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

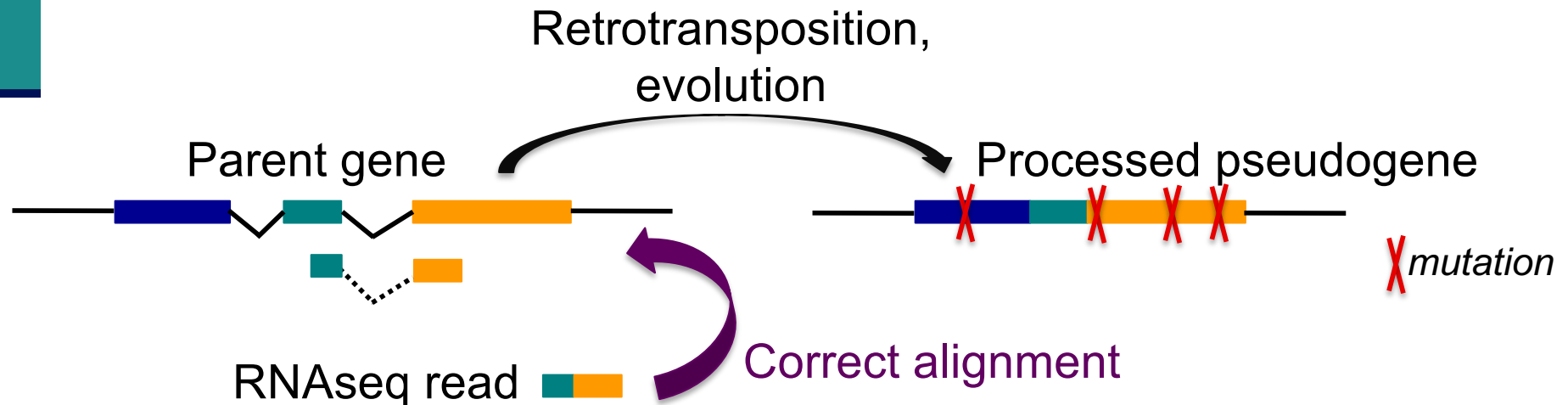


Main limits of *de novo* RNA-seq aligners

- 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](#)
 - Annotations for some species and Ensembl versions already available on GalaxEast
- The main Ensembl site focuses on vertebrate genomes and some other representative 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)

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 :
 - <https://mblab.wustl.edu/GTF22.html>
 - e.g. human Ensembl 95 GTF file
 - http://ftp.ensembl.org/pub/release-95/gtf/homo_sapiens/Homo_sapiens.GRCh38.95.chr.gtf.gz
 - Caution : use annotations corresponding to the version of genome assembly you are working on

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. Import the corresponding FASTQ file in your history
 2. Launch STAR on this FASTQ file

Exercise 1

1. Import the FASTQ file in your history

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

- Import this file in your current history

[Download](#) **to History** [Modify](#) [Permissions](#)

Libraries / NGS data analysis training / RNAseq / rawdata / siLuc2_1000000.fastq

This dataset is unrestricted so everybody can access it. Just share the URL of this page. [To Clipboard](#)

Name	siLuc2_1000000.fastq
Data type	fastqsanger
Genome build	hg38
Size	150.2 MB
Date uploaded (UTC)	2016-09-09 08:27 AM
Uploaded by	keime@igbmc.fr
Miscellaneous blurb	150.2 MB
Miscellaneous information	uploaded fastq file

```
@HWI-ST1136:225:HS140:8:1101:1169:2100 1:N:0:ACTTGA
CTTTGCTATTGTGAATAGTGCTGCATGAACATATACATGCATGTCTTT
+
CCCCFFFFHHHHIIJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ
```

Exercise 1

2. Launch STAR

RNA STAR Gapped-read mapper for RNA-seq data (Galaxy Version 2.6.0b-1)

Options

Single-end or paired-end reads

Single-end

Type of sequencing (single or paired-end)

RNA-Seq FASTQ/FASTA file

6: siLuc2_1000000.fastq

FASTQ file

Custom or built-in reference genome

Use a built-in index

Built-ins were indexed using default options

Reference genome with or without an annotation

use genome reference with builtin gene-model

Reference genome with annotation

Must the index have been created with a GTF file (if not you can specify one afterward).

Select reference genome

hg38+ensembl95

Reference genome (assembly name)
+ annotation source and version

If your genome of interest is not listed, contact the Galaxy team (--ger

Count number of reads per gene

Yes No

column 1: gene ID, column 2: counts for unstranded RNA-seq, column 3: counts for the 1st read strand aligned with RNA, column 4: counts for the 2nd read strand aligned with RNA. This requires either (A) an index that was built with an annotation (GTF or GFF3 file) or (B) having specified an annotation (GTF or GFF3 file above). (--quantMode)

Would you like to set output parameters (formatting and filtering)?

No

Other parameters (seed, alignment, limits and chimeric alignment)

Use Defaults

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
 - *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:1.0 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/segment.
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 AACCCCTAACCCCTAACCCCTAACCCCTAACCCCTAACCCCTAAC
FEJJHHFBJJIHGBJJIIGIJJHGGCJJIIHFJJIIHFHHHHHDFFFFCBB 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 ACGCAGCTCCGCCCTCGCGGTGCTCTCCGGTCTGTGCTGAGGAGAACGCA
BBBBDDDDDFHHJJIGJIIJJIIJJIIJJJJJJJJJJJHHHHHHHFFFCBB 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 CCGCCCTCGCGGTGCTCTCCGGTCTGTGCTGAGGAGAACGCAACTCCGCC
DDDCDDFHIIJJJJIIHJJJJIIJJJJJJJJJJJJGHHHHHFFFCBB 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)

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

- STAR mapping quality

- 255 for uniquely mapped reads

- $\text{int}(-10 \cdot \log_{10}(1 - 1/N_{\text{map}}))$ for multi-mapping reads

- Nmap : the number of loci a read maps to

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 ¹	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, $BAQ_i = Q_i - (BQ_i - 64)$ 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 <code>(uint16_t) round(value * 100.0)</code> .
LB	Z	Library. Value to be consistent with the header RG-LB tag if @RG is present.
HO	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. <i>Regex</i> : <code>[0-9]+((([A-Z] \^[A-Z]+)[0-9]+)*⁶</code>
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

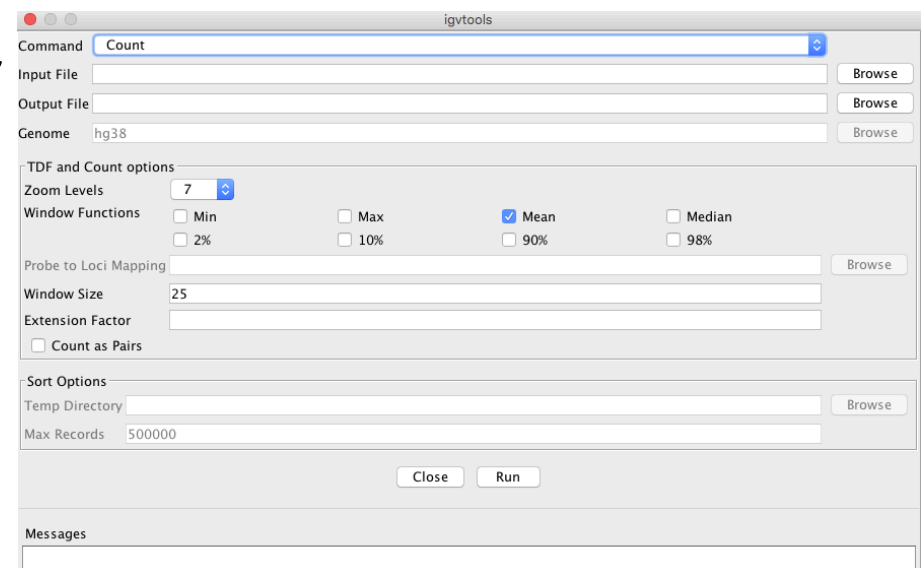
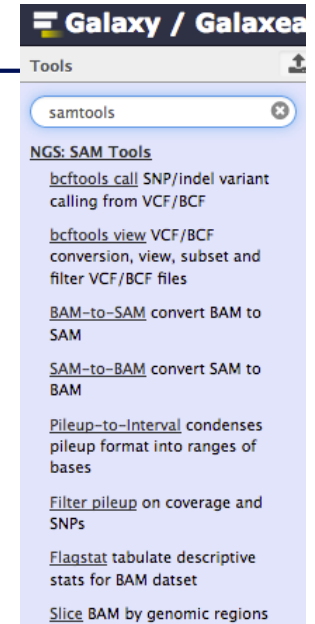
Alignment file format : BAM

- 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



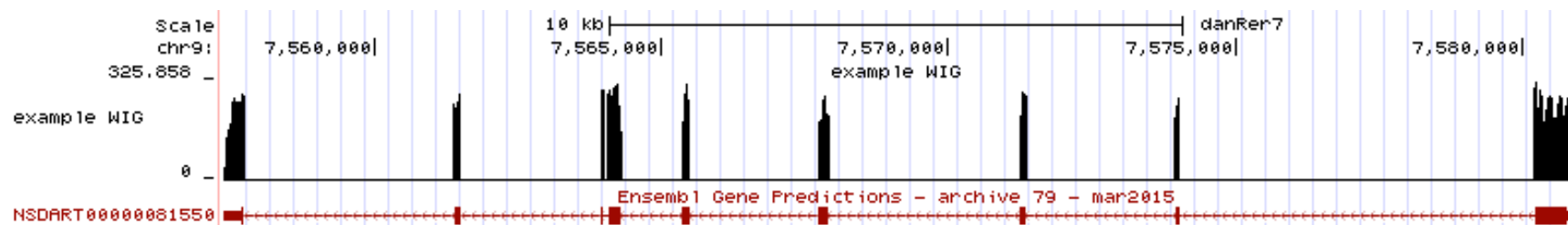
Utilities to manipulate SAM/BAM files

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



TDF file format

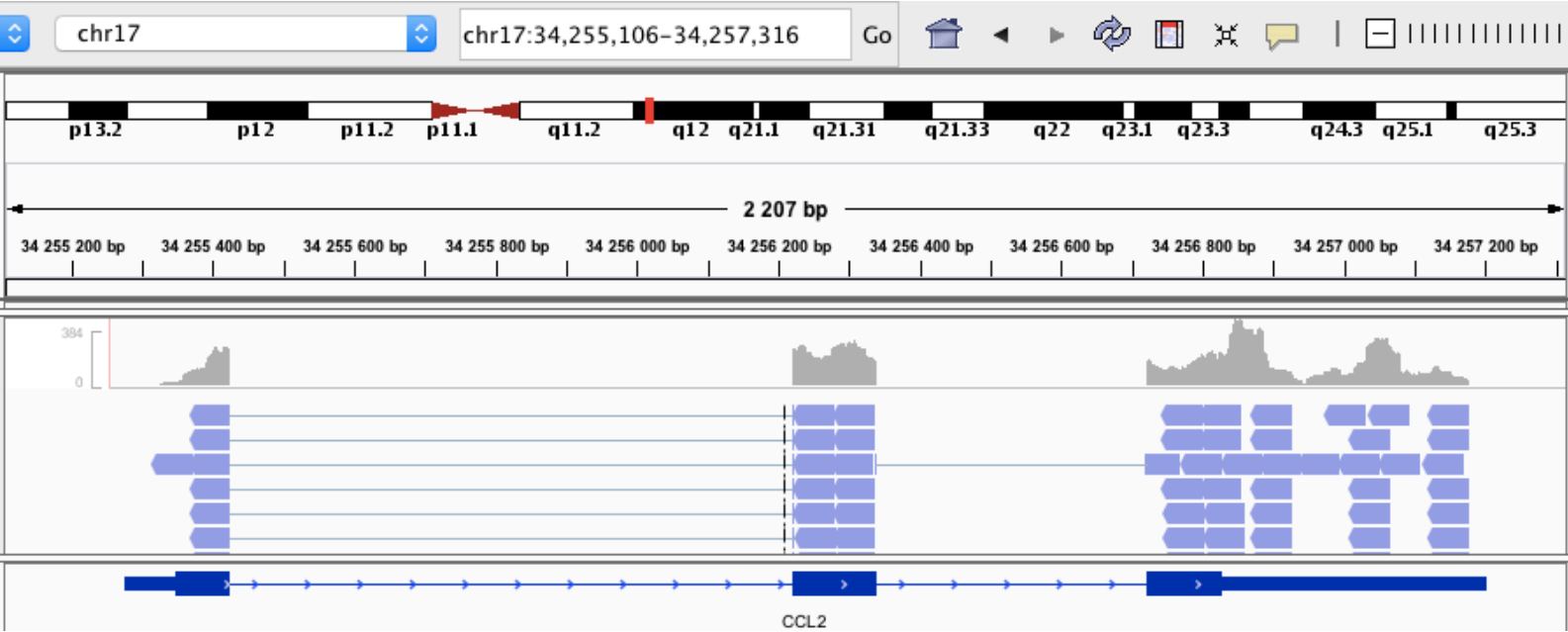
- Tiled data file
- Binary file
- Read count density
 - Pre-processed data for faster display in IGV
- TDF file can be computed from a BAM file using igvtools
 - IGV Tools menu → run igvtools → Count

The screenshot displays the IGV Tools 'Count' window. The 'Command' dropdown is set to 'Count'. The 'Input File' is '/Volumes/rufushome/CNRStraining/analyzeddata/RNAseq/alignment/siLuc2_alignment.bam' and the 'Output File' is '/Volumes/rufushome/CNRStraining/analyzeddata/RNAseq/alignment/siLuc2_alignment.bam.tdf'. The genome is set to 'hg38'. Under 'TDF and Count options', 'Zoom Levels' is 7, and 'Window Functions' includes 'Mean' (checked), 'Min', 'Max', 'Median', '2%', '10%', '90%', and '98%'. 'Probe to Loci Mapping' is empty, 'Window Size' is 25, and 'Extension Factor' is empty. 'Count as Pairs' is unchecked. 'Sort Options' are empty, 'Temp Directory' is empty, and 'Max Records' is 500000. The 'Run' button is visible at the bottom.

The main visualization area shows a genomic track for Human (hg38) on chromosome 4 (chr4) at coordinates 15,958,524-15,964,999. The track includes a cytoband scale (p16.1 to q35.1), a 6,454 bp zoomed-in view with coordinates from 15,959,000 bp to 15,964,000 bp, and four tracks showing read count density for siLuc2, siLuc3, siMitf3, and siMitf4. The gene track at the bottom identifies the gene as FGFBP2.

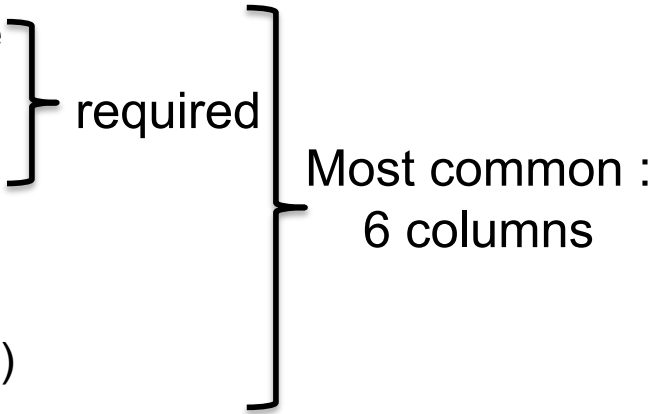
Coverage vs alignment

Coverage
Alignment
Annotation

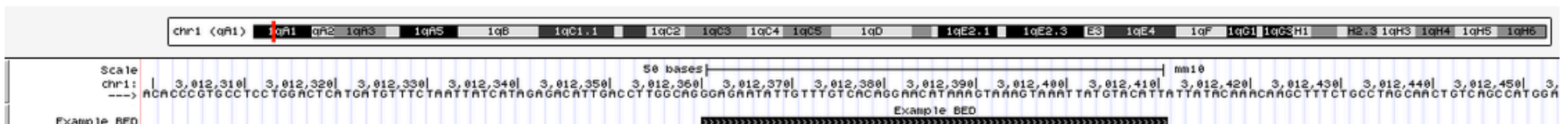
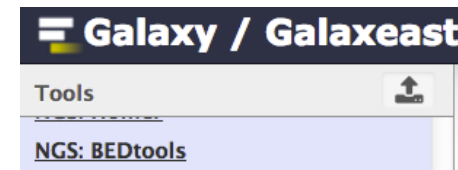


Browser Extensible Data (BED) format

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



- More precise definition and examples
 - <http://genome.ucsc.edu/FAQ/FAQformat.html#format1>
- Manipulation of BED files
 - BEDTools : <https://bedtools.readthedocs.io>



Gene Transfert Format (GTF)

- GTF files can be visualized using IGV
 - e.g. Ensembl 95 annotations
(downloaded from http://ftp.ensembl.org/pub/release-95/gtf/homo_sapiens/Homo_sapiens.GRCh38.95.chr.gtf.gz)
- Sort and index for faster display
 - Tools → Run igvtools → Sort
→ Homo_sapiens.GRCh38.95.sorted.gtf
 - Tools → Run igvtools → Index
→ Homo_sapiens.GRCh38.95.sorted.gtf.idx (in the same directory)
 - File → Load from file and choose Homo_sapiens.GRCh38.95.sorted.gtf

The screenshot displays the IGV interface for the CCL2 gene on chromosome 17. The top track shows the chromosome map with coordinates from 34,255,000 bp to 34,257,000 bp. Below this, the gene model track shows the CCL2 gene structure with exons and introns. The right-hand panel provides detailed metadata for the transcript and exon:

```
Type: transcript
gene_id: ENSG00000108691
gene_version: 9
transcript_id: ENST00000582017
transcript_version: 1
gene_name: CCL2
gene_source: ensembl_havana
gene_biotype: protein_coding
transcript_name: CCL2-203
transcript_source: havana
transcript_biotype: retained_intron
transcript_support_level: NA

Type: exon
gene_id: ENSG00000108691
gene_version: 9
transcript_id: ENST00000582017
transcript_version: 1
exon_number: 1
gene_name: CCL2
gene_source: ensembl_havana
gene_biotype: protein_coding
transcript_name: CCL2-203
transcript_source: havana
transcript_biotype: retained_intron
exon_id: ENSE00002695009
exon_version: 1
transcript_support_level: NA
```

Main NGS file formats : summary

- FASTQ

- Raw data

text

binary

- SAM / BAM

- alignment

- WIG / TDF

- coverage

- BED

- Genomic intervals

- GTF

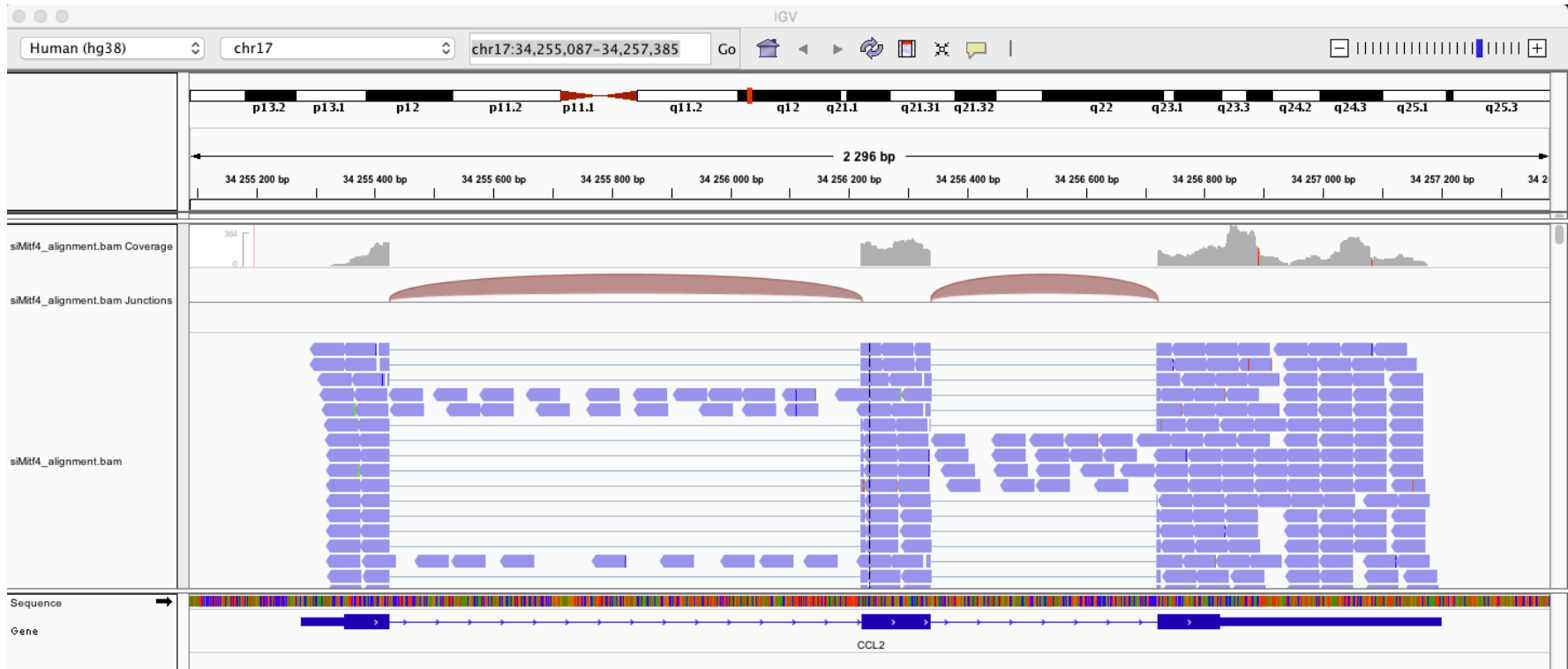
- annotations

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*

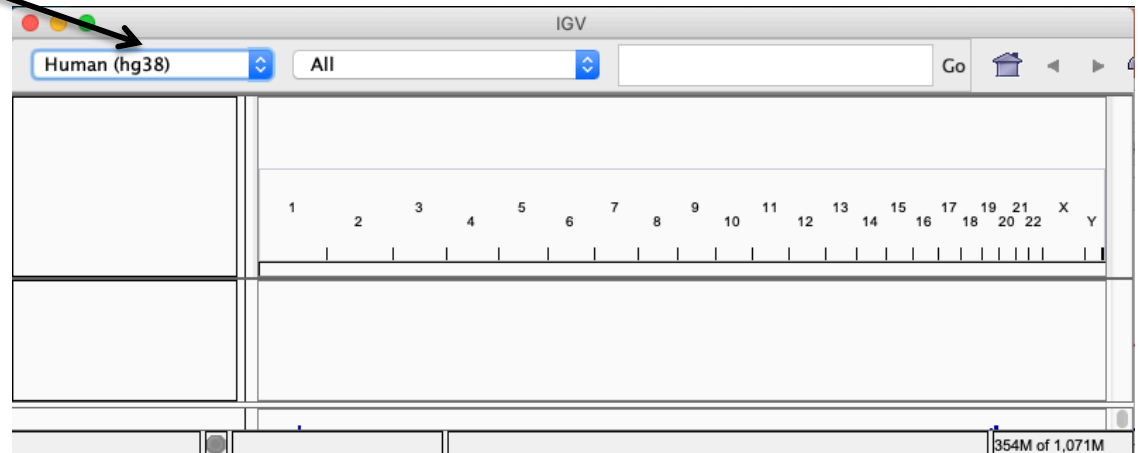
Alignment visualization

- 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

The screenshot displays the IGV 2.3.81 interface. At the top is a menu bar with options: File, Genomes, View, Tracks, Regions, Tools, GenomeSpace, and Help. Below the menu is a tool bar containing a dropdown menu set to 'Human (hg38)', a dropdown for 'chr17', a text input field with 'chr17:34,255,130-34,257,331', a 'Go' button, and navigation icons (home, left, right, refresh, zoom). The main view is divided into several tracks:

- Chromosome ideogram:** Shows the human genome with chromosome 17 highlighted in red. A zoomed-in view of a 2,177 bp region is shown below, with coordinates from 34,255,200 bp to 34,257,200 bp.
- Data tracks:** Includes 'siMitf4_alignment.bam Coverage' (a histogram showing read depth) and 'siMitf4_alignment.bam Junctions' (a track showing structural variant junctions with blue arcs).
- Annotation tracks:** Shows the 'CCL2' gene structure with exons as blue boxes and introns as lines with arrows indicating the direction of transcription.

Menu

Tool bar

Chromosome ideogram

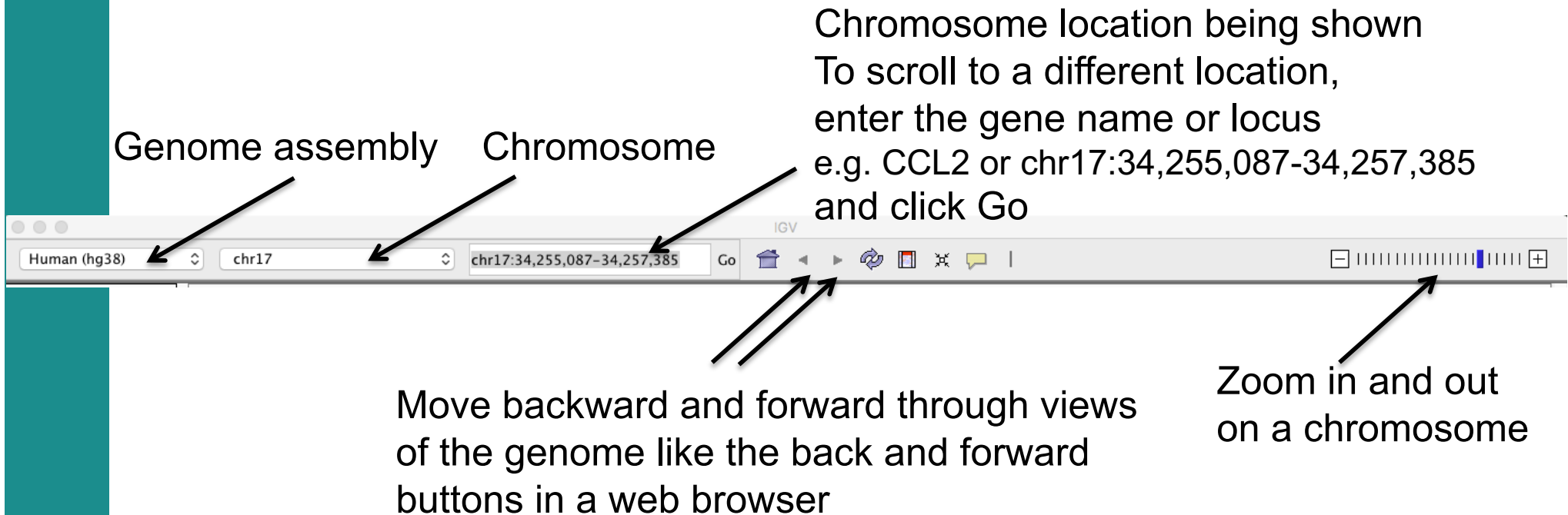
Data tracks

Annotation tracks

IGV menu : main features

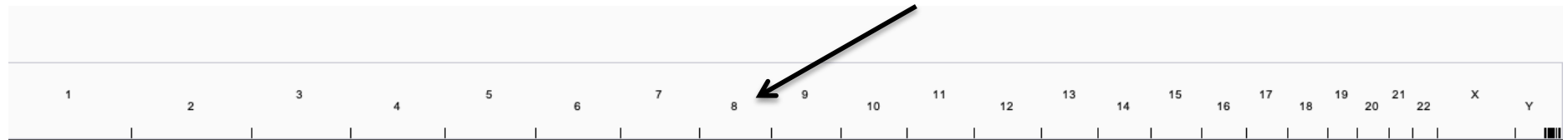
- File
 - Load files into IGV
 - Manage sessions (e.g. save your current settings to a named session file)
 - Save an image
- Genome
 - Manage genomes available on IGV data server (<http://software.broadinstitute.org/software/igv/Genomes>)
 - Create new genomes (required : FASTA file, optional : annotation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

IGV tool bar : main features



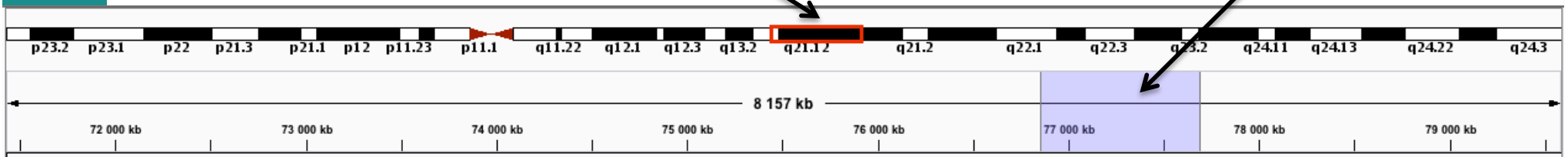
IGV : chromosome ideogram

Click on a chromosome number to jump to this chromosome



Chromosome location being shown

Click and drag to define a region to zoom in

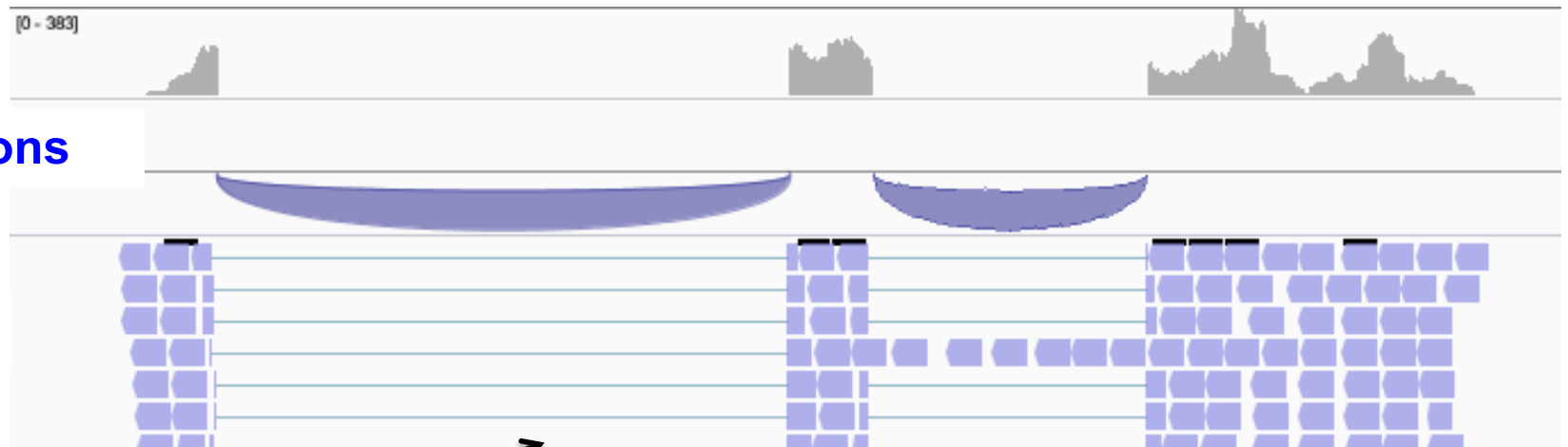


IGV : Data track

Coverage

Splice junctions

Reads



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

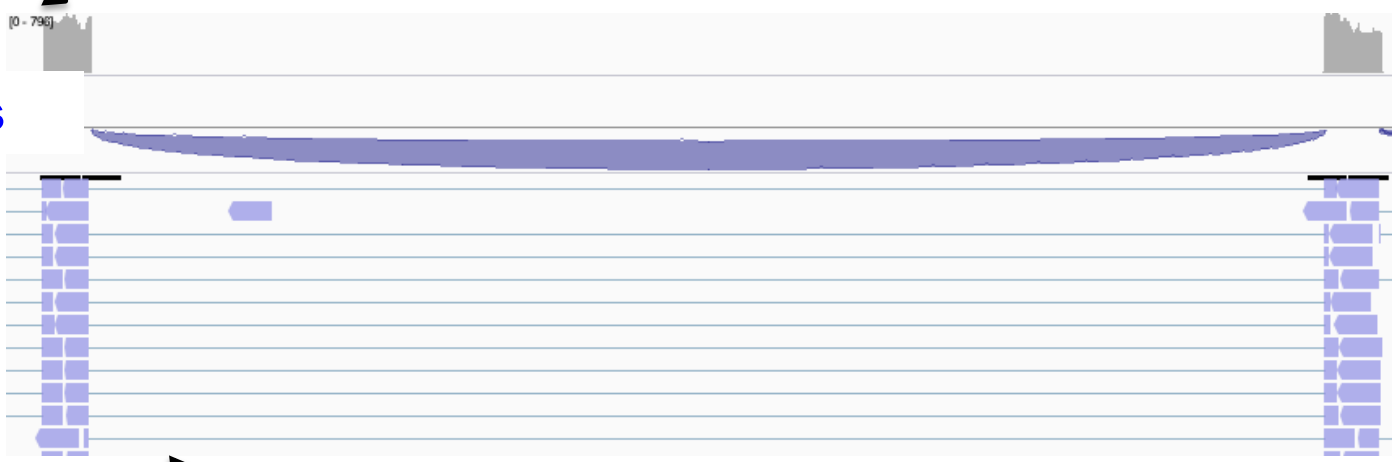
IGV : Data track

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

A context menu is open over the 'siMitf4_alignment.bam' track. The menu items are:

- Rename Track...
- Copy read details to clipboard
- Group alignments by
- Sort alignments by
- Color alignments by** (highlighted)
- Re-pack alignments
- ✓ Shade base by quality
- ✓ Show mismatched bases
- Show all bases
- View as pairs
- Go to mate
- View mate region in split screen
- Set insert size options ...

The 'Color alignments by' sub-menu is open, showing the following options:

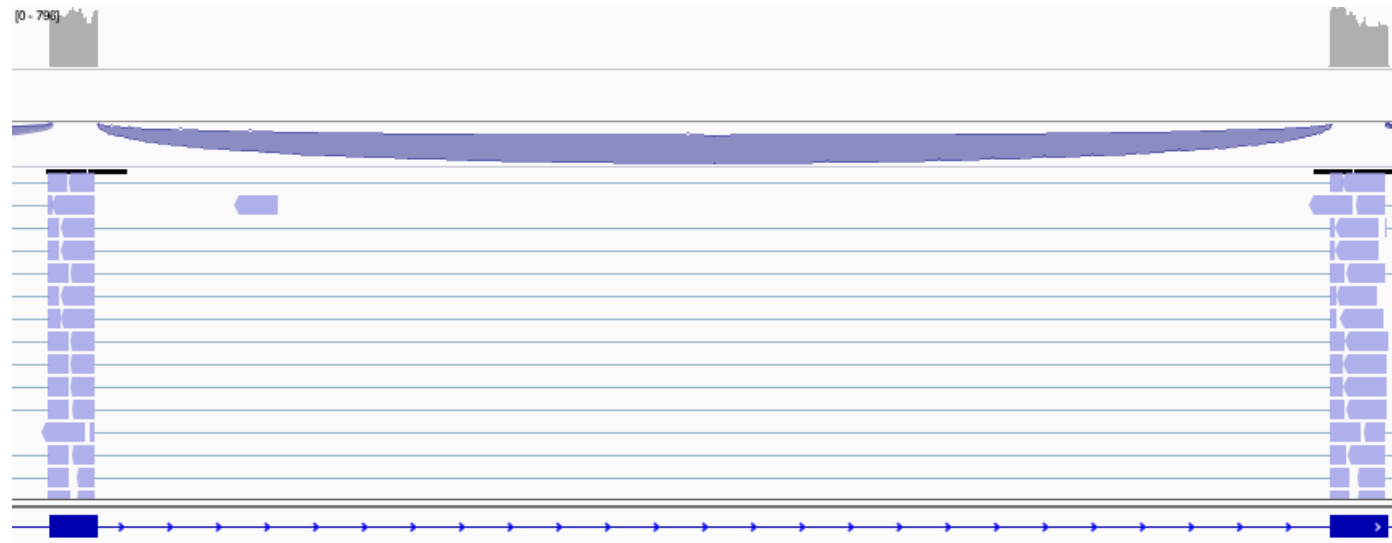
- no color
- ✓ read strand
- read group
- sample
- library
- tag
- bisulfite mode ▶

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

[0 - 252]

Splice junctions

Reads

```
chr17:34,256,279
-----
Total count: 217
A : 0
C : 1 (0%, 0+, 1-)
G : 0
T : 216 (100%, 0+, 216-)
N : 0
-----
```

```
chr17:34255425-34256221
Strand: -
Depth = 194, Flanking Widths: (47,47)
```

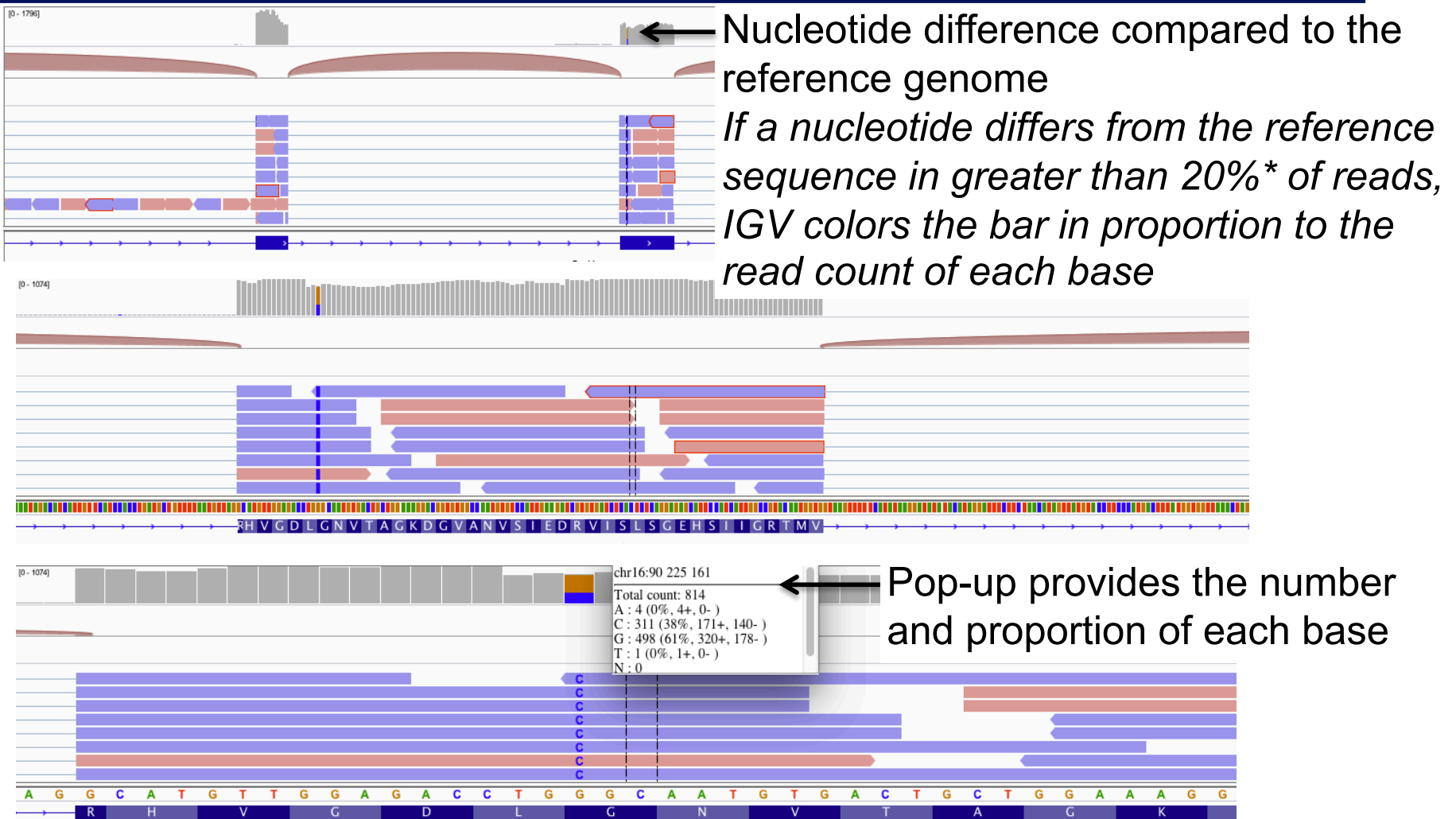
```
Read name = HWI-ST1136:225:HS140:8:1208:10751:43645
Read length = 50bp
-----
Mapping = Primary @ MAPQ 255
Reference span = chr17:34,255,406-34,256,251 (-) = 846bp
Cigar = 20M796N30M
Clipping = None
-----
NH = 1
HI = 1
nM = 0
AS = 50
```

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

IGV data track differences vs reference genome

Zoom in

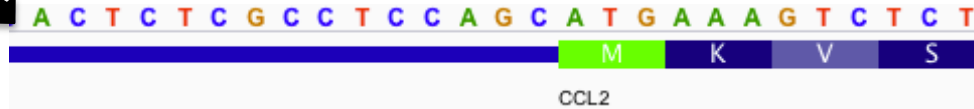
Zoom in



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

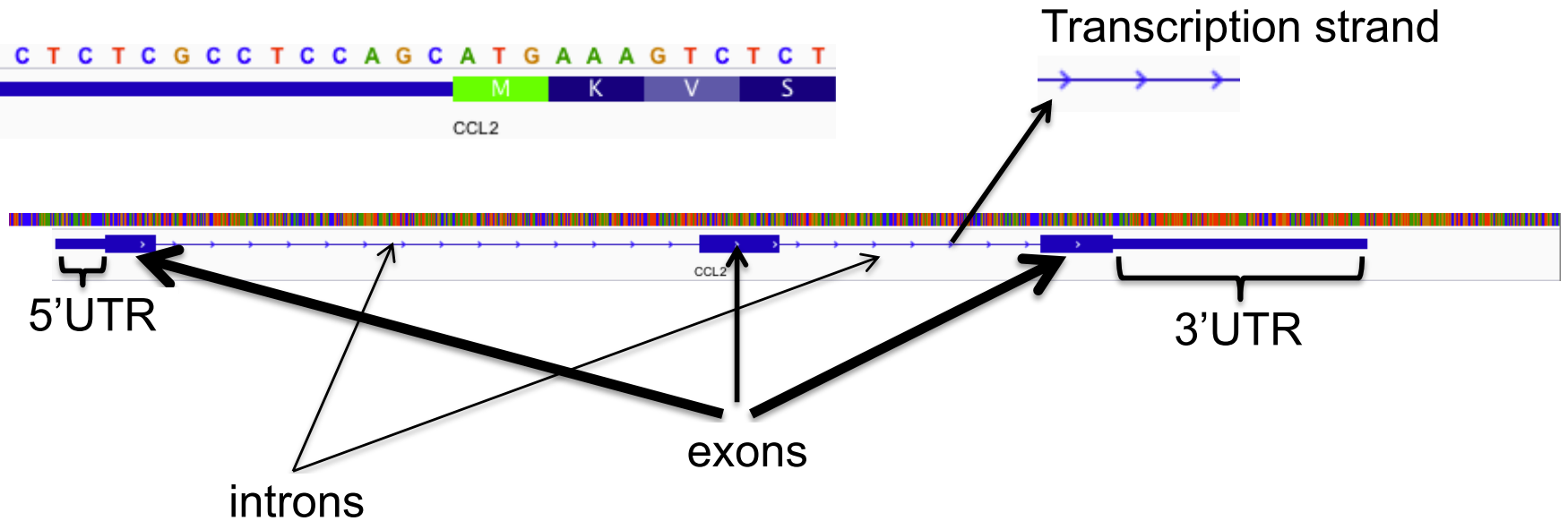
IGV annotation track

Zoom in



Sequence

Annotation



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

CCL2
chr17:34255277-34257201
id = NM_002982

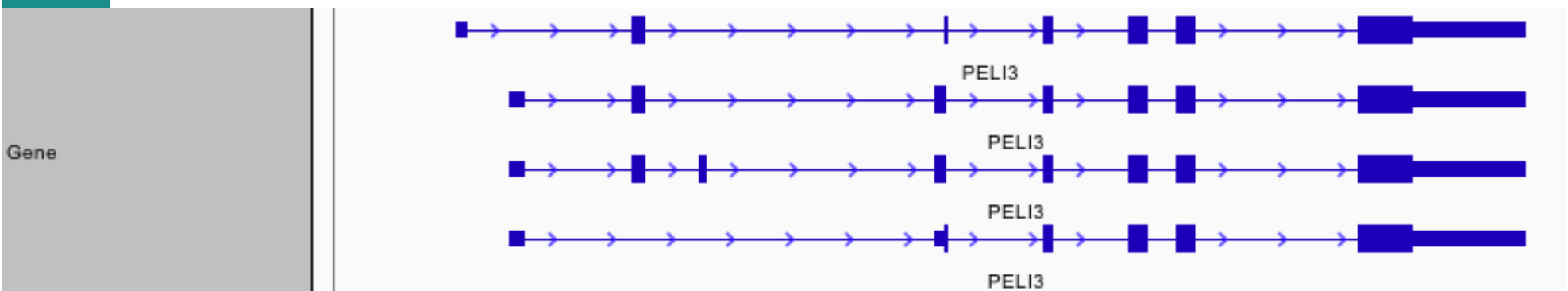
Exon number: 2
Amino acid coding number: 51
chr17:34256222-34256339

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

<u>15: RNA STAR on data</u> <u>6: mapped.bam</u>	  	→ Reads alignment
<u>14: RNA STAR on data</u> <u>6: splice junctions.bed</u>	  	→ Splice junction
<u>13: RNA STAR on data</u> <u>6: log</u>	  	→ 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
 - Shared Data → Data Libraries → NGS data analysis training
 - RNAseq → alignment
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)
 - Start a new IGV session (File → new session)
 - Verify that “Normalize coverage data” is selected (in View → Preferences → Tracks tab)
 - 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

- Verify that “Show junction track” is checked
- In Splice junction track section choose Minimum junction coverage: 10

Quit and open again IGV, 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 chr4:6707960-6707961 ?
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_alignment.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 available on GalaxEast

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

NGS: RSeQC

Inner Distance calculate the inner distance (or insert size) between two paired RNA reads

Read Duplication determines reads duplication rate with sequence-based and mapping-based strategies

Infer Experiment speculates how RNA-seq were configured

Gene Body Coverage (BAM) Read coverage over gene body.

Read NVC to check the nucleotide composition bias

Read Quality determines Phred quality score

Read Distribution calculates how mapped reads were distributed over genome feature

Read GC determines GC% and read count

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

- Launch **RSeQC read distribution** on the mapping results from siLuc2 sample
 - Alignment file
 - Shared Data → Data Libraries → NGS data analysis training → RNAseq → alignment → STAR on siLuc2 : mapped.bam
 - Annotations
 - Shared Data → Data Libraries → NGS data analysis training → RNAseq → annotation_files → Homo_sapiens.GRCh38.95_UCSC_chr.bed

Exercise 3 – Question 1




Read Distribution calculates how mapped reads were distributed over genome feature (Galaxy Version 2.4galaxy1) Options

input bam/sam file


   23: STAR on siLuc2: mapped.bam ▼

(--input-file)

reference gene model

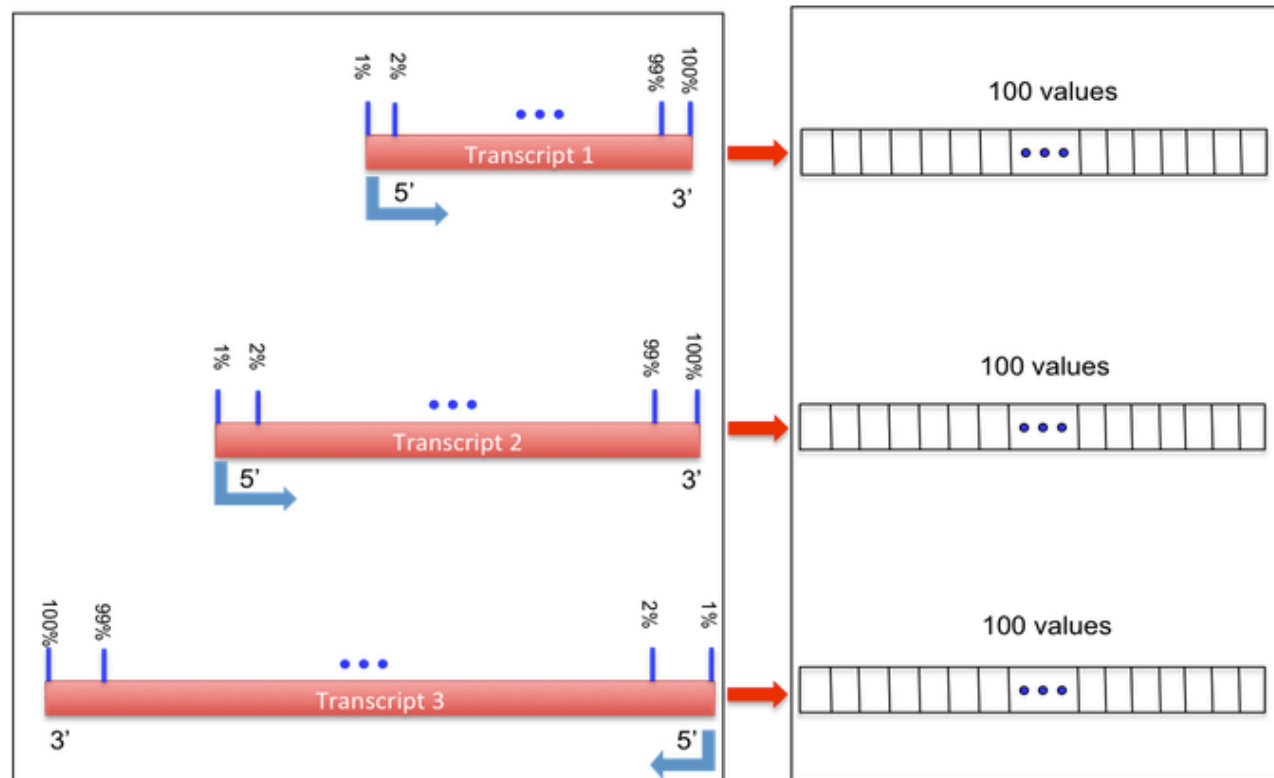
   25: Homo_sapiens.GRCh38.95_UCSC_chr.bed ▼

(--refgene)



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

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 2.4galaxy1) Options

Input .bam File

23: STAR on siLuc2: mapped.bam

(--input-file)

Additional input .bam files

1: Additional input .bam files

Additional input .bam file

27: STAR on siLuc3: mapped.bam

2: Additional input .bam files

Additional input .bam file

28: STAR on siMitf3: mapped.bam

3: Additional input .bam files

Additional input .bam file

29: STAR on siMitf4: mapped.bam

+ Insert Additional input .bam files

reference gene model

25: Homo_sapiens.GRCh38.95_UCSC_chr.bed

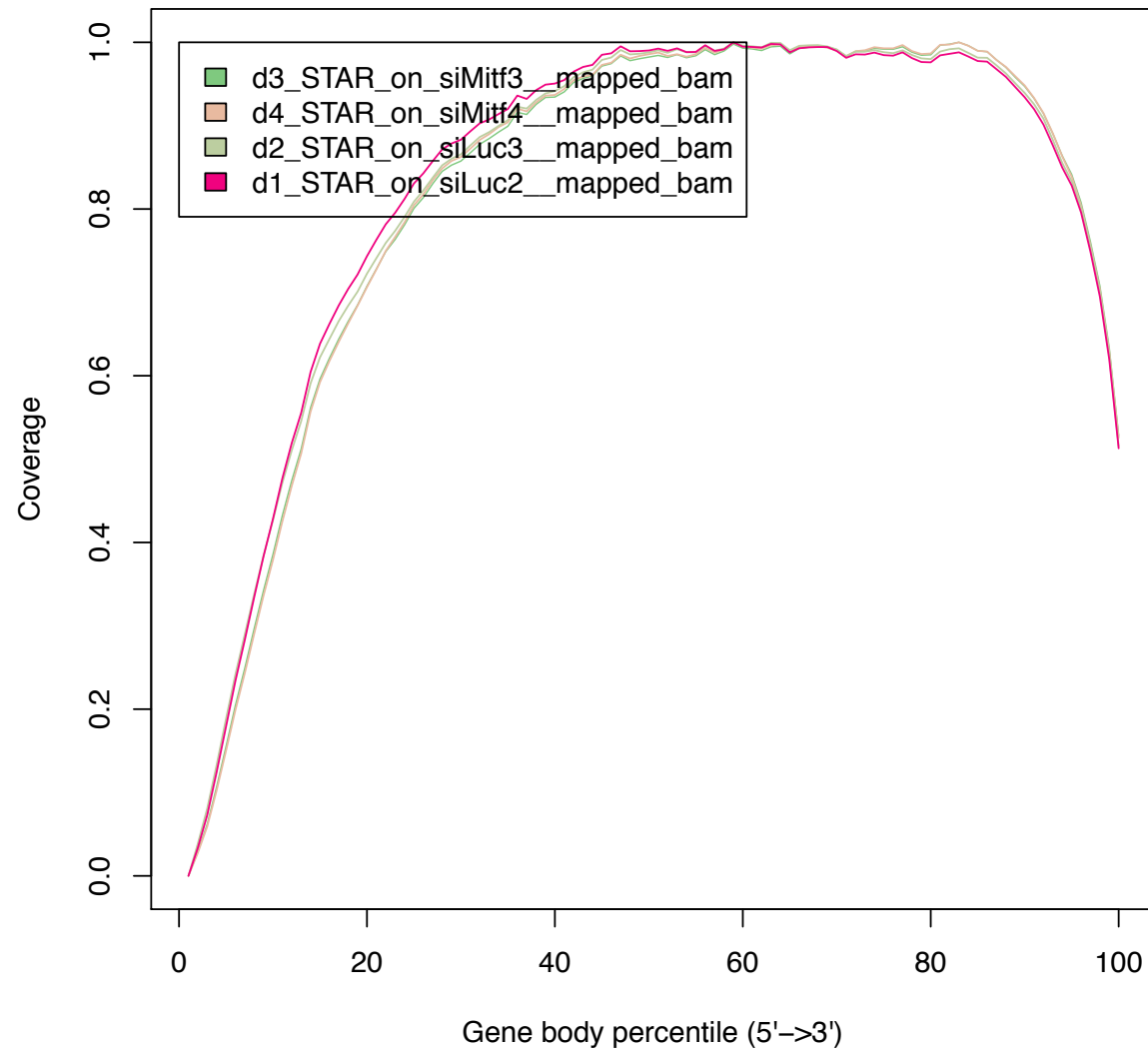
(--refgene)

Minimum mRNA length in bp (default: 100)

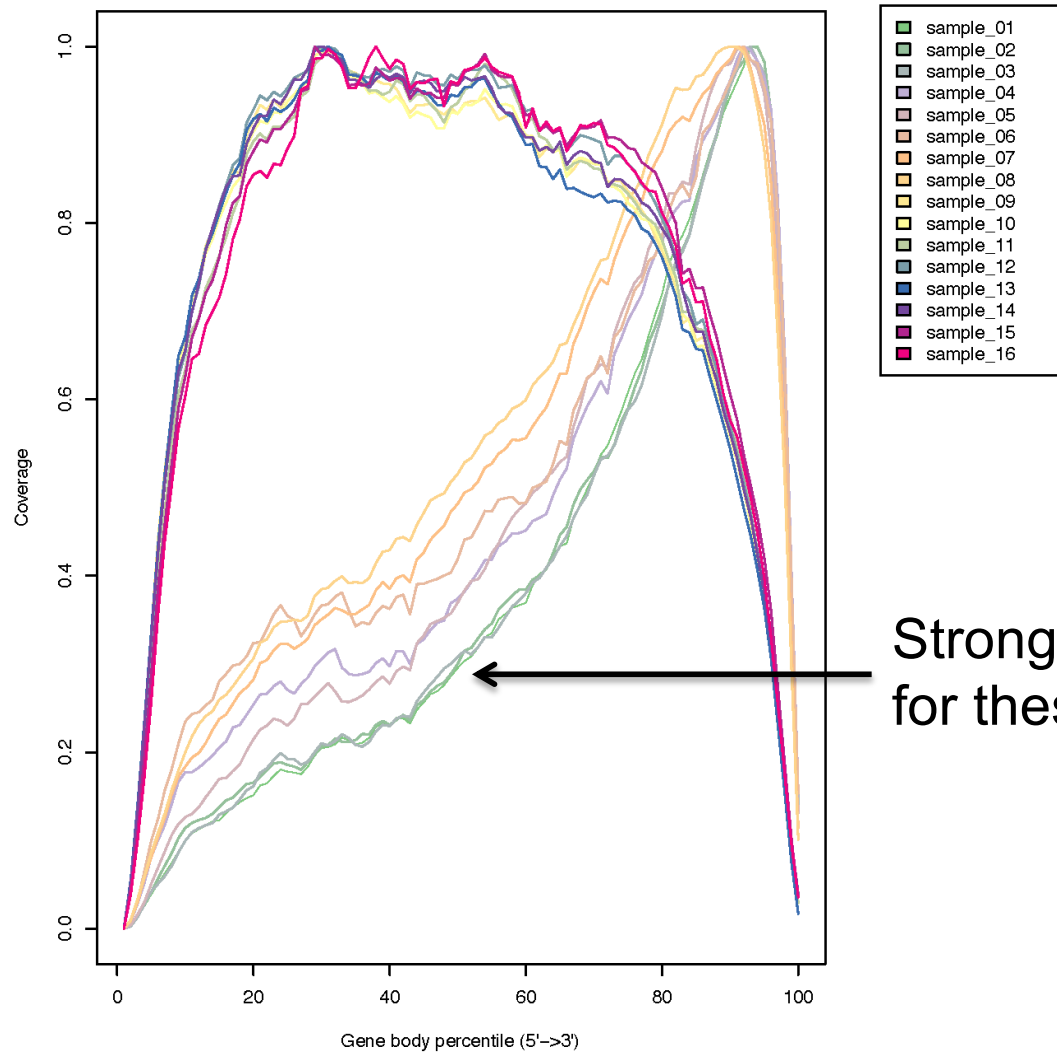
100

mRNA that are shorter than this value will be skipped (--minimum_length).

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 **RSeQC infer experiment** on the mapping results obtained on siLuc2 data from the two different protocols and compare the two results
 - Alignment files
 - Shared Data → Data Libraries → NGS data analysis training → RNAseq → alignment → STAR on siLuc2 : mapped.bam
 - Shared Data → Data Libraries → NGS data analysis training → RNAseq → other_protocol → siLuc2_other_protocol_alignment.bam
 - Annotations
 - Shared Data → Data Libraries → NGS data analysis training → RNAseq → annotation_files → Homo_sapiens.GRCh38.95_UCSC_chr.bed

Exercise 3 – Question 2

- RSeQC infer experiment on siLuc2 mapping results :

Infer Experiment speculates how RNA-seq were configured (Galaxy Version 2.4galaxy1) Options

Input BAM/SAM file
23: STAR on siLuc2: mapped.bam
(--input-file)

Reference gene model in bed format
25: Homo_sapiens.GRCh38.95_UCSC_chr.bed
(--refgene)

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

Minimum mapping quality (default=30)
30

Minimum phred scale mapping quality to consider a read 'uniquely mapped' (--mapq)




✓ Execute

Exercise 3 – Question 2

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




Infer Experiment speculates how RNA-seq were configured (Galaxy Version 2.4galaxy1) ▼ Options

Input BAM/SAM file

   2: siLuc2_other_protocol_alignment.bam ▼

(--input-file)

Reference gene model in bed format

   3: Homo_sapiens.GRCh38.95_UCSC_chr.bed ▼

(--refgene)

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


200000

(--sample-size)

Minimum mapping quality (default=30)

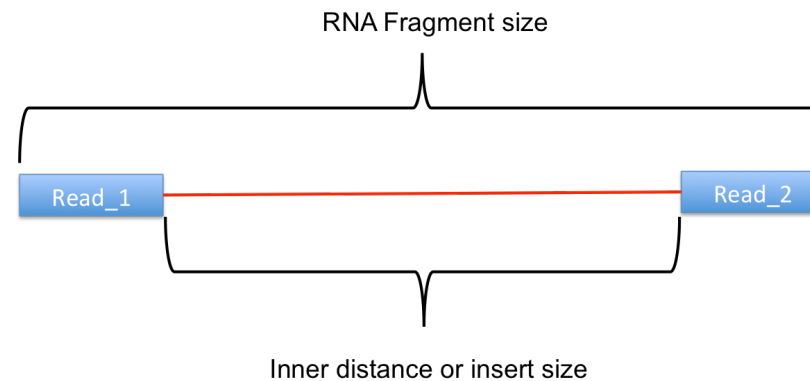
30

Minimum phred scale mapping quality to consider a read 'uniquely mapped' (--mapq)



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

