



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
reads

CACGTACC
CACGT**T**CC

mismatch

CACGTA_CC
CACGT**A**TCC

indels (insertion/deletion)

CACGTACC
CACGT_**_**CC

Challenges of short read mapping

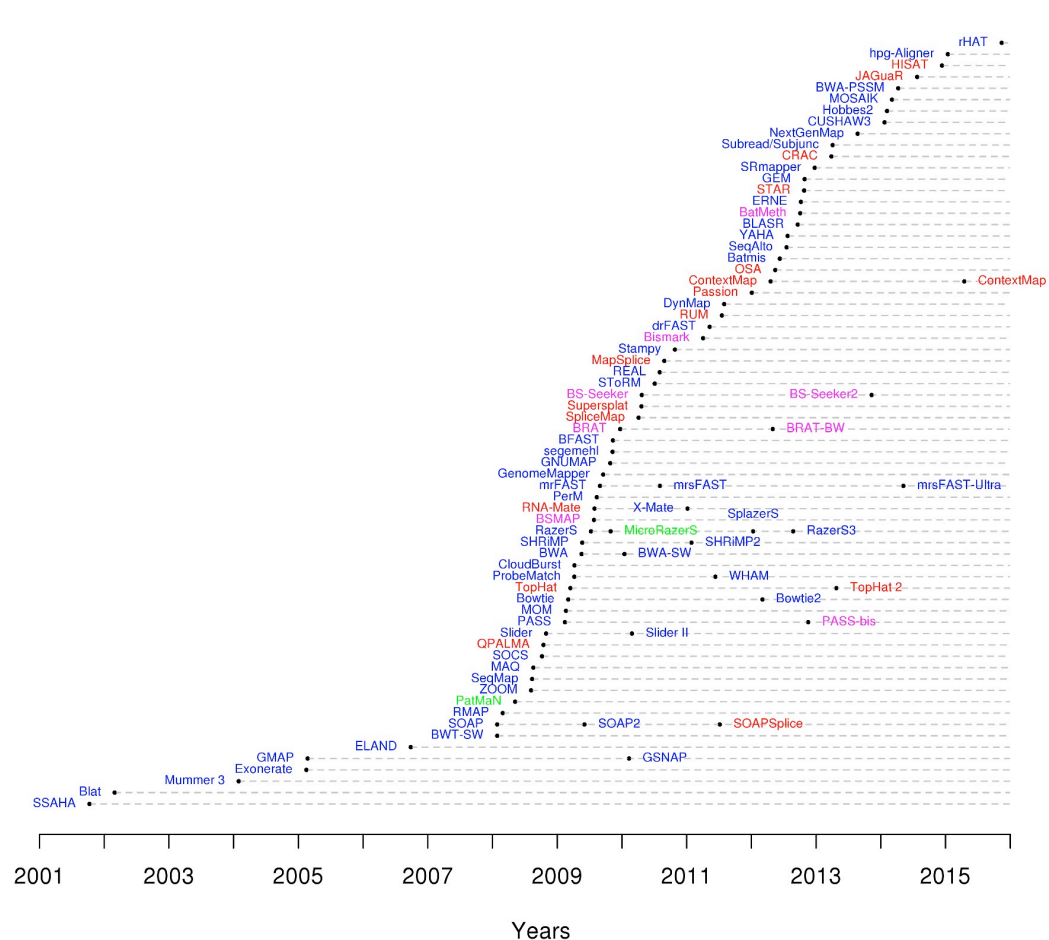
- Reference sequence can be large (~3 Gb for human)
 - Short reads → several, equally likely places in reference sequence from which they could have been read
e.g. repetitive regions
 - The genome from which reads have been generated may be different from the reference genome
→ Need to allow mismatches 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

A lot of tools developed ...

- More than 90 mapping tools



DNA mappers
RNA mappers
miRNA mappers
bisulfite mappers

Two main strategies

■ Indexing

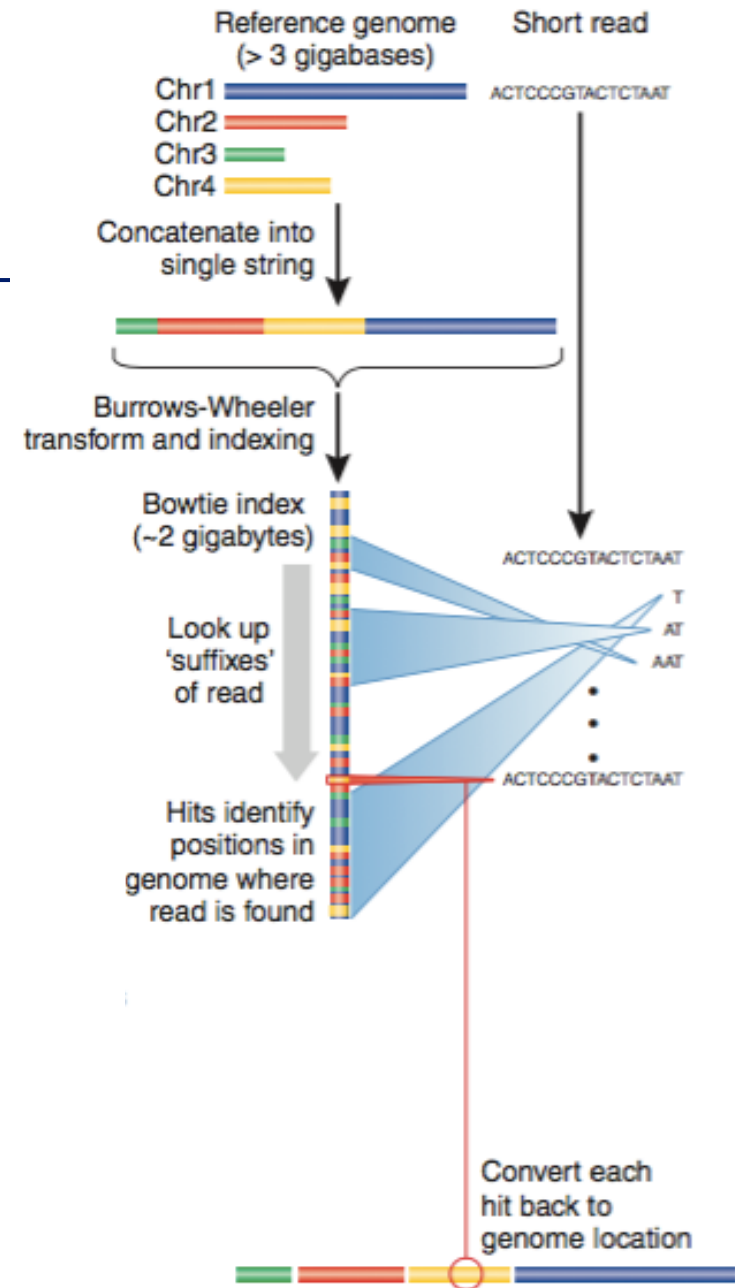
- Like the index at the end of a book
 - ➔ an index of a large DNA sequence allows one to rapidly find shorter sequences embedded within it
- 2 strategies : index the reads or the genome

■ 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 FM 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)

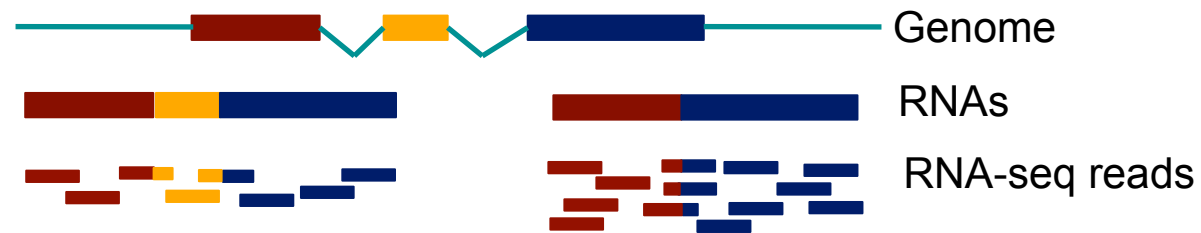
How to choose a mapper ?

- 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

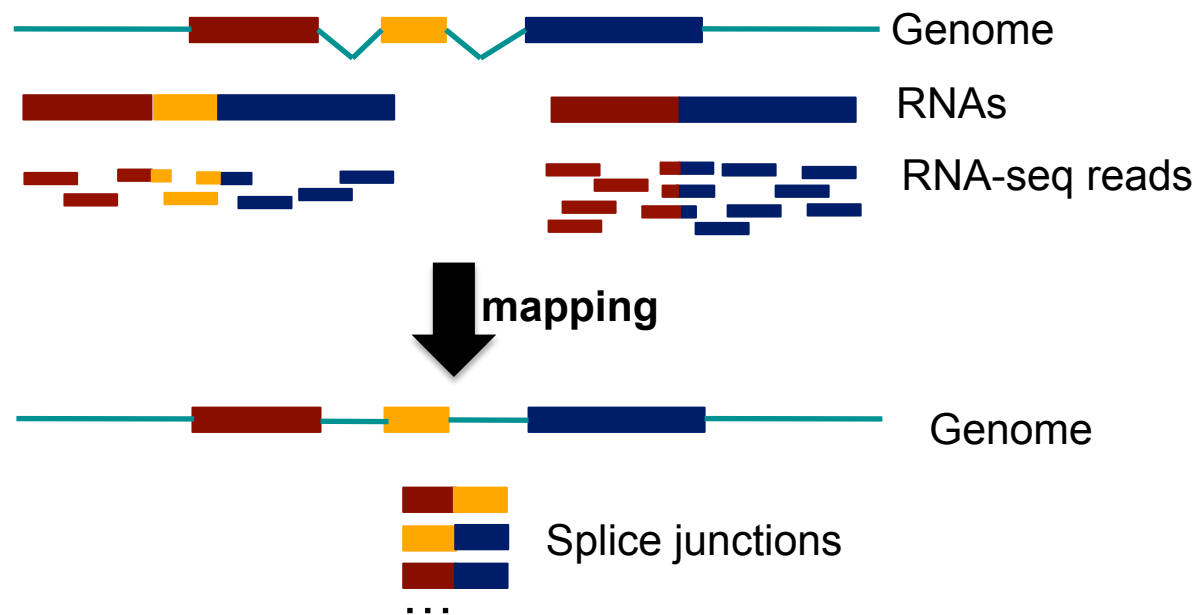
Specificity of RNA-seq reads



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

Map onto the genome and splice junctions ?

■ ERANGE, RNA-Mate

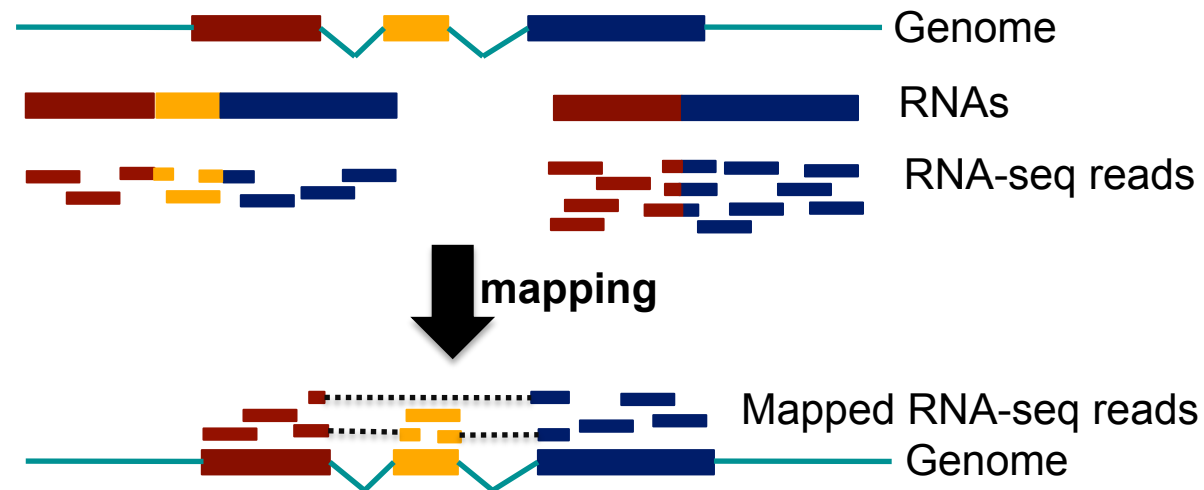


■ But

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

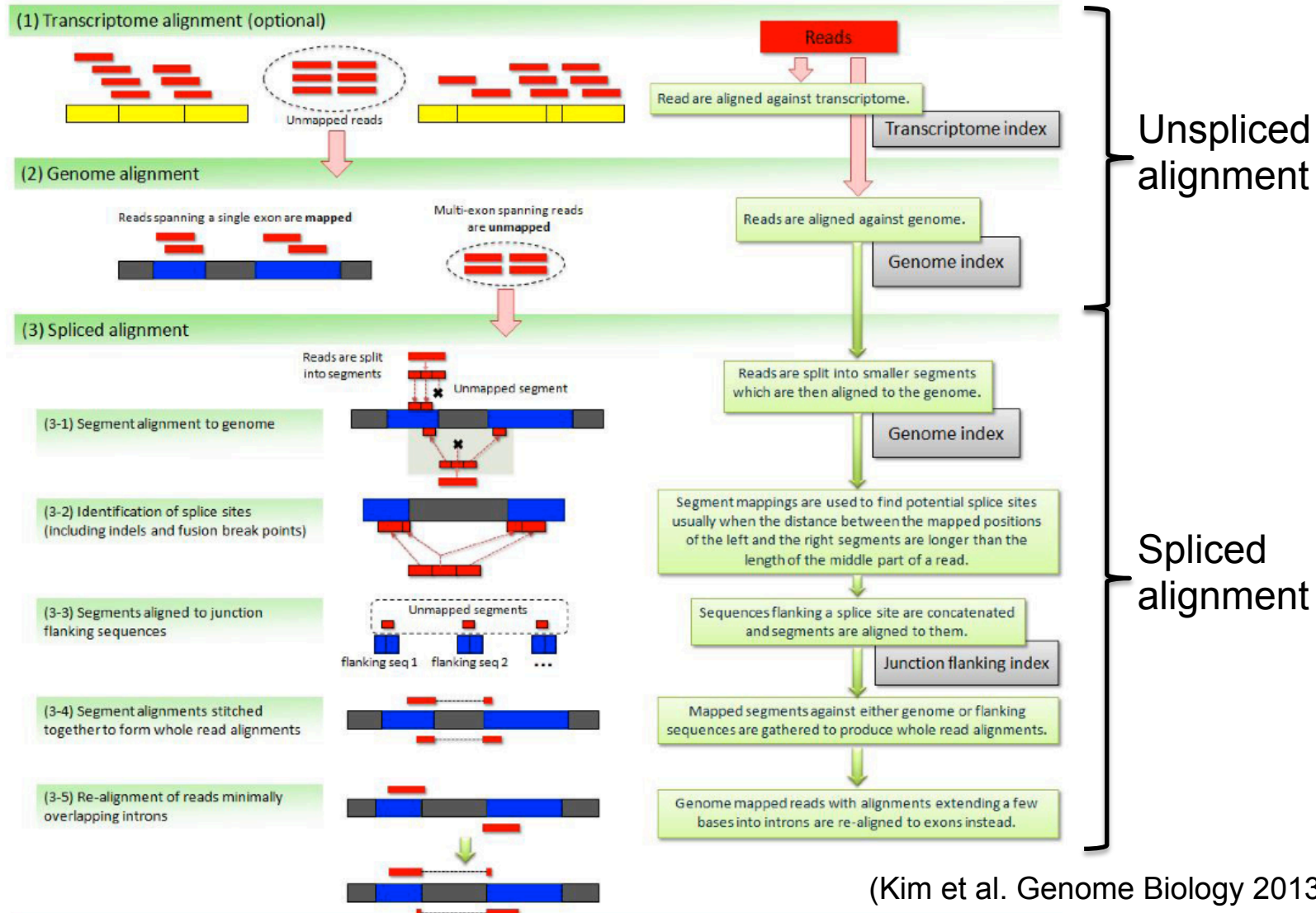
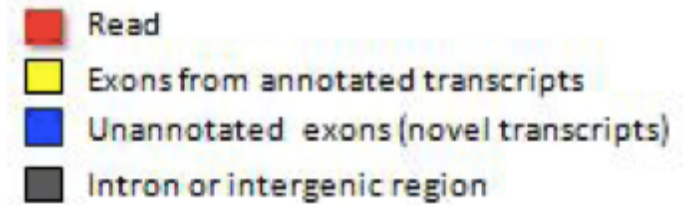
Spliced mapping

- 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)

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
 1. Import the corresponding FASTQ file in your history
 2. Launch Tophat2 on this FASTQ file

Exercise 1

1. Import the 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

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

[Libraries](#) / [CNRS 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
Uploaded by	keime@igbmc.fr
Miscellaneous blurb	150.2 MB
Miscellaneous information	uploaded fastq file

Exercise 1

2. Launch Tophat2

Tophat2 Gapped-read mapper for RNA-seq data (Galaxy Version 0.13) Options

Is this library mate-paired?

Single-end Type of sequencing (single or paired-end)

RNA-Seq FASTQ file

6: siLuc2_1000000.fastq FASTQ file

Nucleotide-space: Must have Sanger-scaled quality values with ASCII offset 33

Use a built in reference genome or own from your history

Use a built-in genome

Built-ins genomes were created using default options

Select a reference genome

hg38 Reference genome (assembly name)

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

Library Type

FR First Strand

--library-type; TopHat will treat the reads as strand options below to select the correct RNA-seq proto

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

2. Launch Tophat2

Use Own Junctions

Yes

Use Gene Annotation Model

Yes

TopHat will use the exon records in this file to build a set of known splice junctions for each gene, and will attempt to align reads to these junctions even if they would not normally be covered by the initial mapping.

Use one of our GFF file

Select a reference annotation

hg38_version_85_ensembl

if your annotation of interest is n

Annotation file

→ Using this file, TopHat will first extract the transcript sequences and use Bowtie to align reads to this virtual transcriptome first.

✓ 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

- **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

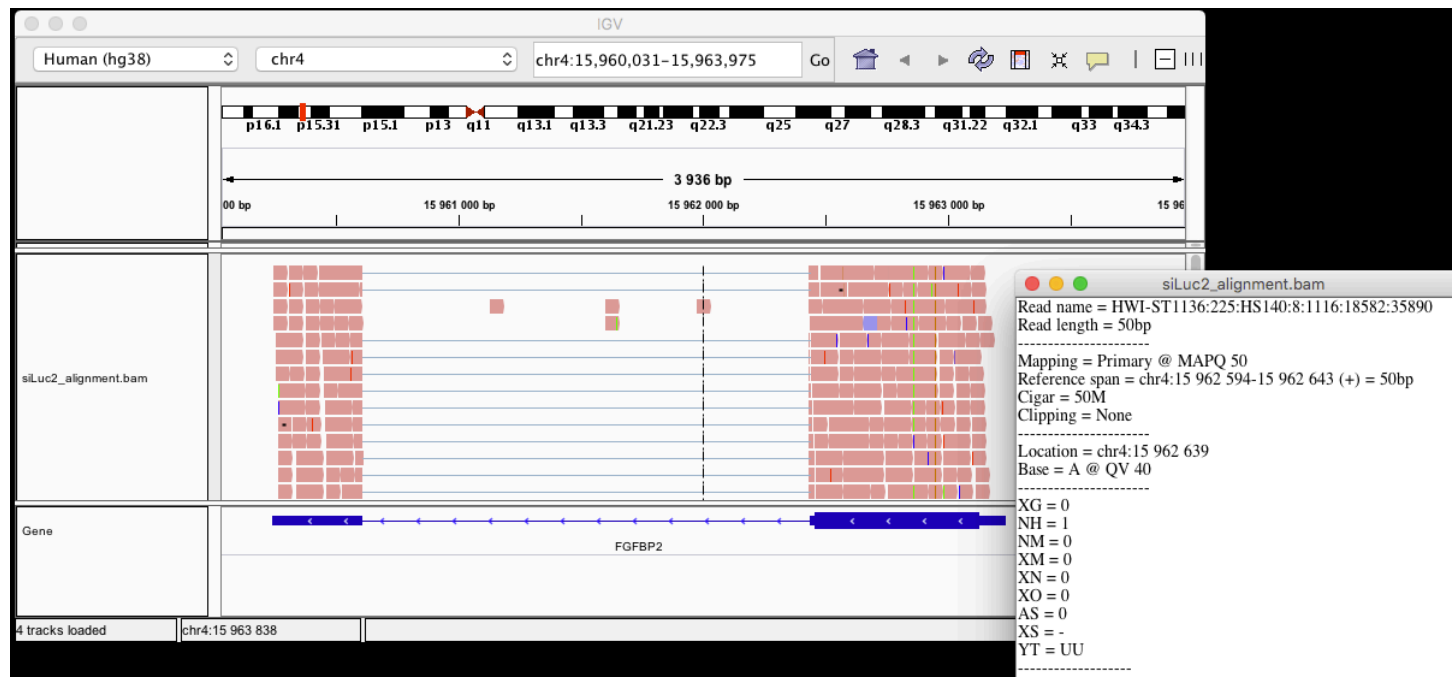
Alignment file format : SAM

■ Additional tags (format tag:type:value)

Tag 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 (uint16.t) <code>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]+))*</code> 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

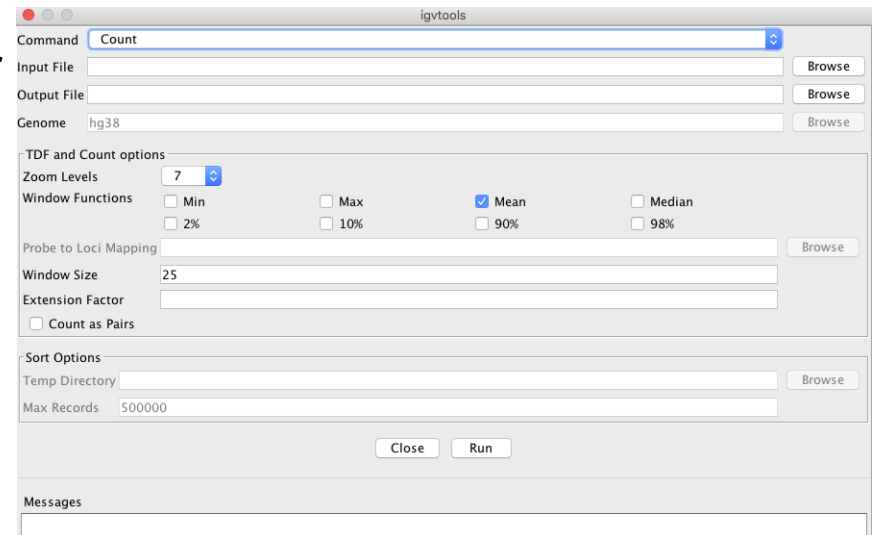
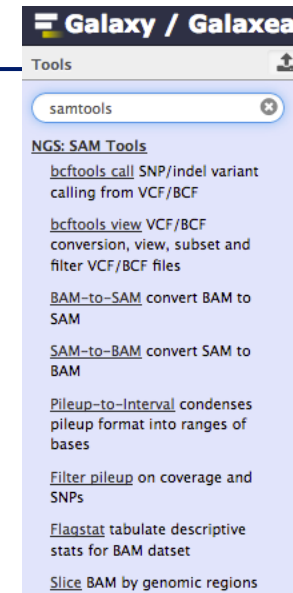
Alignment file format : BAM

- 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



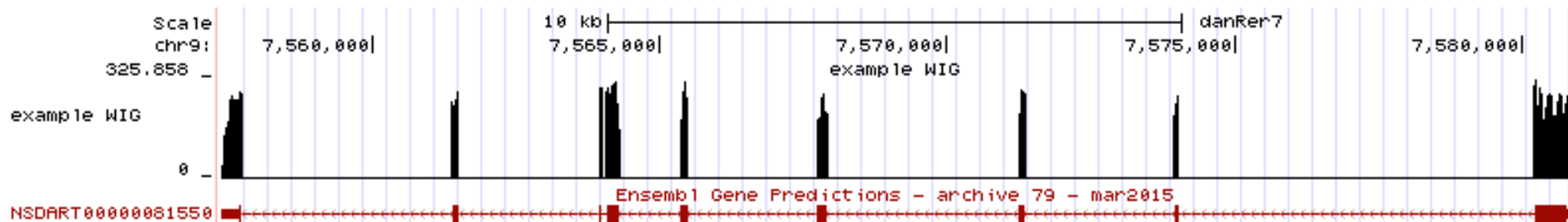
Utilities to manipulate SAM/BAM files

- 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

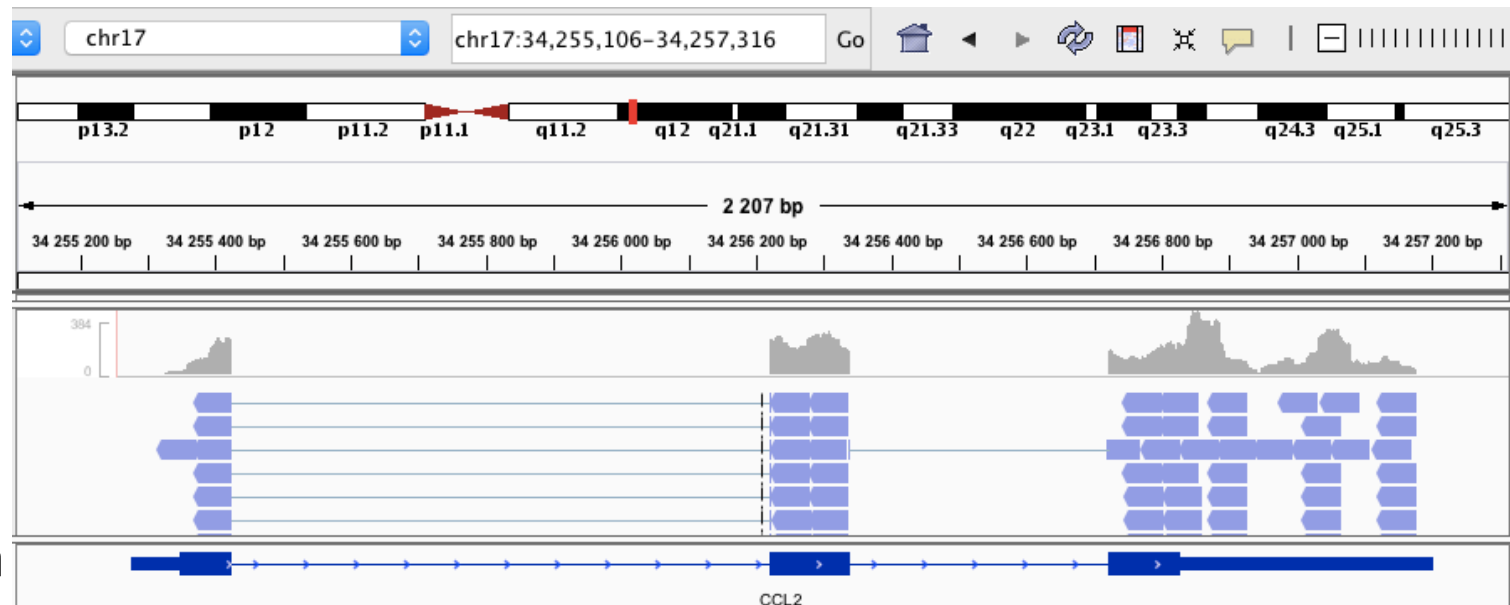
The screenshot displays the igvtools application window. The 'Command' dropdown is set to 'Count'. The 'Input File' is '/Volumes/rufushome/CNRStraining/analyzeddata/RNAseq/alignment/siLuc2_alignment.bam' and the 'Output File' is '/Volumes/rufushome/CNRStraining/analyzeddata/RNAseq/alignment/siLuc2_alignment.bam.tdf'. 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. Under 'Sort Options', 'Temp Directory' is empty and 'Max Records' is 500000. The 'Run' button is visible. The main visualization area shows 'Human (hg38)' with 'chr4' selected, covering the region 'chr4:15,958,524-15,964,999'. A scale bar at the top shows cytobands from p16.1 to q35.1. A 6,454 bp window is highlighted. Below the scale bar are four tracks: 'siLuc2_alignment.bam.tdf', 'siLuc3_alignment.bam.tdf', 'siMif3_alignment.bam.tdf', and 'siMif4_alignment.bam.tdf'. The tracks show read counts as black and green bars. The 'Gene' track at the bottom shows the 'FGFBP2' gene structure.

Coverage vs alignment

Coverage

Alignment

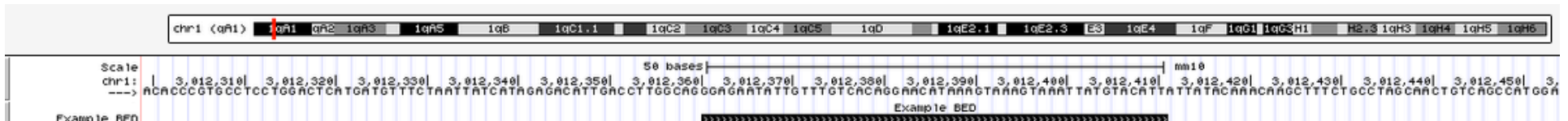
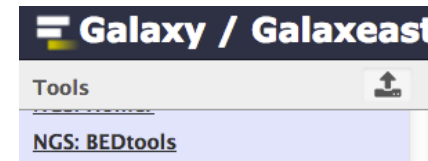
Annotation



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 : <http://code.google.com/p/bedtools/>

required
Most common :
6 columns

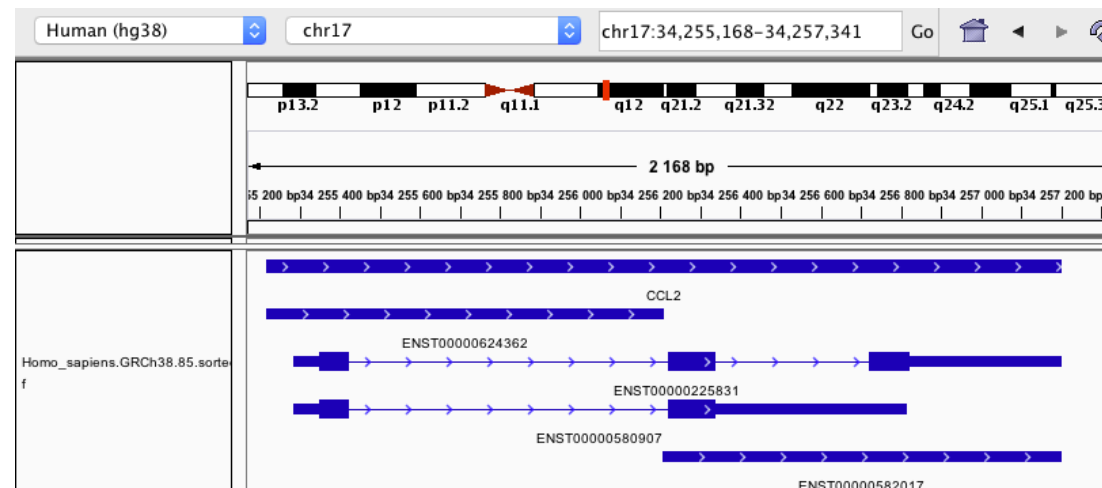


General Feature Format (GFF)

- GFF: General Feature Format
- 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

- Raw data

text

binary

- SAM / BAM

- 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://www.broadinstitute.org/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 1.5x 3x 10x base zoom out 1.5x 3x 10x 100x

chr11:82,035,577-82,037,452 1,876 bp. enter position, gene symbol or search terms go

chr11 (qC) 11qA1 11qA2 11qA3 11qA4 11qA5 11qB1.3 11qB3 11qB5 11qC 11qD 11qE1 11qE2

The screenshot displays a genomic browser interface for a region on mouse chromosome 11. The main visualization area contains several tracks:

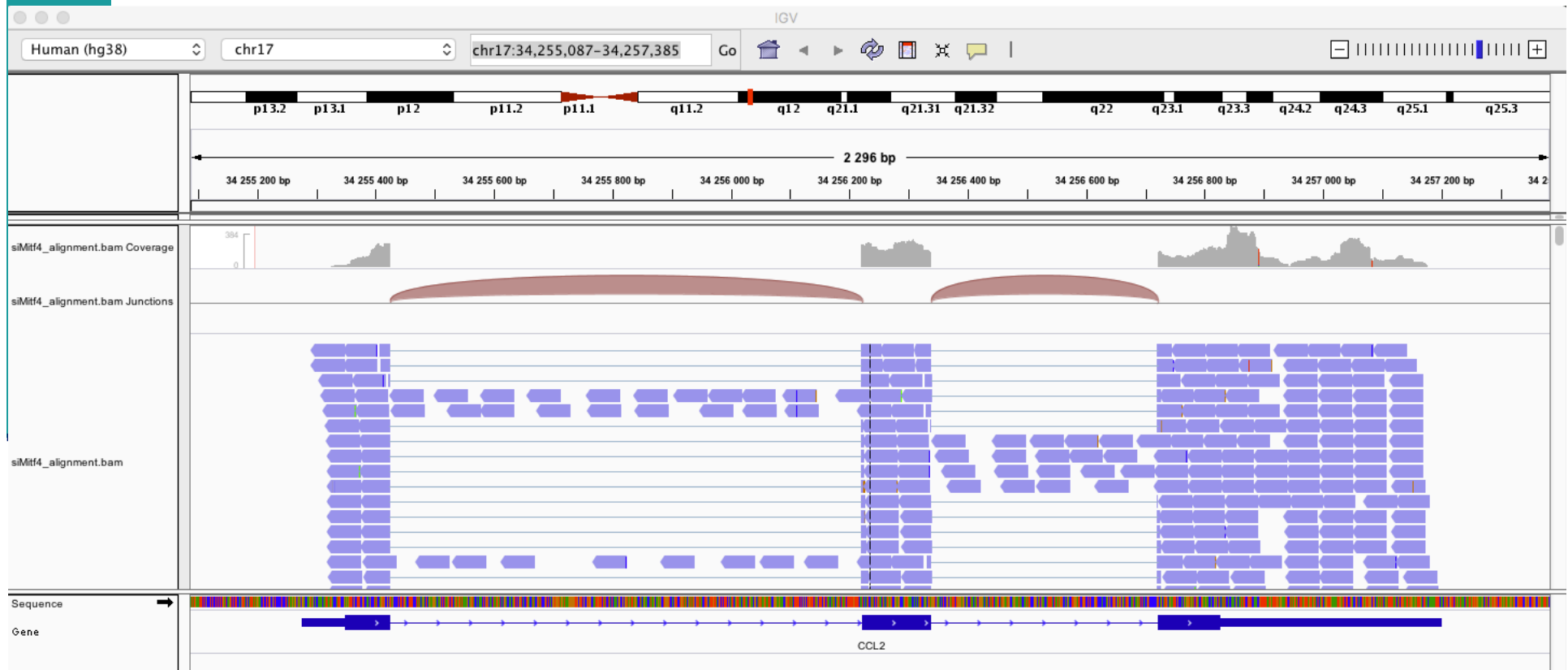
- Scale:** Shows a 500-base scale with coordinates 82,036,000 and 82,037,000.
- Sample 1:** A signal track showing a peak at approximately 82,036,500.
- UCSC Genes (RefSeq, GenBank, tRNAs & Comparative Genomics):** Shows gene models for *Cc12* and *Gm17266*.
- Basic Gene Annotation Set from ENCODE/GENCODE Version M9 (Ensembl 84):** Shows gene models for *Cc12* and *RefSeq Genes*.
- RefSeq Genes:** Shows gene models for *Cc12* and *RefSeq Genes*.
- Retroposed Genes V6, Including Pseudogenes:** Shows gene models for *DQ973566*, *RK153469*, *RK150937*, *RK153520*, *RK151789*, *RK153443*, *RK132590*, *BC145867*, *BC145869*, *BC355970*, *AF065929*, *AF065930*, *AF065931*, *AF065932*, *AF065933*, and *CT018187*.
- Mouse ESTs That Have Been Spliced:** Shows spliced ESTs for the listed genes.
- Common SNPs (142):** Shows single nucleotide polymorphisms found in at least 1% of samples.
- RepeatMasker Viz.:** Shows a detailed visualization of RepeatMasker annotations.
- RepeatMasker:** Shows repeating elements identified 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 < 2.0 > move end < 2.0 >

track search default tracks default order hide all manage custom tracks track hubs configure multi-region reverse resize refresh

IGV (Integrative Genomics Viewer)



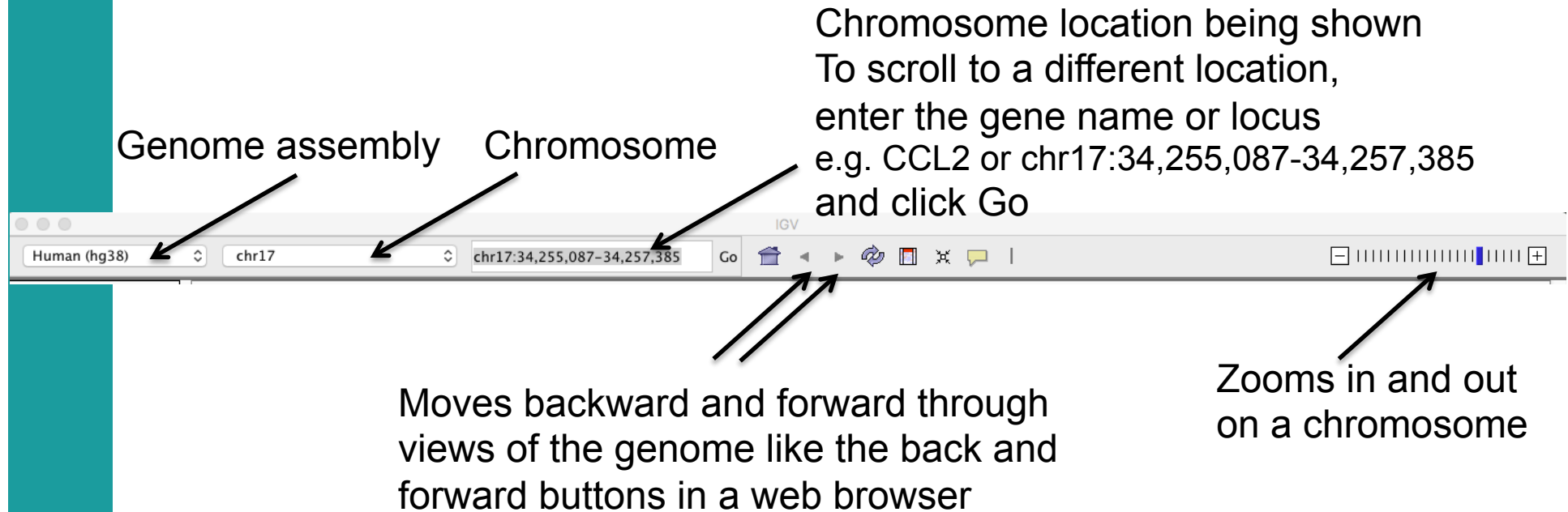
IGV



IGV 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

Chromosome location being shown



Click and drag to define a new region to zoom in

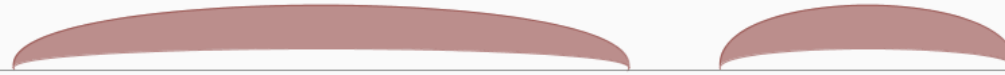


IGV : Data track

Coverage



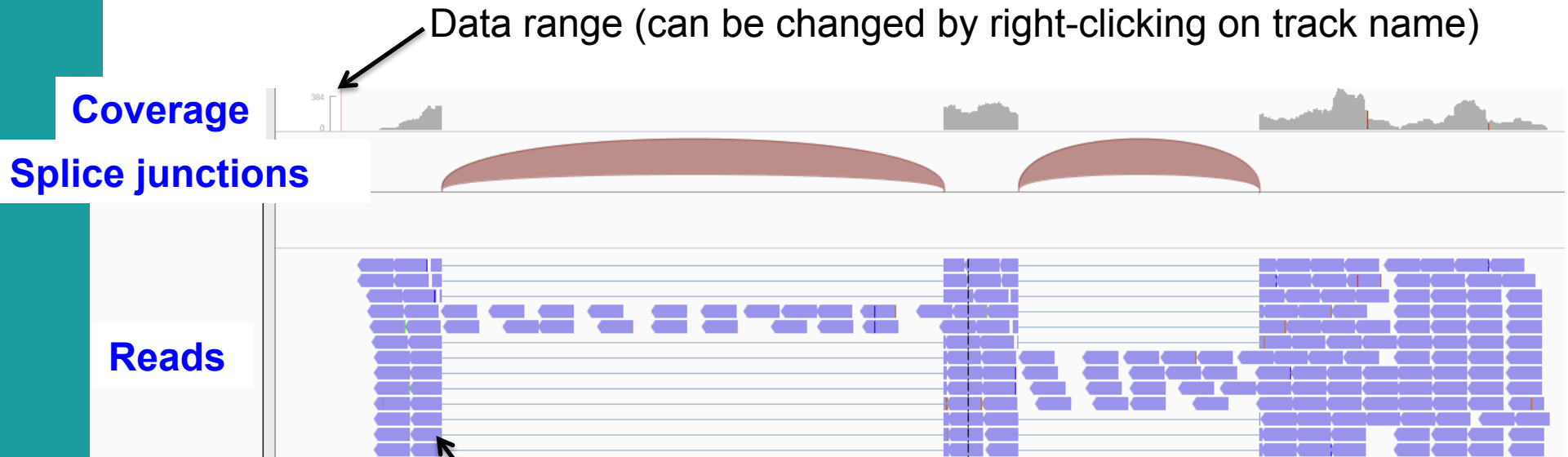
Splice junctions



Reads



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

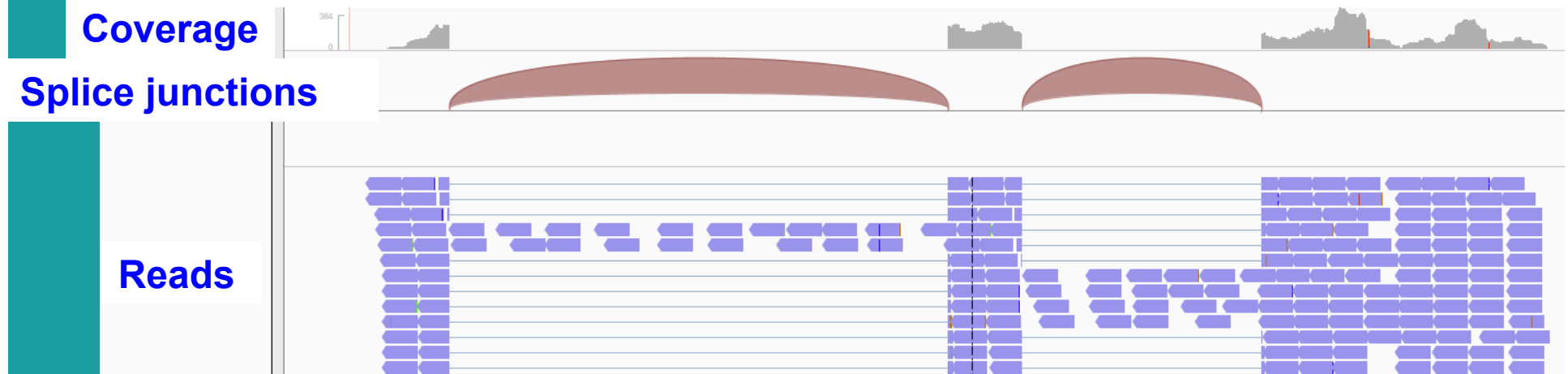
The image shows a context menu for the track 'siMitf4_alignment.bam'. The menu items are:

- 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 ...

On the right, a sub-menu for 'Color alignments by' is open, showing options:

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

IGV : Data track



- Display of splice junctions

- Strand

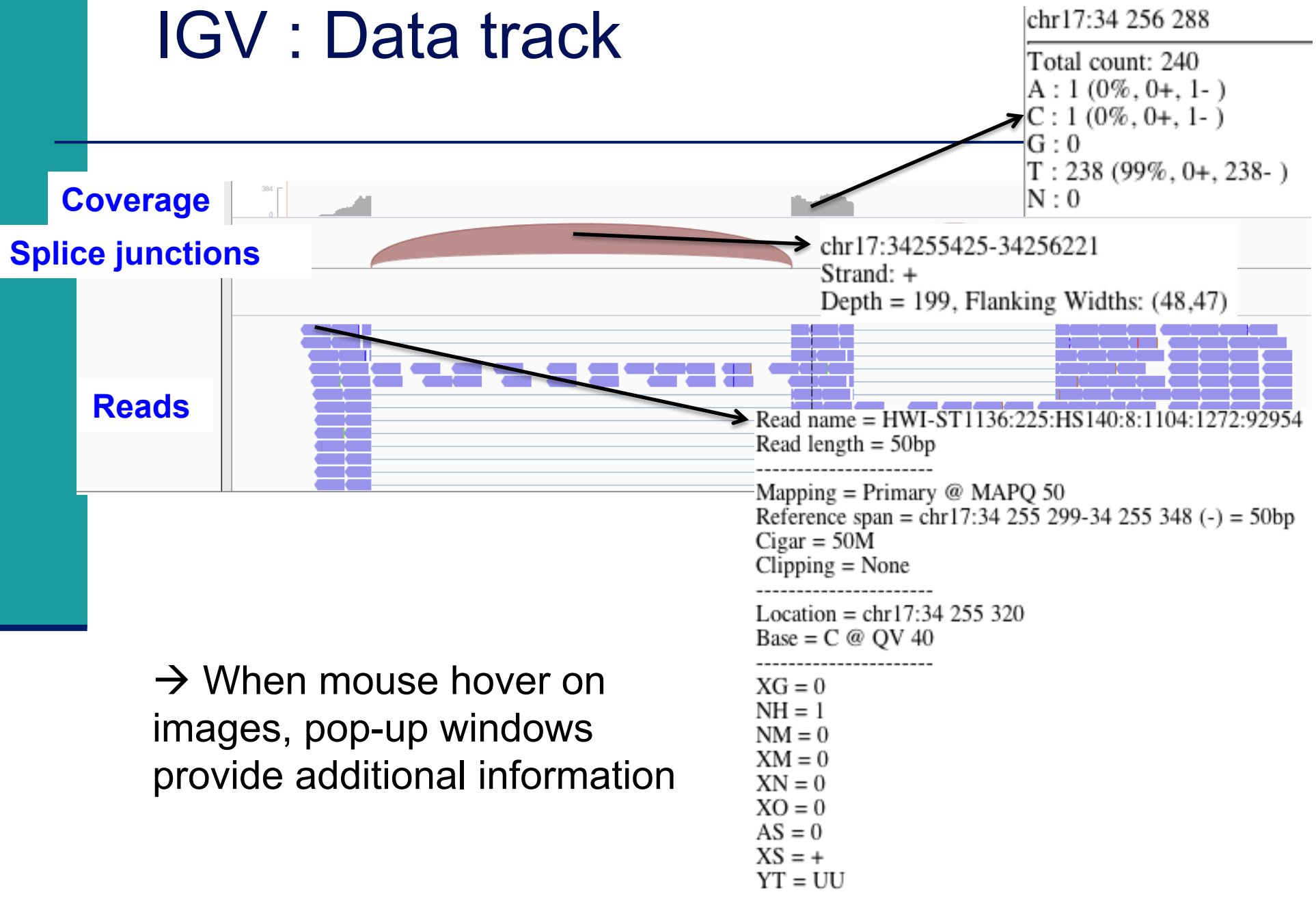
- Blue junctions : + strand
 - Red junctions : - strand

- 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

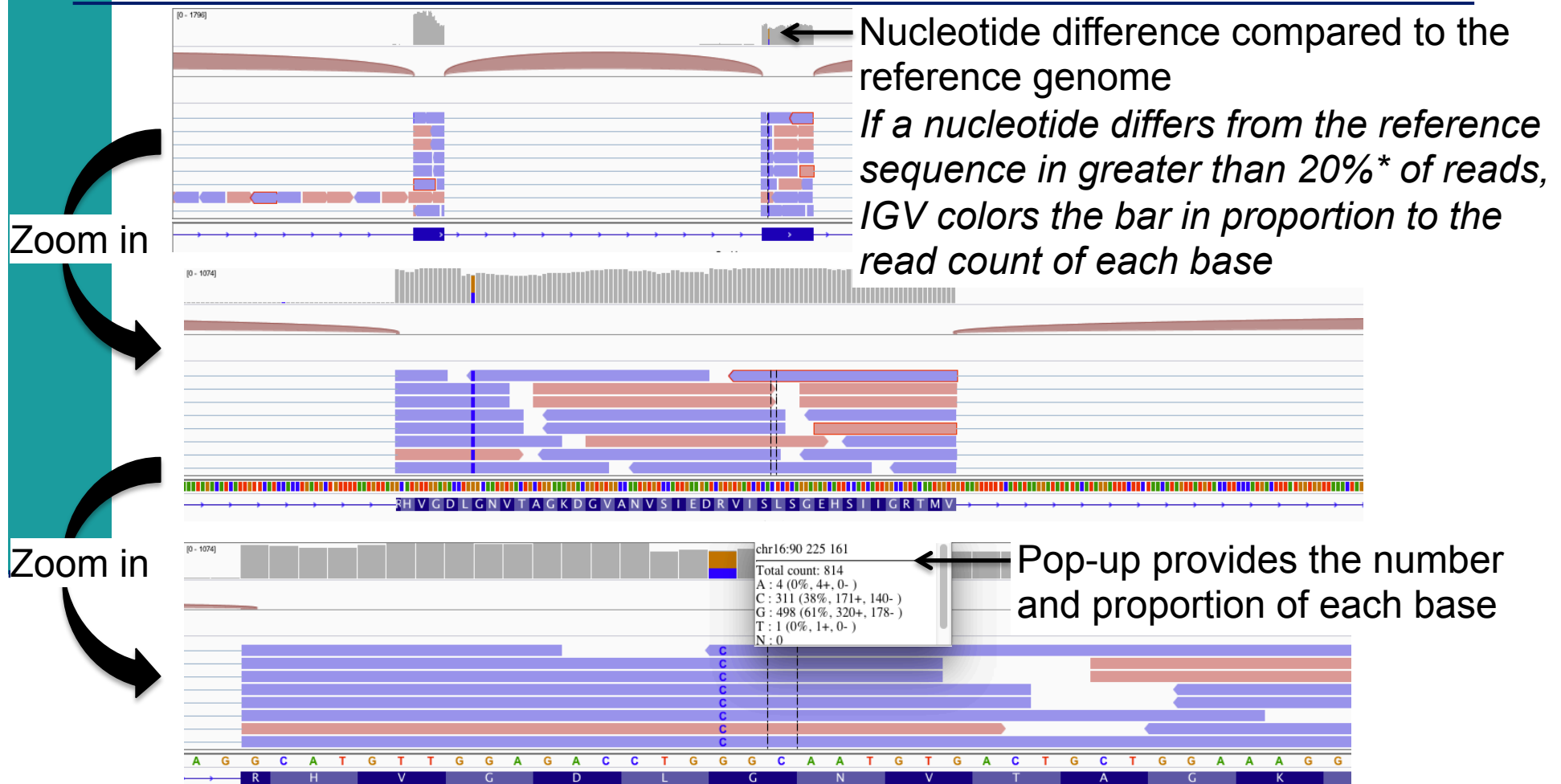


IGV : Data track



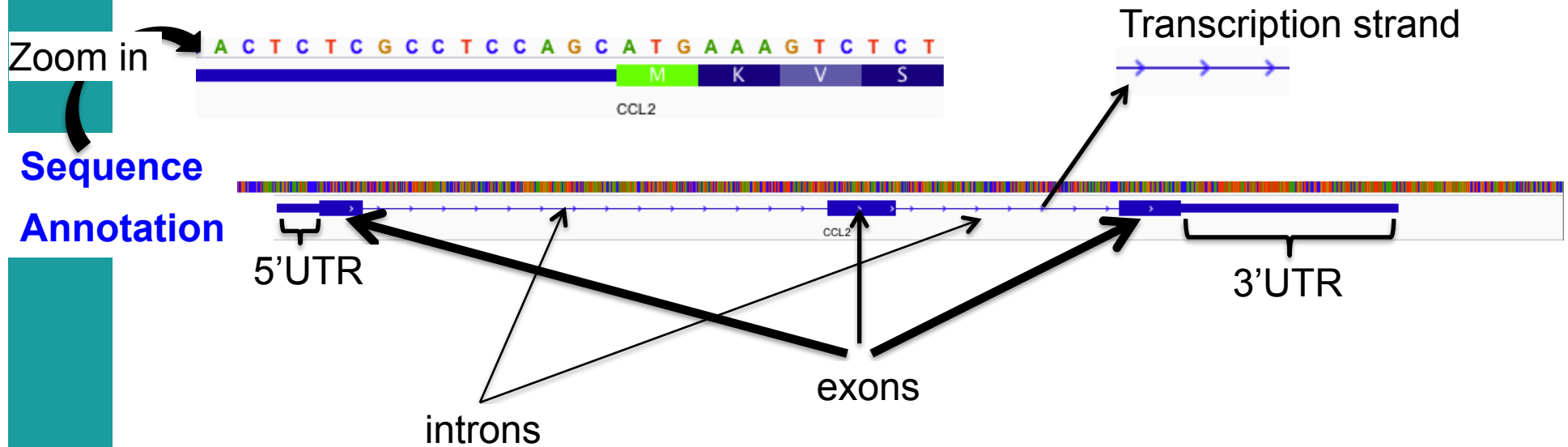
→ When mouse hover on images, pop-up windows provide additional information

IGV data track differences vs reference genome



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

IGV annotation track



Zoom in

Sequence
Annotation

→ 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

9: Tophat2 on data 4: <u>accepted hits</u>	  
8: Tophat2 on data 4: <u>splice junctions</u>	  
7: Tophat2 on data 4: <u>deletions</u>	  
6: Tophat2 on data 4: <u>insertions</u>	  
5: Tophat2 on data 4: <u>align summary</u>	  

→ 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 ?
 - 3.6. 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 2 : whole dataset alignments (1/2)

- 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 ?

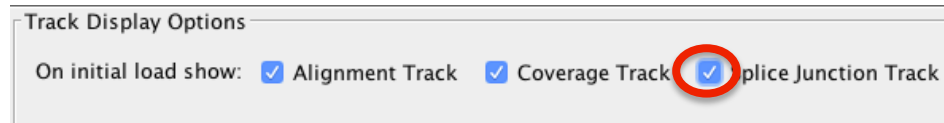
3. Load the 4 BAM files on IGV.

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 (2/2)

4. What do you observe in exons 11 and 13 of *Eef2* gene ?
5. Look at splice junctions identified on *Acp5* gene

- For this purpose verify in View → Preferences → Alignments that “Splice Junction Track is selected



- 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
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 *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 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

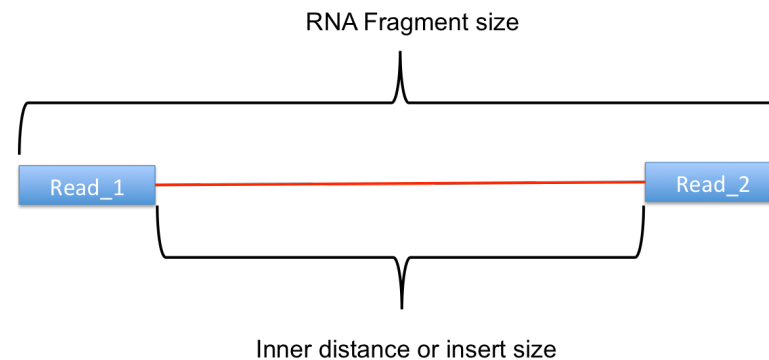
Read Quality determines Phred quality score

Read Distribution calculates how mapped reads were distributed over genome feature

Read GC determines GC% and read count

Distance between reads (paired-end sequencing)

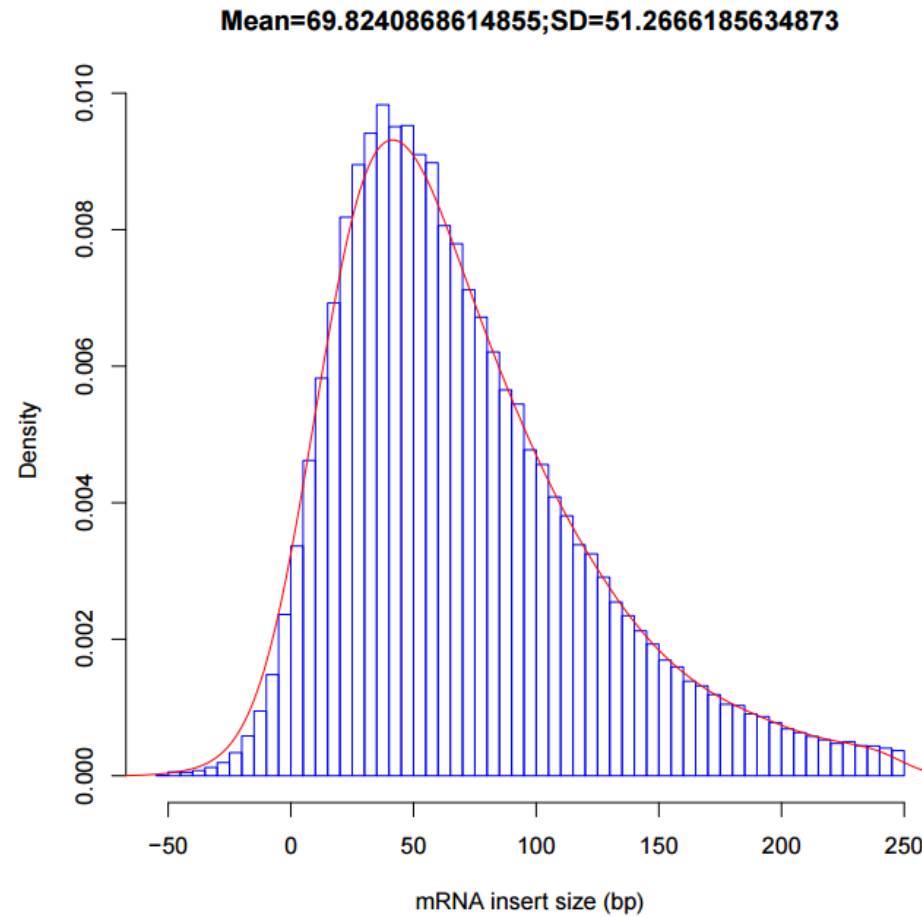
- To know inner distance (insert size) between paired reads
 - The distance is the mRNA length between two paired fragments



- 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 “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	-+	--

RSeQC infer experiment : examples of result

Result on siLuc2 (directional protocol)

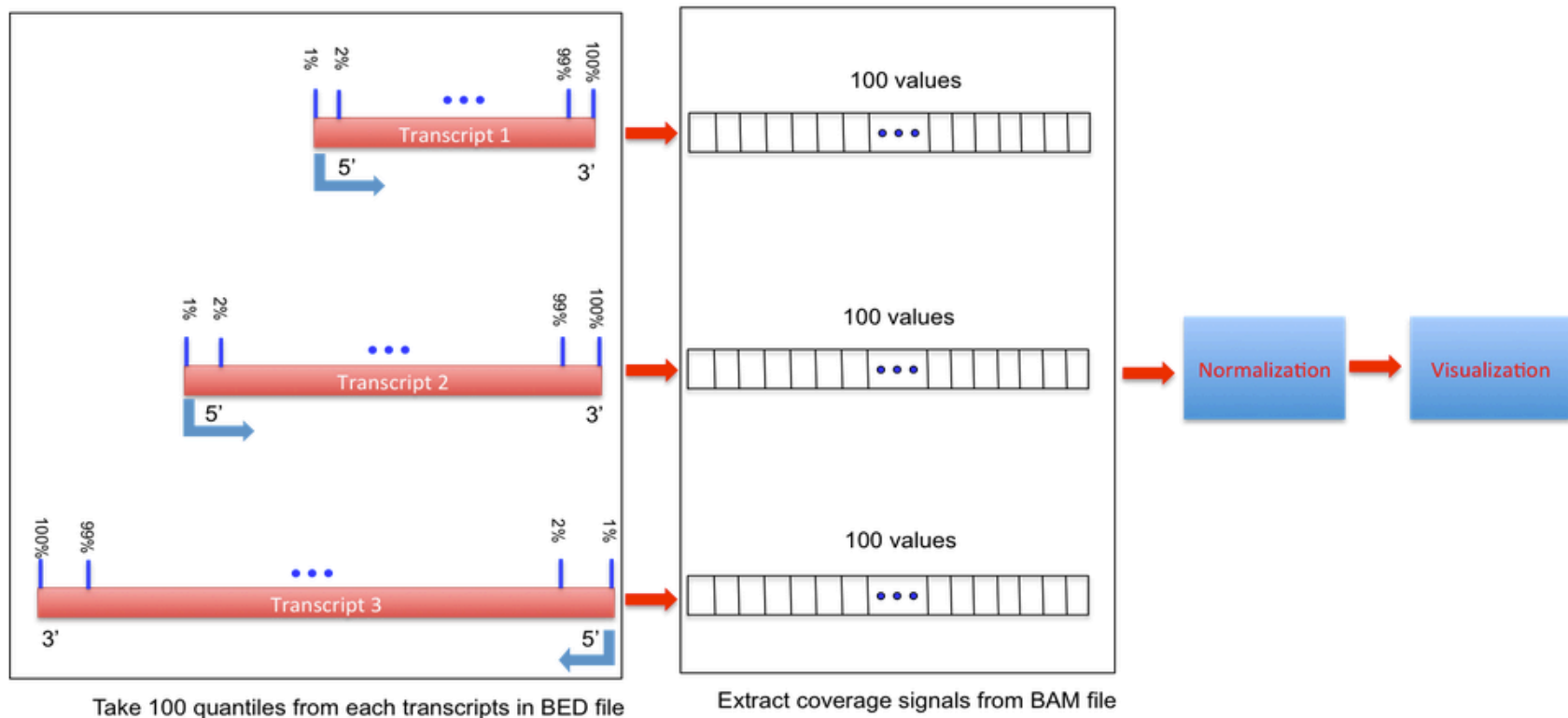
```
This is SingleEnd Data  
Fraction of reads explained by "++,--": 0.0090  
Fraction of reads explained by "+-,-+": 0.9910  
Fraction of reads explained by other combinations: 0.0000
```

Result on siLuc2 (standard protocol)

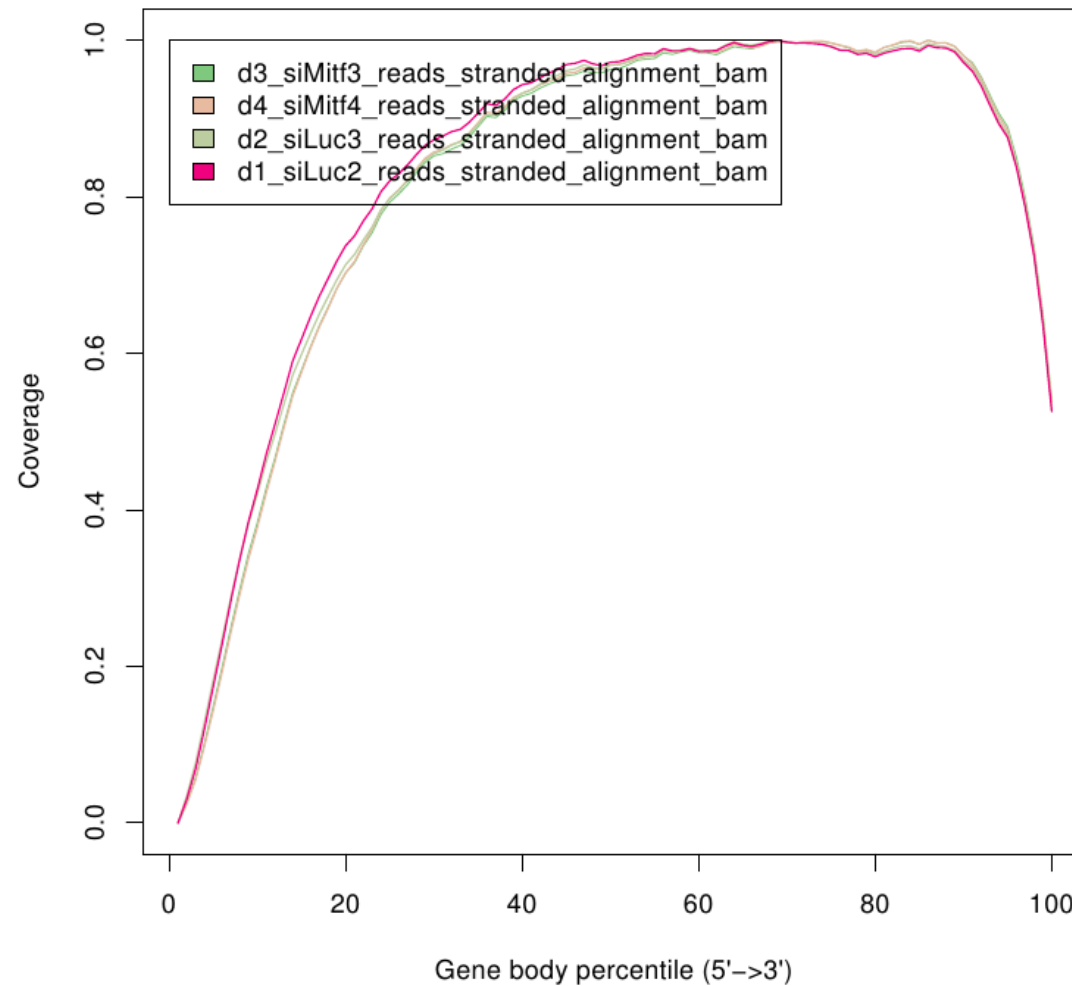
```
This is SingleEnd Data  
Fraction of reads explained by "++,--": 0.4984  
Fraction of reads explained by "+-,-+": 0.5016  
Fraction of reads explained by other combinations: 0.0000
```

Read coverage over genes

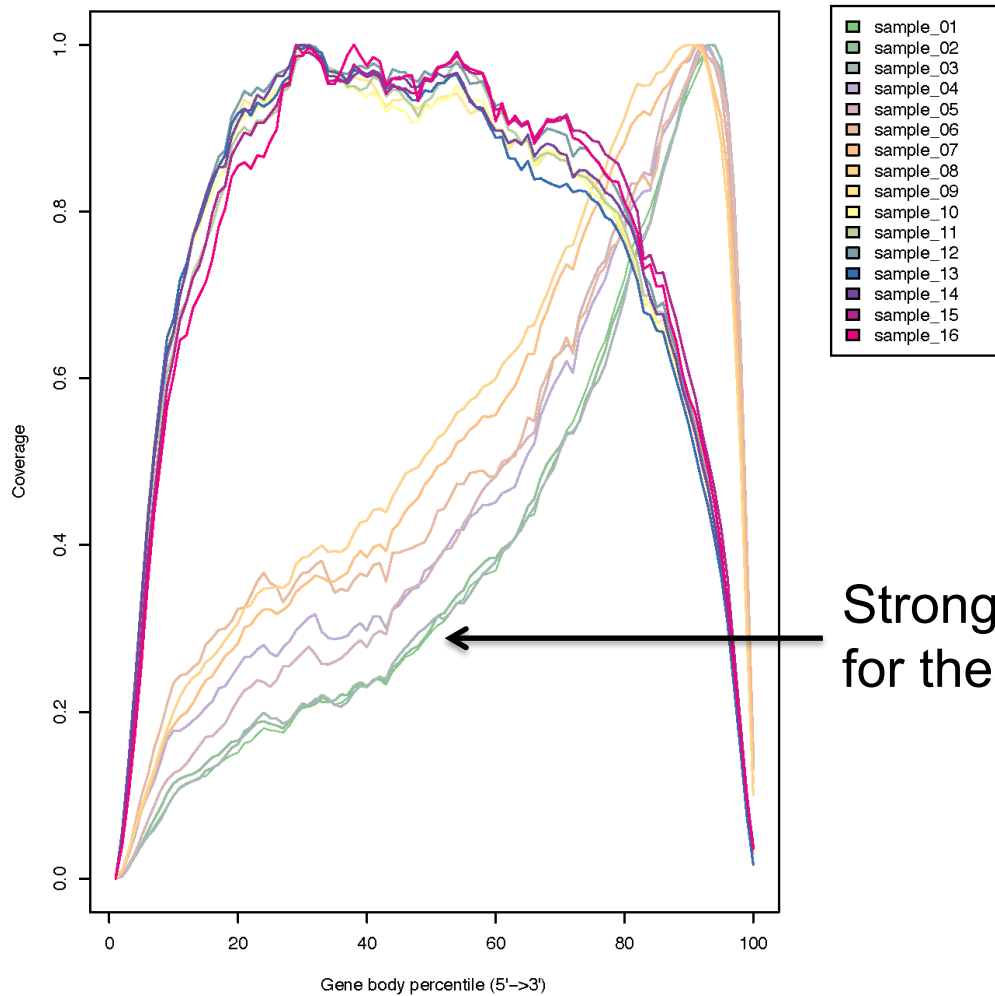
- To identify any bias in read coverage over genes
- RSeQC Gene Body Coverage



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”