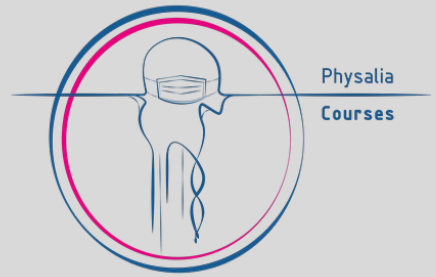
Processing NGS data

Epigenomics Data Analysis

Jacques Serizay

Physalia 2023



NGS processing workflow

- Get .bcl files
- Create fastq files
- QC: remove/trim low quality reads
- Align fastq to BAM
- Filter duplicates, artifacts, …
- Generate tracks
- Assay-specific downstream analysis

NGS processing workflow

- Get .bcl files
- Create fastq files
- QC: remove/trim low quality reads
- Align fastq to BAM
- Filter duplicates, artifacts
- Generate tracks
- Assay-specific downstream analysis

CHECK YOUR DATA

CHECK YOUR DATA

CHECK YOUR DATA

CHECK YOUR DATA

CHECK YOUR DATA

CHECK YOUR DATA

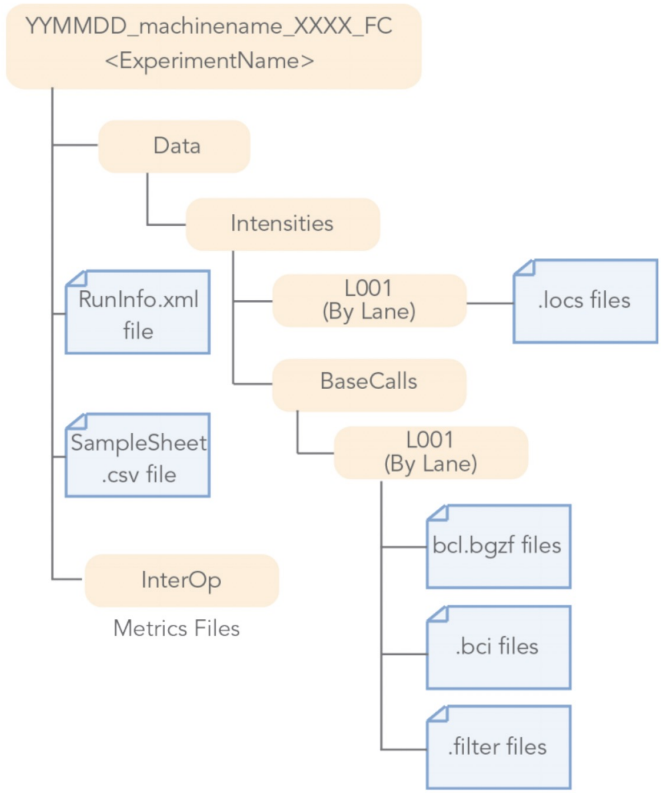
.bcl files

.bcl:

- **Raw data** output of a sequencing run
- **Binary**, non-human-readable file

- Contains the **base calling** and **quality score per cluster, per sequencing lane, per cycle**

- **Huge files**
- **No aggregated sequence per read**



NGS processing workflow

- Get .bcl files 
- Create fastq files
- QC: remove/trim low quality reads
- Align fastq to BAM
- Filter duplicates, artifacts, …
- Generate tracks
- Assay-specific downstream analysis

Fastq files

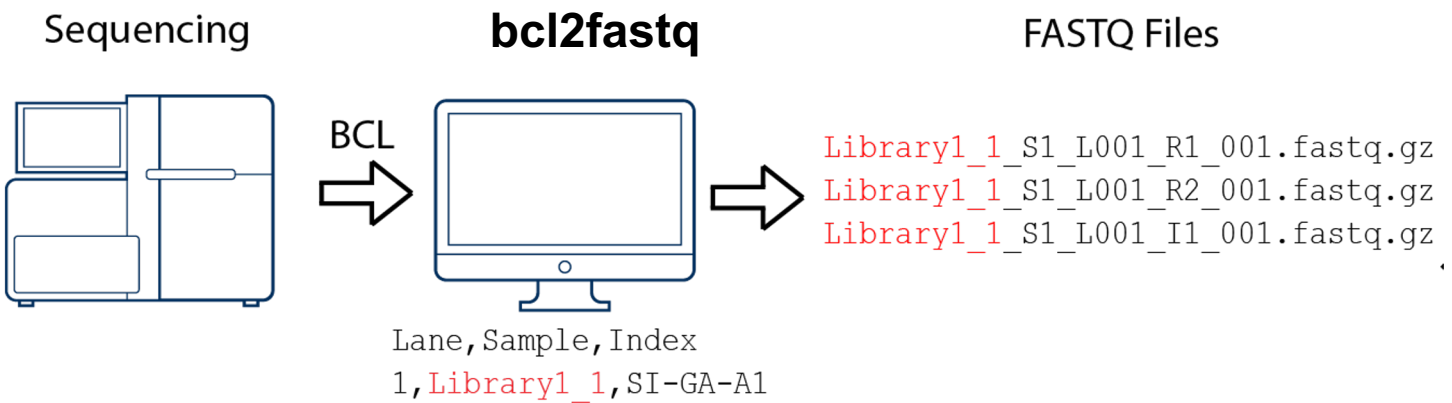
A fastq file contains reads, each read is composed of 4 lines:

1. A sequence identifier with information about the sequencing run
2. The sequence (the base calls; A, C, T, G and N).
3. A separator, which is simply a plus (+) sign.
4. The base call quality scores, using ASCII characters to represent the numerical quality scores.

```
jacquesserizay@LOCAL[12:46:19]:~ $ cat SRR11575369_1.fastq.gz | zcat |  head -n 8
@SRR11575369.1 1/1
ANCAACAGTGGAATTCGTTTTGATGAAAAAAATAAATTTGTTCTTCAAAGCAGAGTGAATGATGCAGTACGAGCTCTTGCTCTTGAAAACCCATCACAACTTTATAATTTAATAATTAGTGAAAAATTAAAAAAATAATATTTCTTATATT
+
F#F:FFFFF:FFFFFFFFFF:F,FF,FF::F,::FFFFFFFFFFF:FF:FF:F,FFF,F:F,FFF,FF:FF:F:FFF:FF:FFFF,:F::FFFF:FF,FF:F:FF,,:F:F,::,F:F:FF,,::,FF:,,F,,,,,:FFFF::F:,,,FF
@SRR11575369.2 2/1
TNGCCAGTCAATACCGCCCTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTGTGTTTTTTTTTTTTATATAATAAAATTTTTTTTTTTTTAATAAAATTTTTTTTTTATTTT
+
F#FFFFFFF::FF:FF::FFFFFFFFFFFFFFFF:FFFFFFFFFFFFF,,FFFFFFFFFF:FFFFFF:FFFFFFFFFFF:F,FFFFF,,,:,,::FFFFFFFF,,:,,,F,,,,,,FFFFFF,F,,FF:,F,:,F,FFFFF,,,:,,F::F
```

bcl2fastq

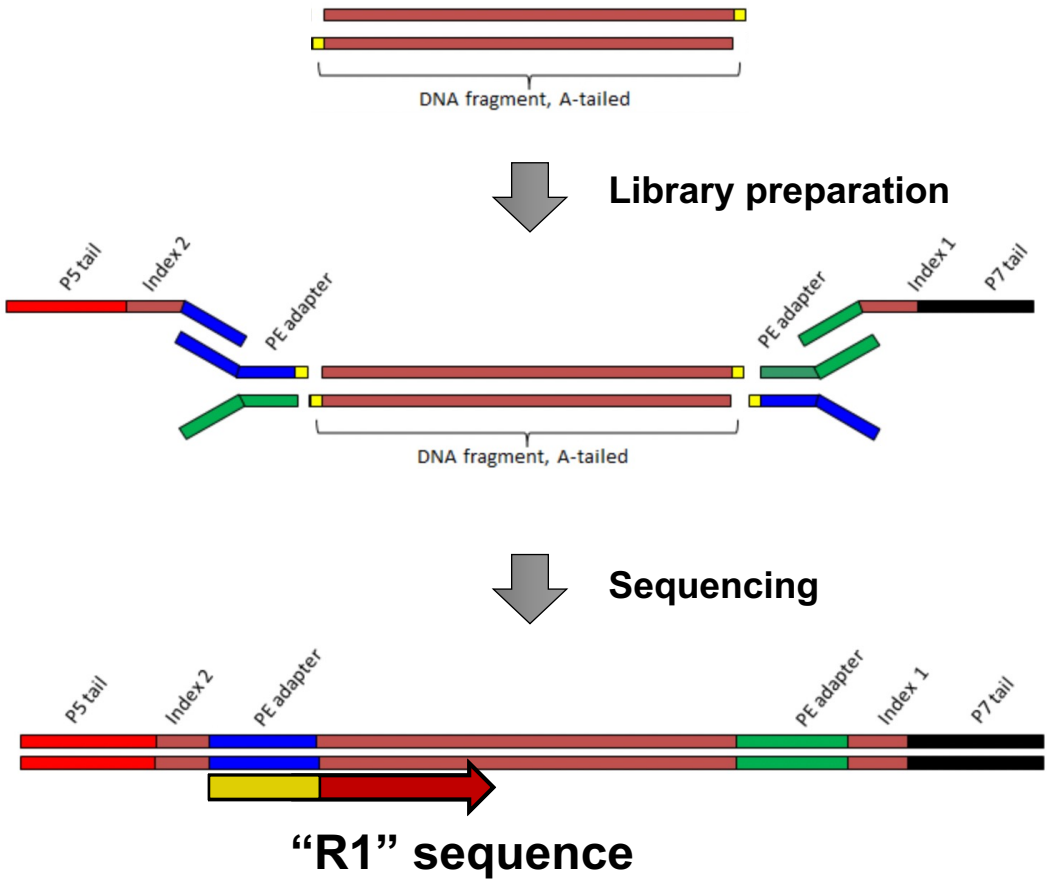
```
bcl2fastq --run-folder-dir <bcl_files_folder> --output-dir <fastq_files_folder>
```



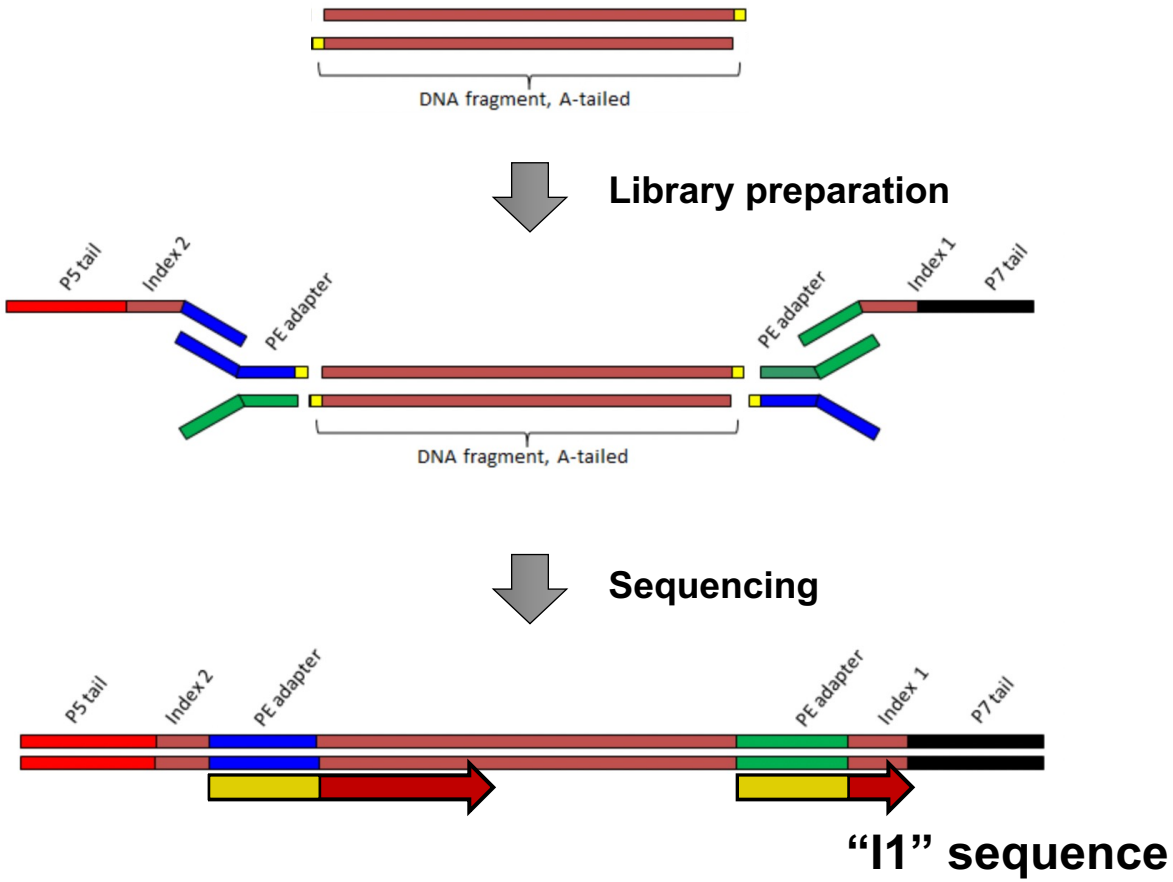
User guide:

https://support.illumina.com/content/dam/illumina-support/documents/documentation/software_documentation/bcl2fastq/bcl2fastq_letterbooklet_15038058brpmi.pdf

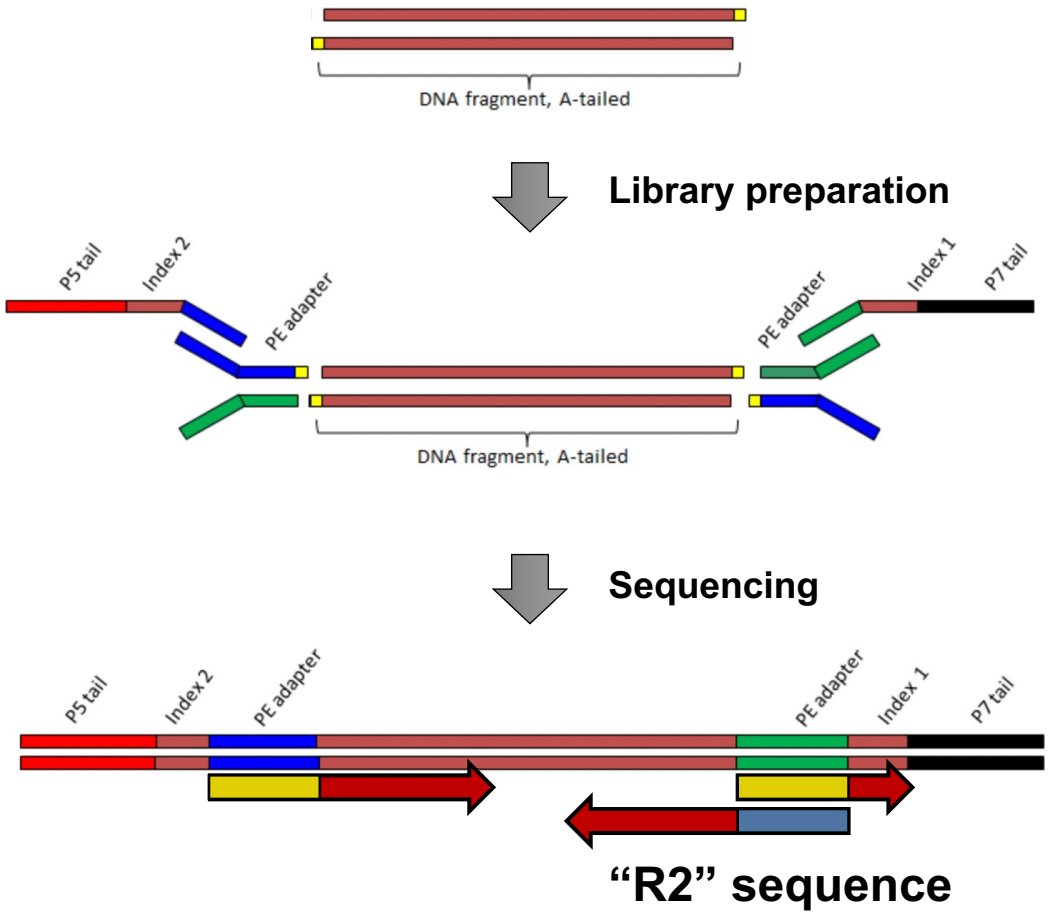
Why so many fastq files?



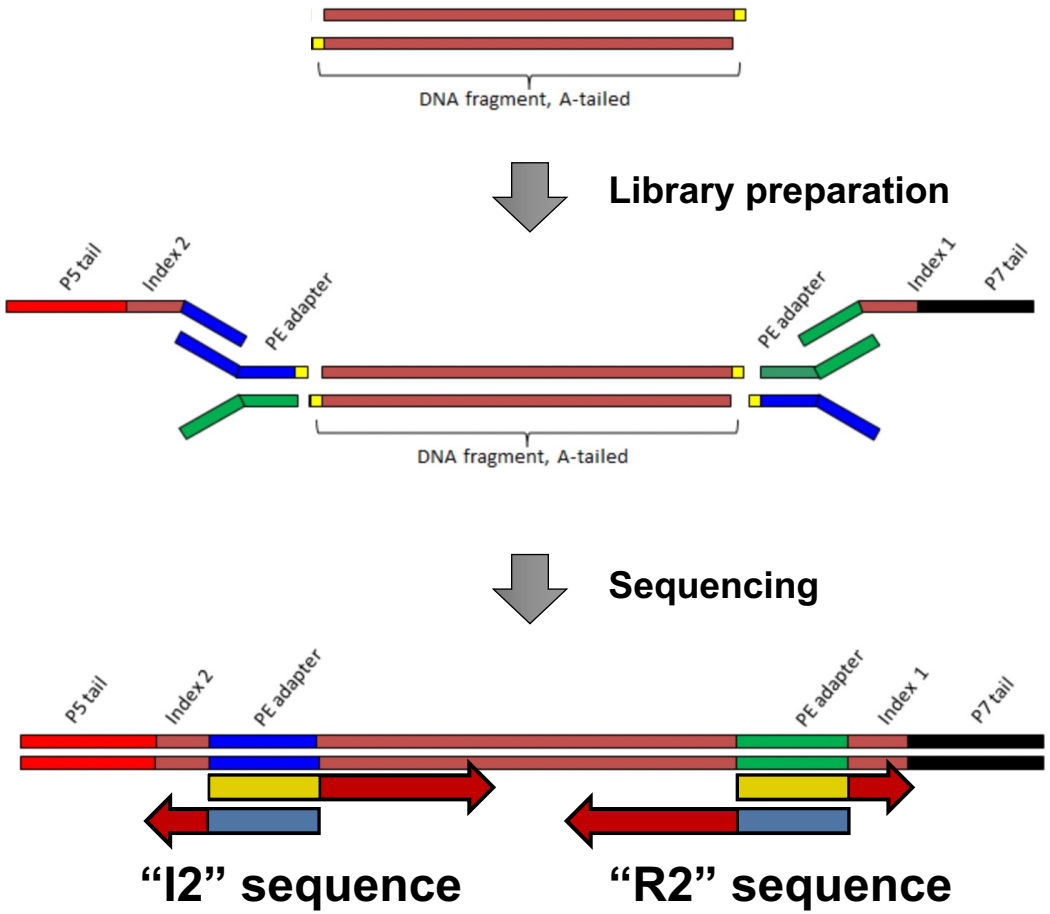
Why so many fastq files?



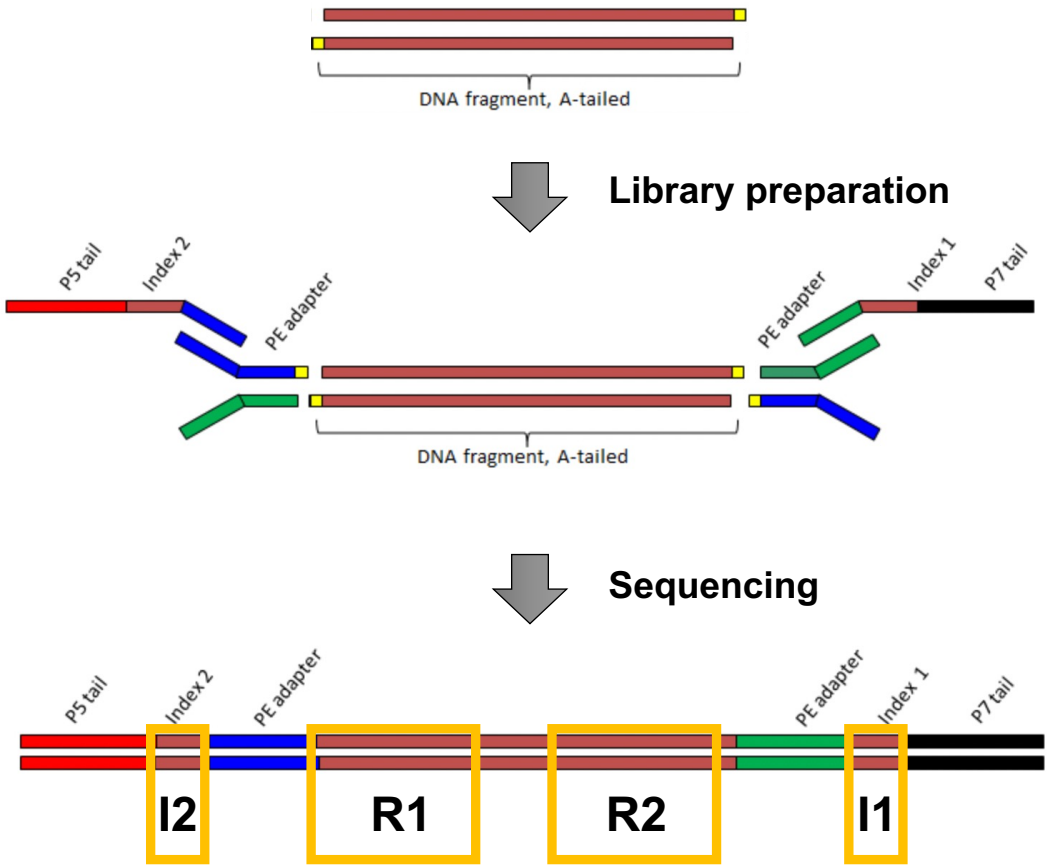
Why so many fastq files?



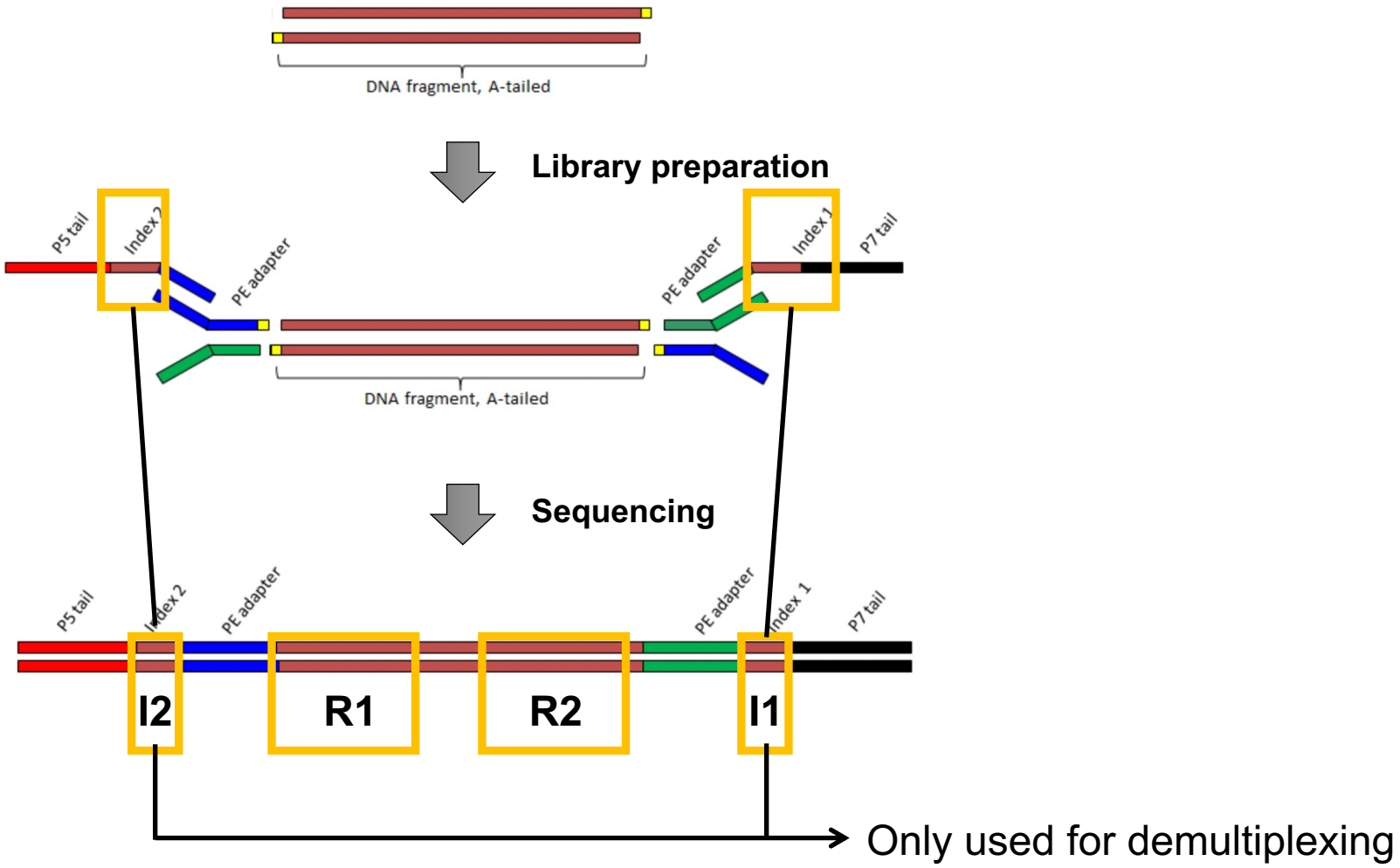
Why so many fastq files?



Why so many fastq files?



Why so many fastq files?



NGS processing workflow

- Get .bcl files 
- Create fastq files 

Or **`bcl2fastq`**

- QC: remove/trim low quality reads
- Align fastq to BAM
- Filter duplicates, artifacts, …
- Generate tracks
- Assay-specific downstream analysis

NGS processing workflow

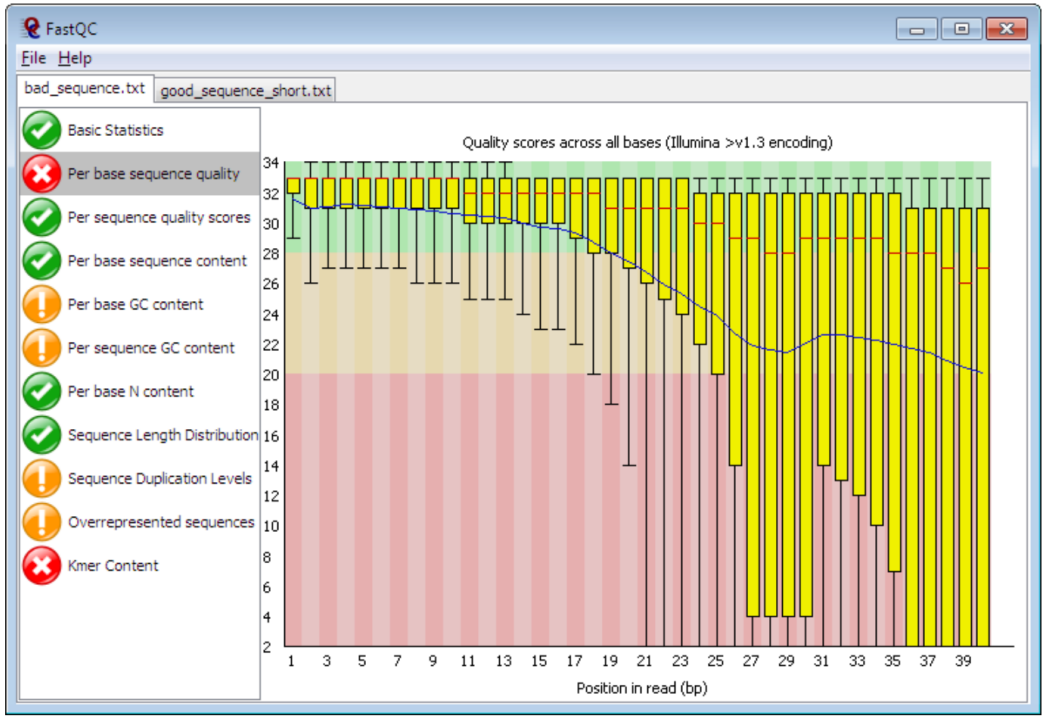
- ✓ Get .bcl files 
- ✓ Create fastq files  Or **`bcl2fastq`**

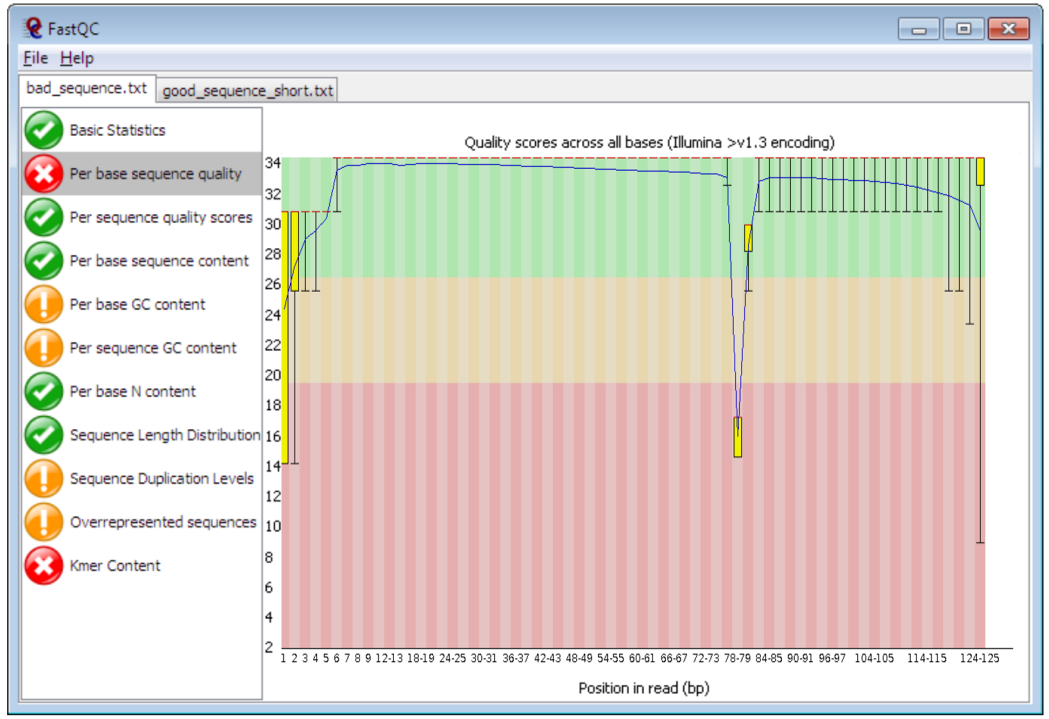
CHECK YOUR DATA

- QC: remove/trim low quality reads
- Align fastq to BAM
- Filter duplicates, artifacts, …
- Generate tracks
- Assay-specific downstream analysis

FastQC

FastQC is a program designed to spot potential problems in high througput sequencing datasets. It runs a set of analyses on one or more raw sequence files in fastq or bam format and produces a report which summarises the results.

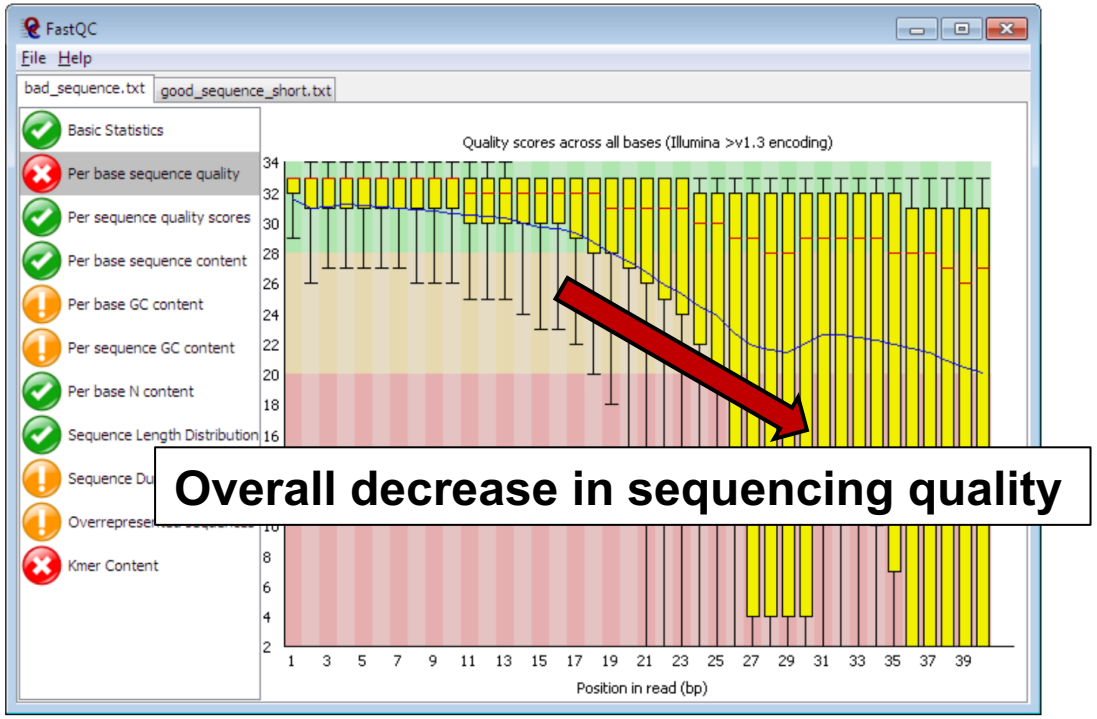


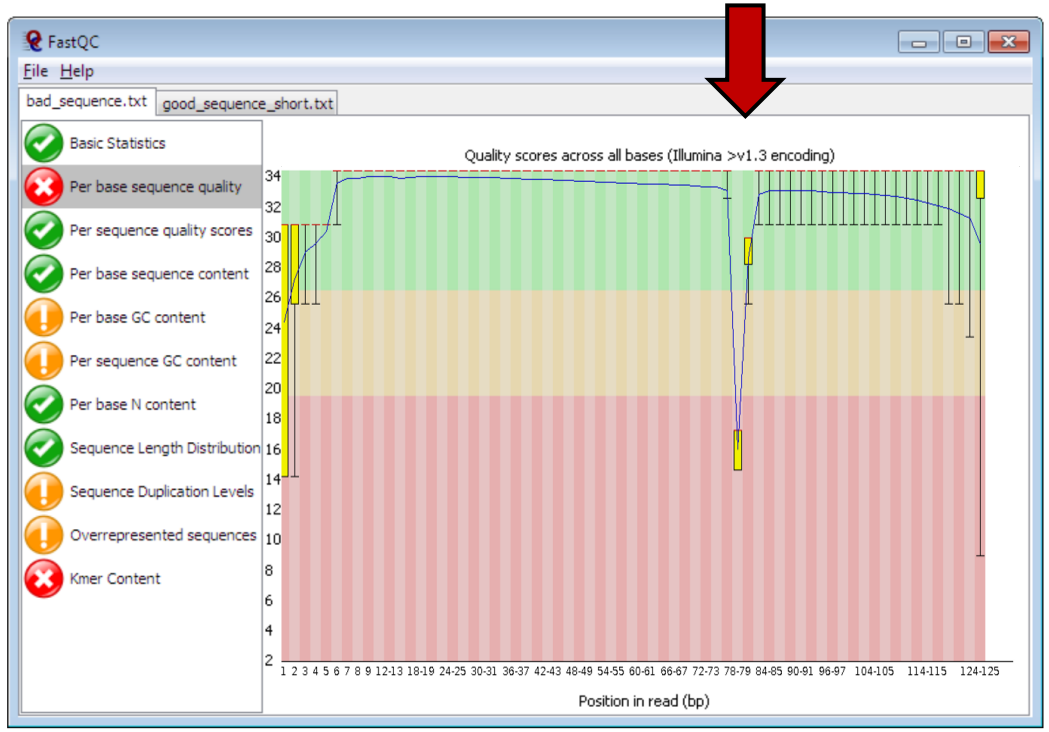


FastQC will highlight any areas where this library looks unusual and where you should take a closer look. The program is not tied to any specific type of sequencing technique and can be used to look at libraries coming from a large number of different experiment types (Genomic Sequencing, ChIP-Seq, RNA-Seq, BS-Seq etc etc).

FastQC

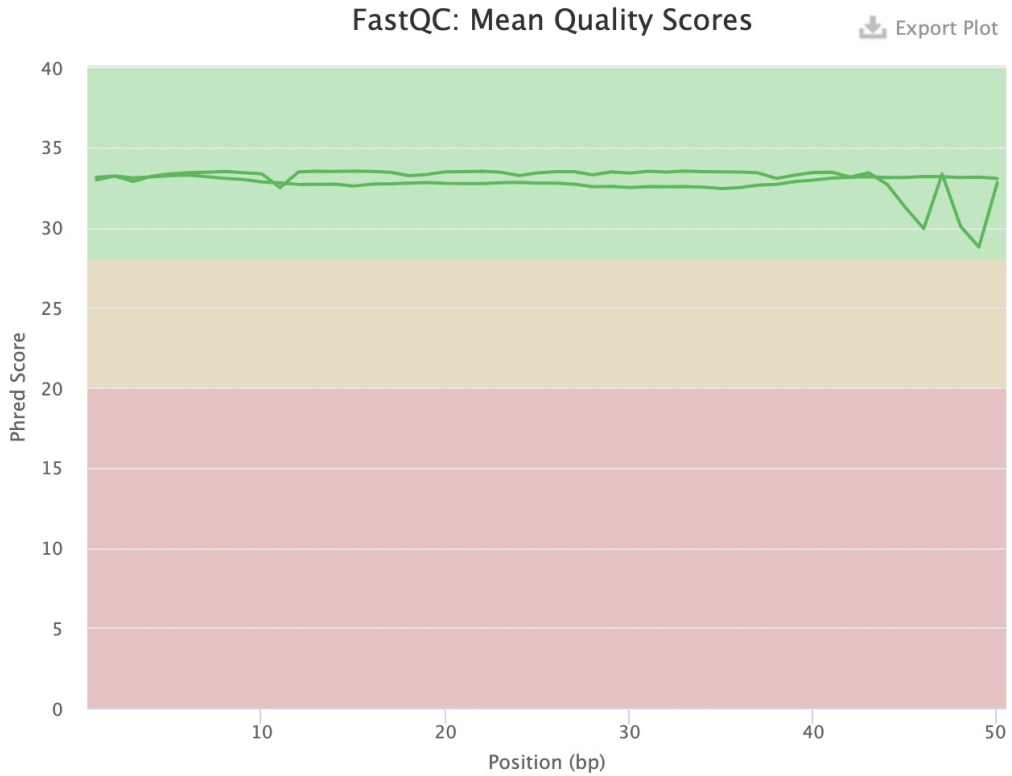
FastQC is a program designed to spot potential problems in high througput sequencing datasets. It runs a set of analyses on one or more raw sequence files in fastq or bam format and produces a report which summarises the results.





FastQC will highlight any areas where this library looks unusual and where you should take a closer look. The program is not tied to any specific type of sequencing technique and can be used to look at libraries coming from a large number of different experiment types (Genomic Sequencing, ChIP-Seq, RNA-Seq, BS-Seq etc etc).

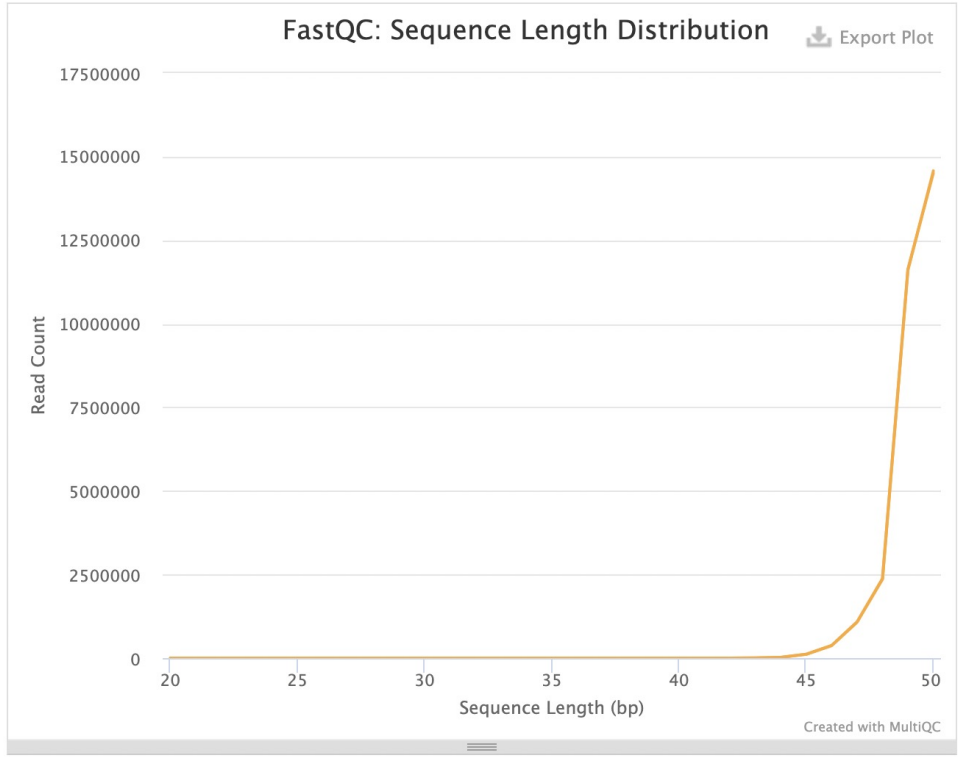
Cutadapt: trim away adapter sequences and low-quality ends



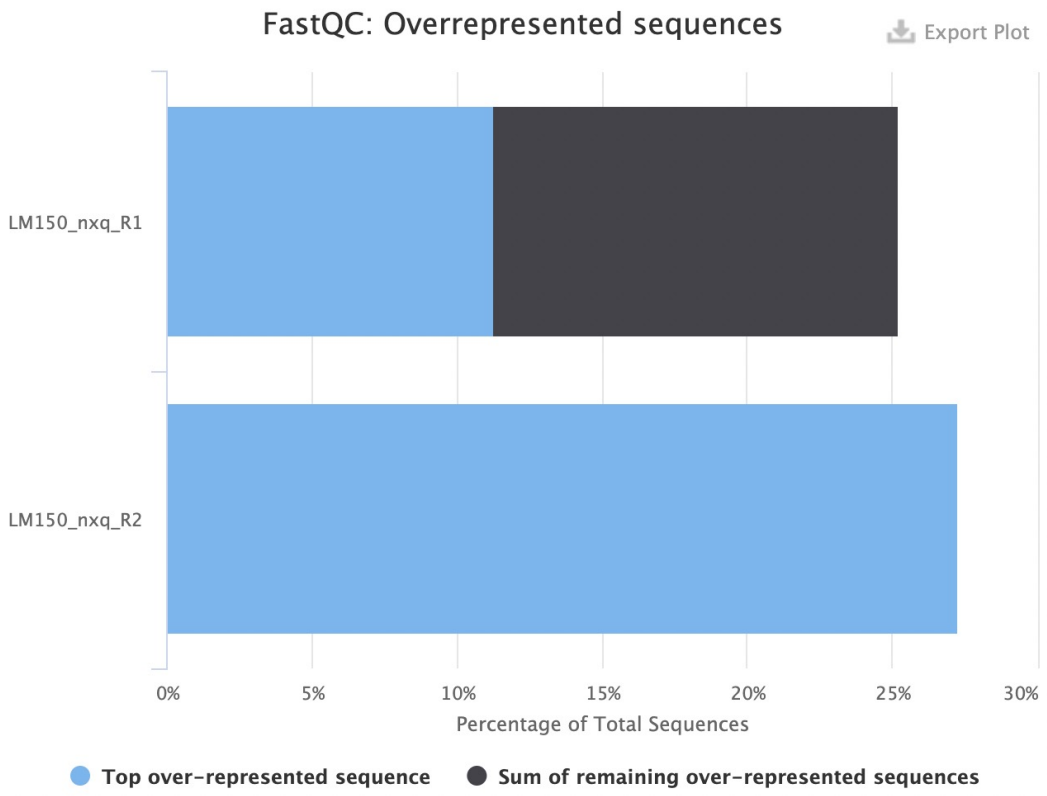
Cutadapt: trim away adapter sequences and low-quality ends

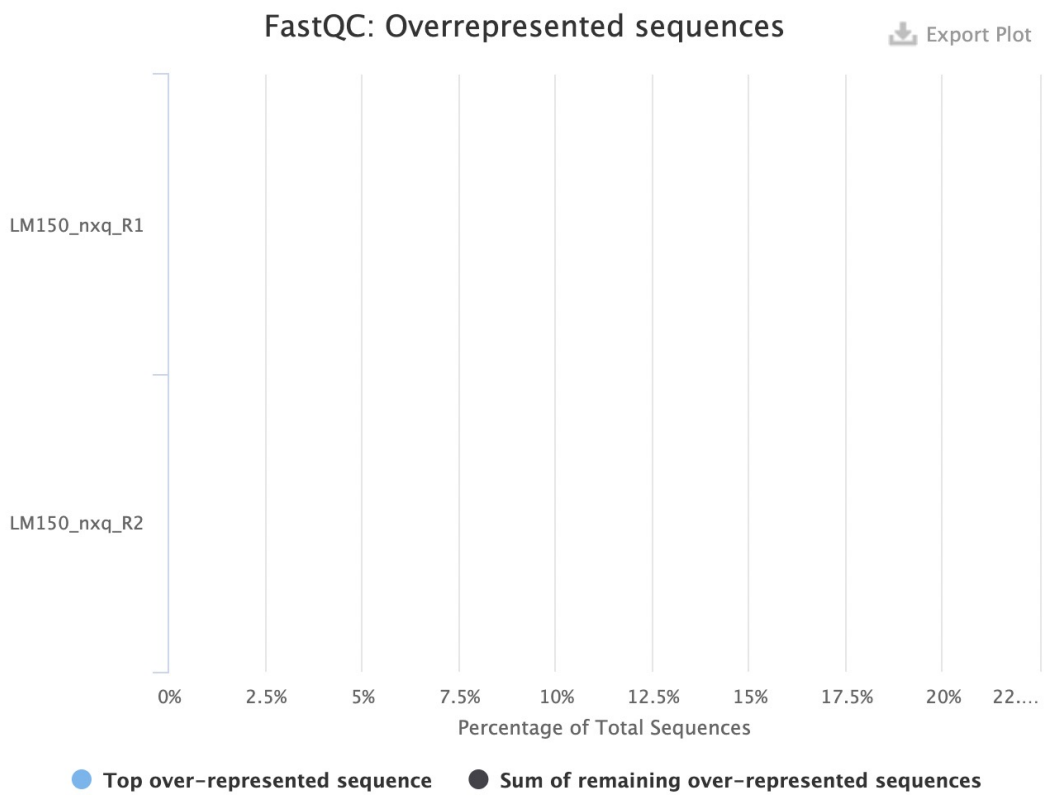
Sequence Length Distribution 80

All samples have sequences of a single length (50bp).



Cutadapt: trim away adapter sequences and low-quality ends

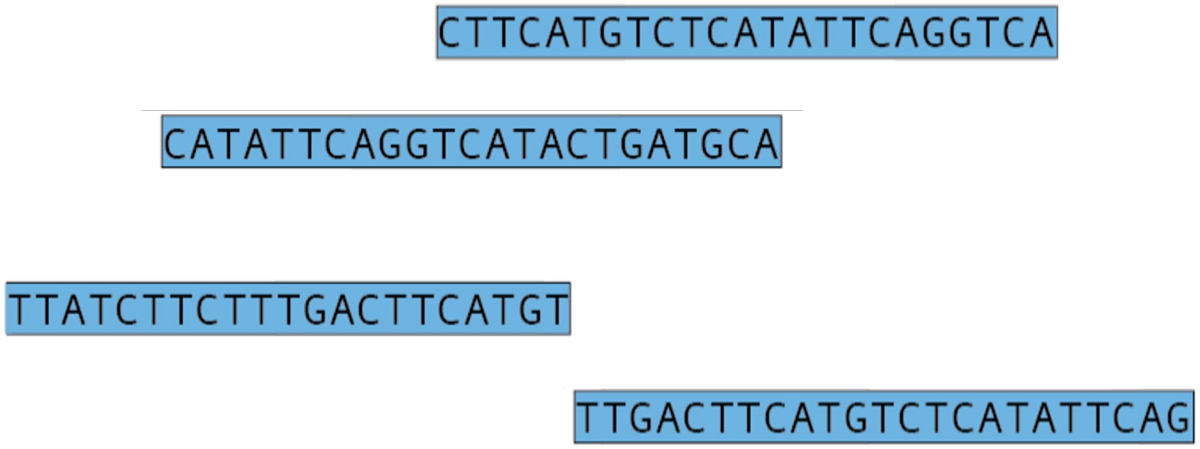




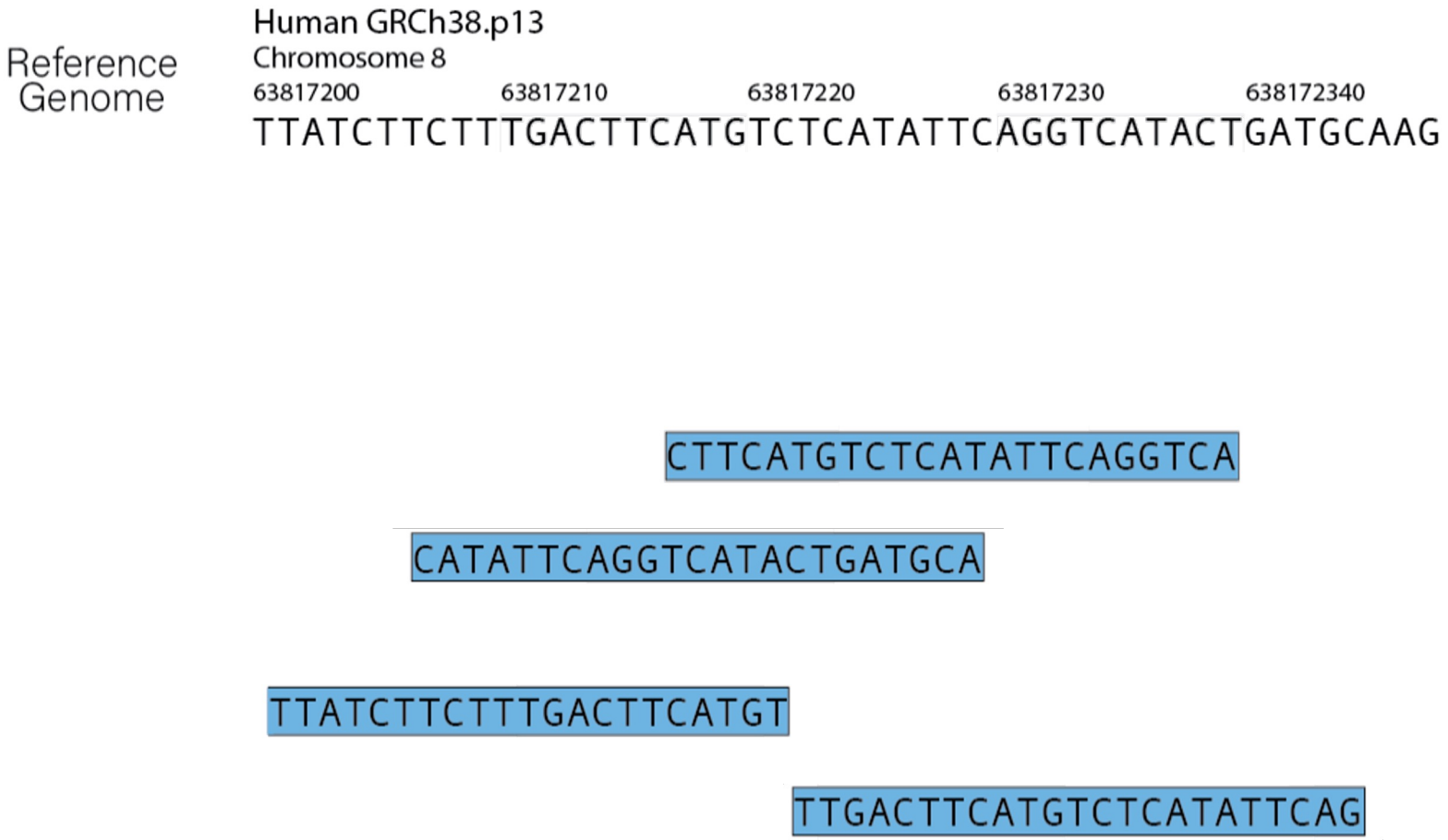
NGS processing workflow

- Get .bcl files 
- Create fastq files  Or **`bcl2fastq`**
- CHECK YOUR DATA E.g. **`FastQC`**
- QC: remove/trim low quality reads E.g. **`cutadapt`**
- Align fastq to BAM
- Filter duplicates, artifacts, …
- Generate tracks
- Assay-specific downstream analysis

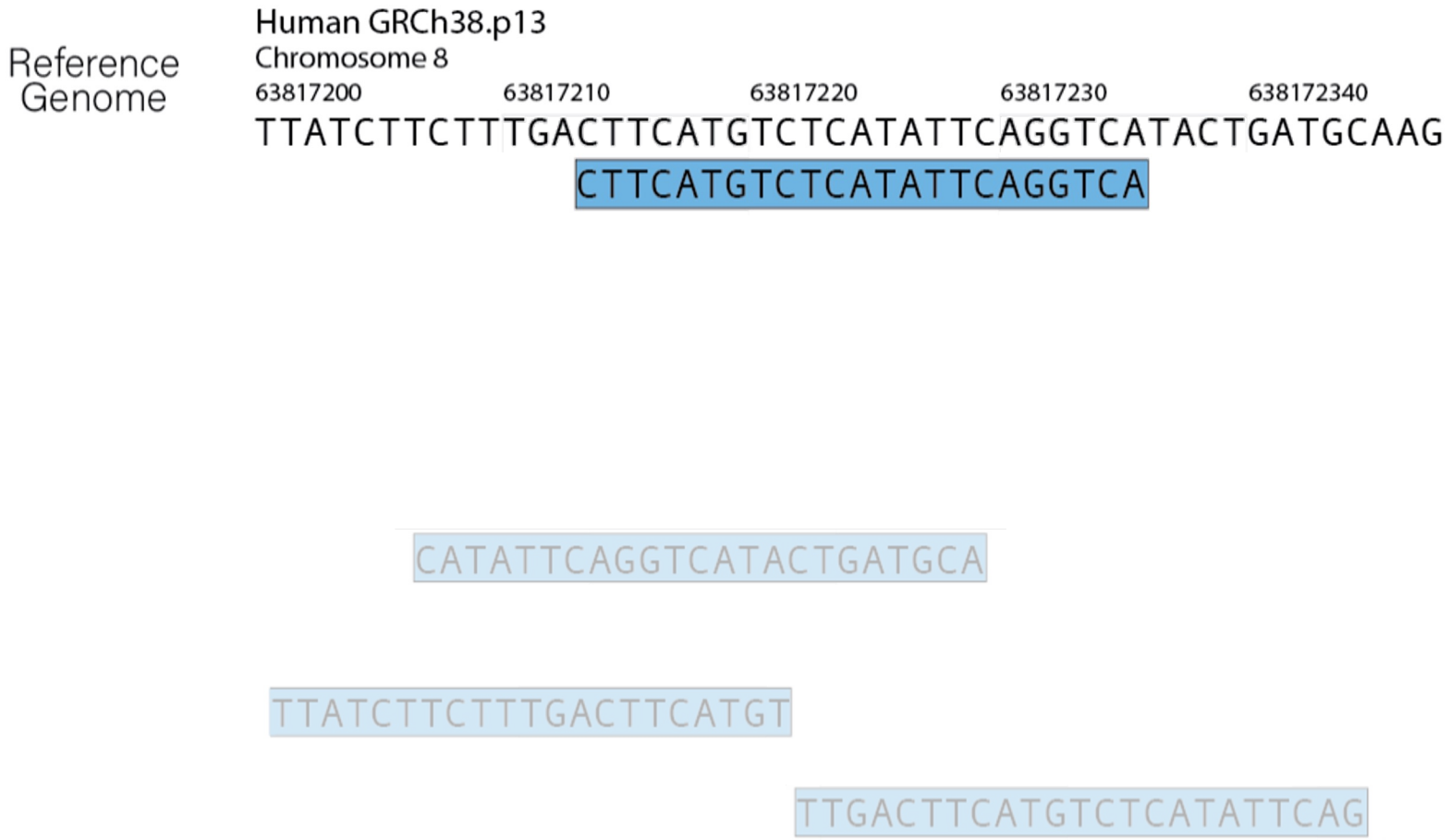
Mapping sequencing reads to a reference



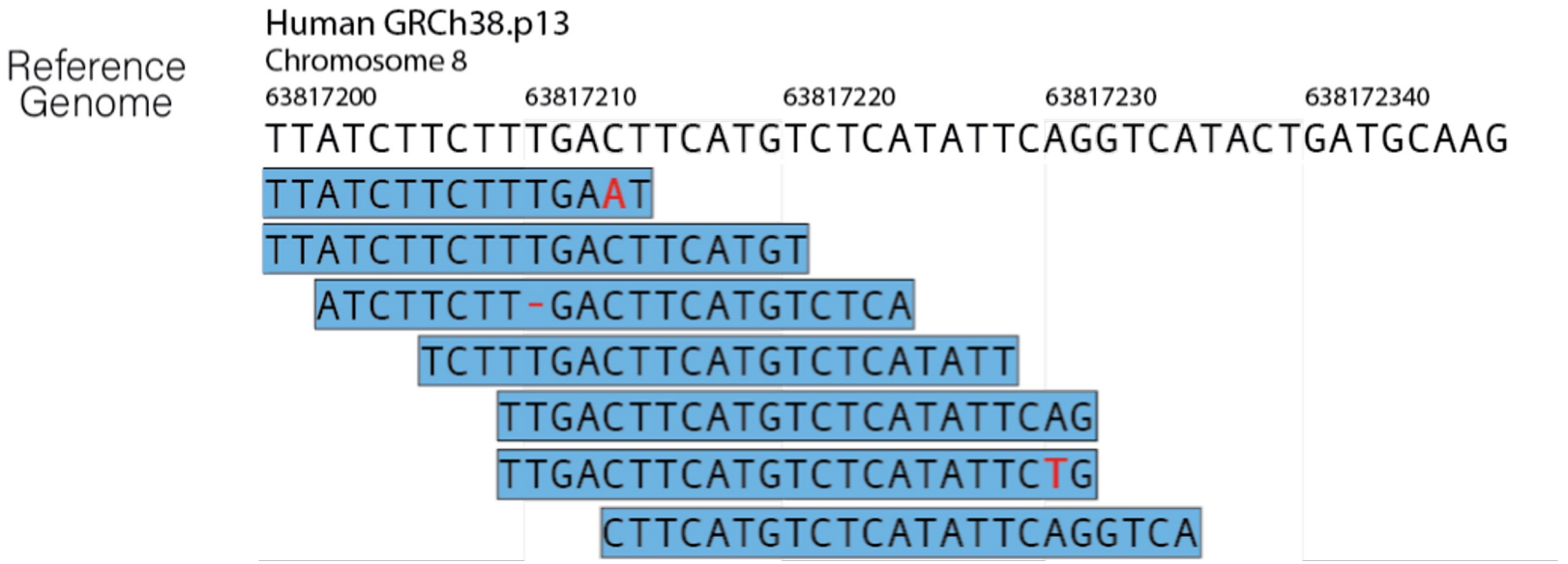
Mapping sequencing reads to a reference



Mapping sequencing reads to a reference



Mapping sequencing reads to a reference



SAM file format

Sequence Alignment Map (SAM) is a human-readable, rectangular, text-based format for storing biological sequences aligned to a reference sequence.

Each entry (line) describes where a read is mapped on the reference and how it is mapped

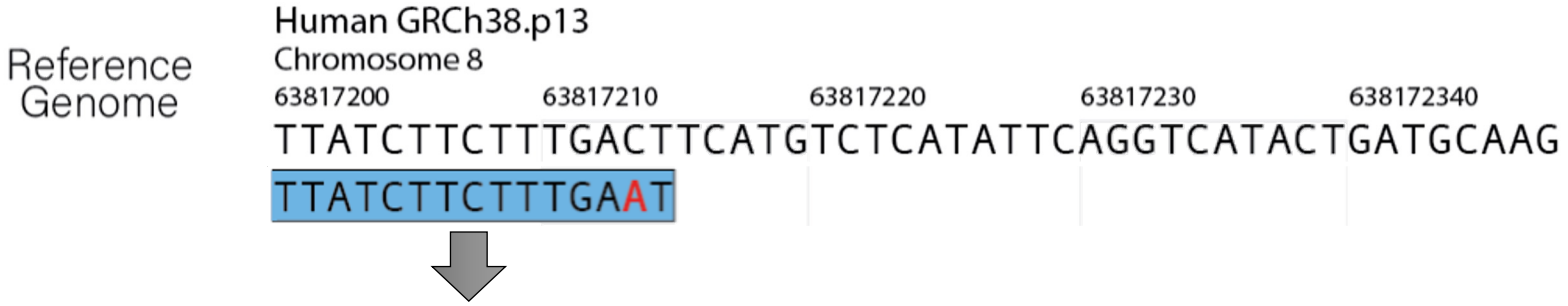
Col	Field	Type	Brief description
1	QNAME	String	Query template NAME
2	FLAG	Int	bitwise FLAG
3	RNAME	String	References sequence NAME
4	POS	Int	1- based leftmost mapping POSition
5	MAPQ	Int	MAPping Quality
6	CIGAR	String	CIGAR string
7	RNEXT	String	Ref. name of the mate/next read
8	PNEXT	Int	Position of the mate/next read
9	TLEN	Int	observed Template LENgth
10	SEQ	String	segment SEQuence
11	QUAL	String	ASCII of Phred-scaled base QUALity+33

SAM file format

Sequence Alignment Map (SAM) is a human-readable, rectangular, text-based format for storing biological sequences aligned to a reference sequence.

Each entry (line) describes where a read is mapped on the reference and how it is mapped

Col	Field	Type	Brief description
1	QNAME	String	Query template NAME
2	FLAG	Int	bitwise FLAG
3	RNAME	String	References sequence NAME
4	POS	Int	1- based leftmost mapping POSition
5	MAPQ	Int	MAPping Quality
6	CIGAR	String	CIGAR string
7	RNEXT	String	Ref. name of the mate/next read
8	PNEXT	Int	Position of the mate/next read
9	TLEN	Int	observed Template LENgth
10	SEQ	String	segment SEQuence
11	QUAL	String	ASCII of Phred-scaled base QUALity+33



Read name	Flag	Chr	Position	Length	CIGAR	Read name (mate)	Chr (mate)	Position (mate)	Sequence	Per base sequencing quality
HWI-ST330:304:H045HADXX:2093#1	2	chr8	63817200	50	14M1X	2093#2	chr8	6381932	TTATCTTCTTTGAAT	?????BBBBBBDD=?

BAM file format

BAM files are **binarized** SAM files, allowing great compression of the alignment results.

However, bam files are not directly human-readable.

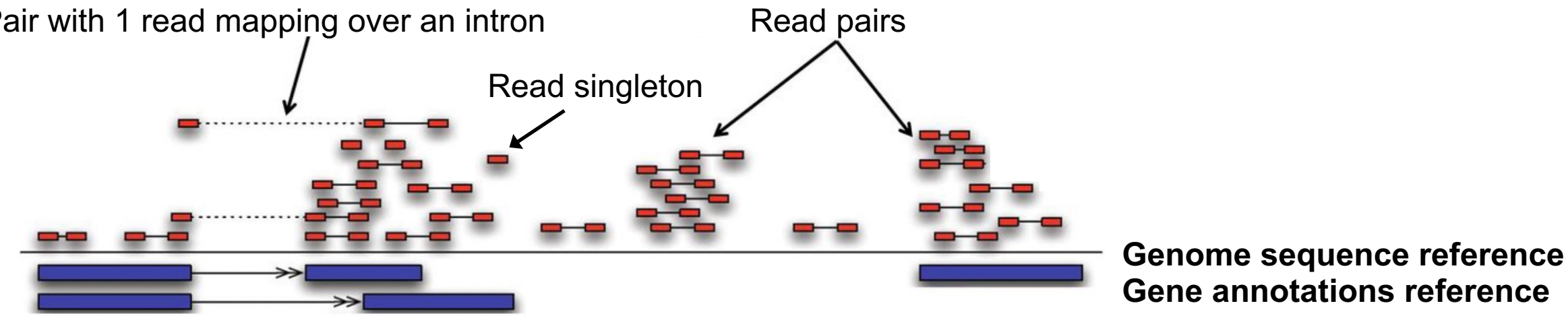
```
samtools view --bam ….sam > ….bam
```

Mapping tools

There are a plethora of alignment tools.

Each one requires the genome reference to be indexed first.

Some mappers can be ”**splice-aware**”, allowing reads to be mapped over annotated introns.



NGS processing workflow

- Get .bcl files 
- Create fastq files  Or **`bcl2fastq`**
- QC: remove/trim low quality reads E.g. **`cutadapt`**
- Align fastq to BAM E.g. **`bowtie2`**
- Filter duplicates, artifacts, …
- Generate tracks
- Assay-specific downstream analysis

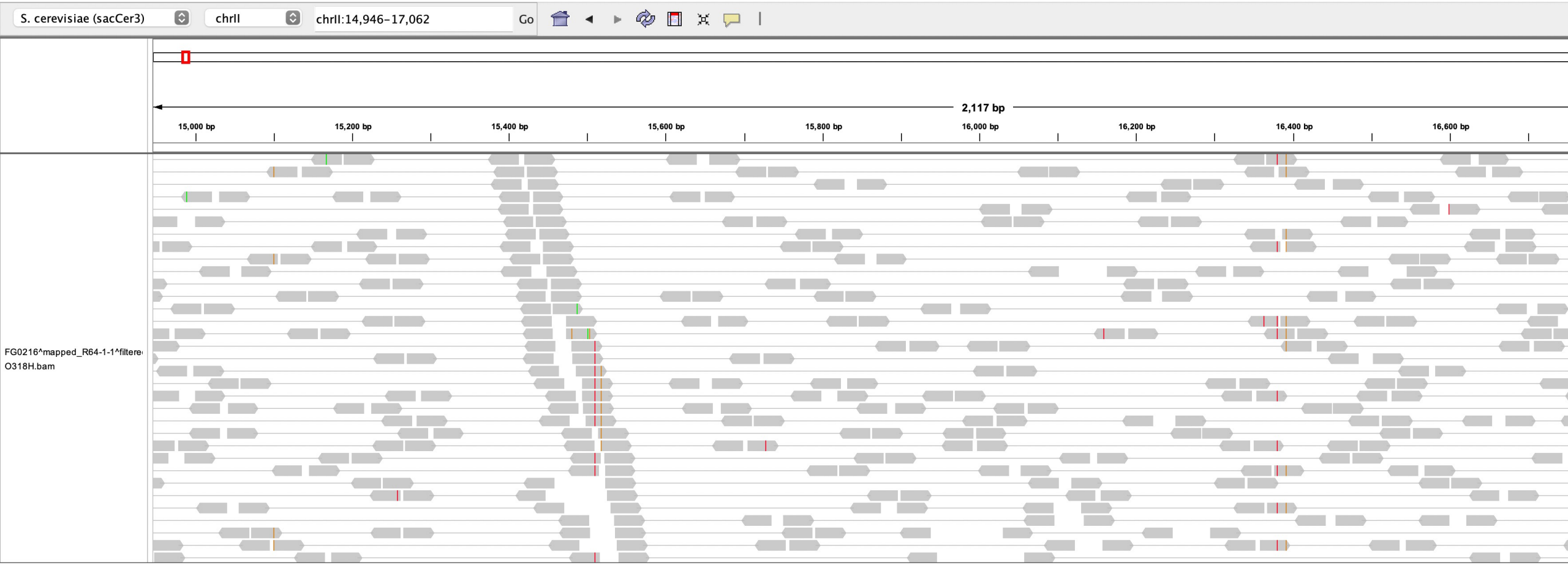
NGS processing workflow

- Get .bcl files 
- Create fastq files  Or **`bcl2fastq`**
- QC: remove/trim low quality reads E.g. **`cutadapt`**
- Align fastq to BAM E.g. **`bowtie2`**

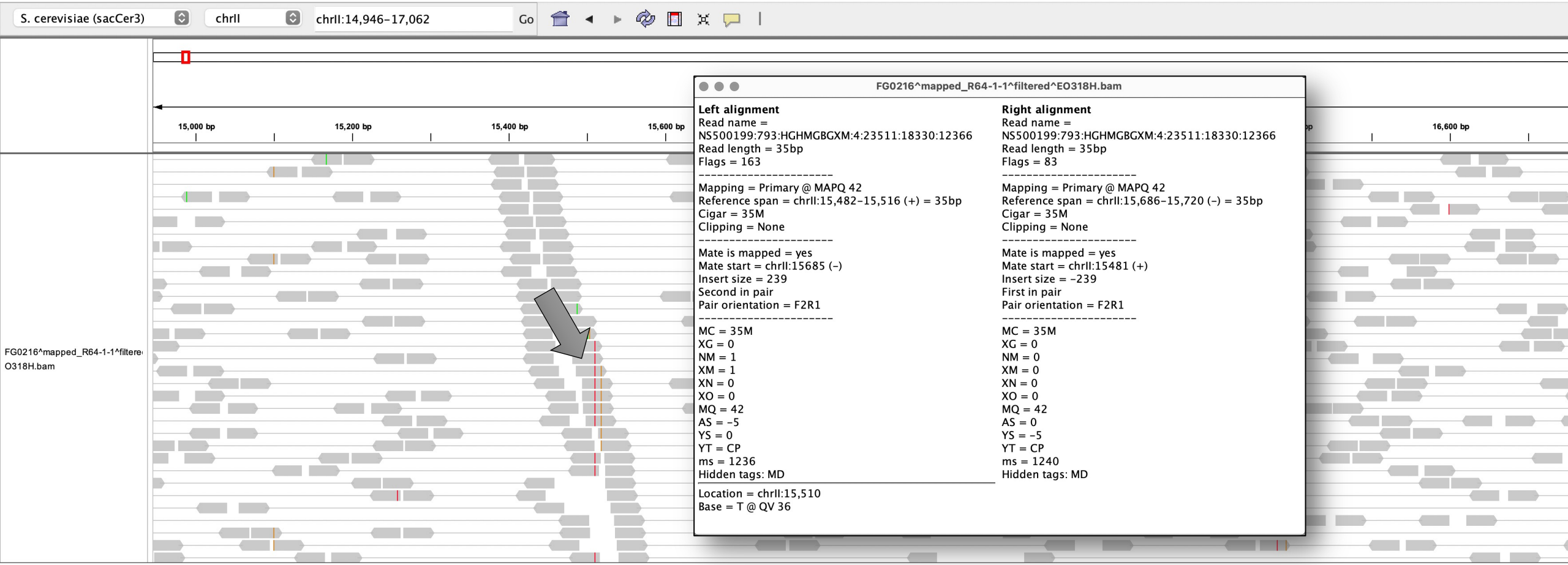
CHECK YOUR DATA

- Filter duplicates, artifacts, …
- Generate tracks
- Assay-specific downstream analysis

IGV: Integrative Genome Browser

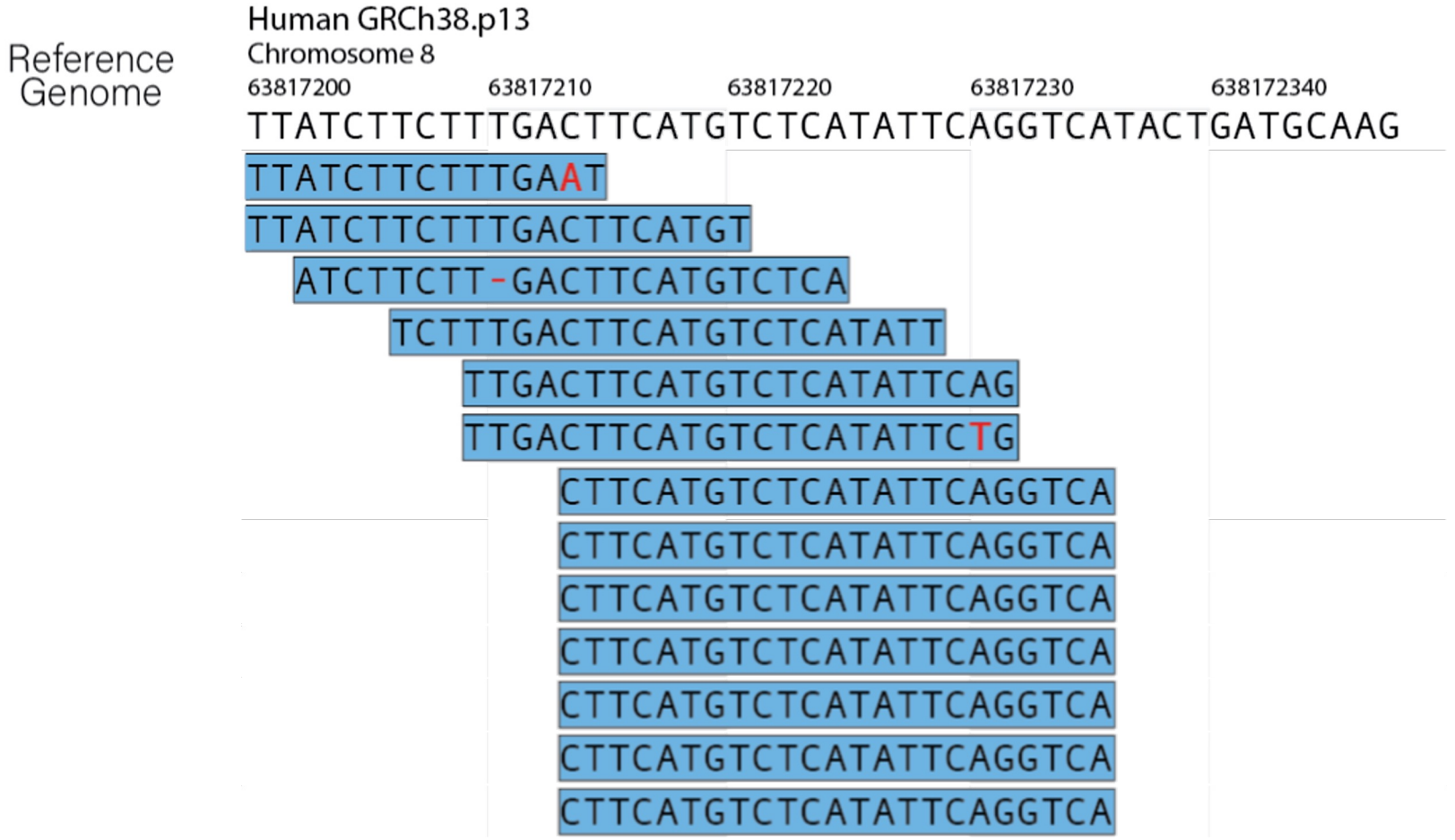


IGV: Integrative Genome Browser



Filtering duplicates

Multiple reads (fragments) with same mapping position (start & end) can be viewed as PCR duplicates.



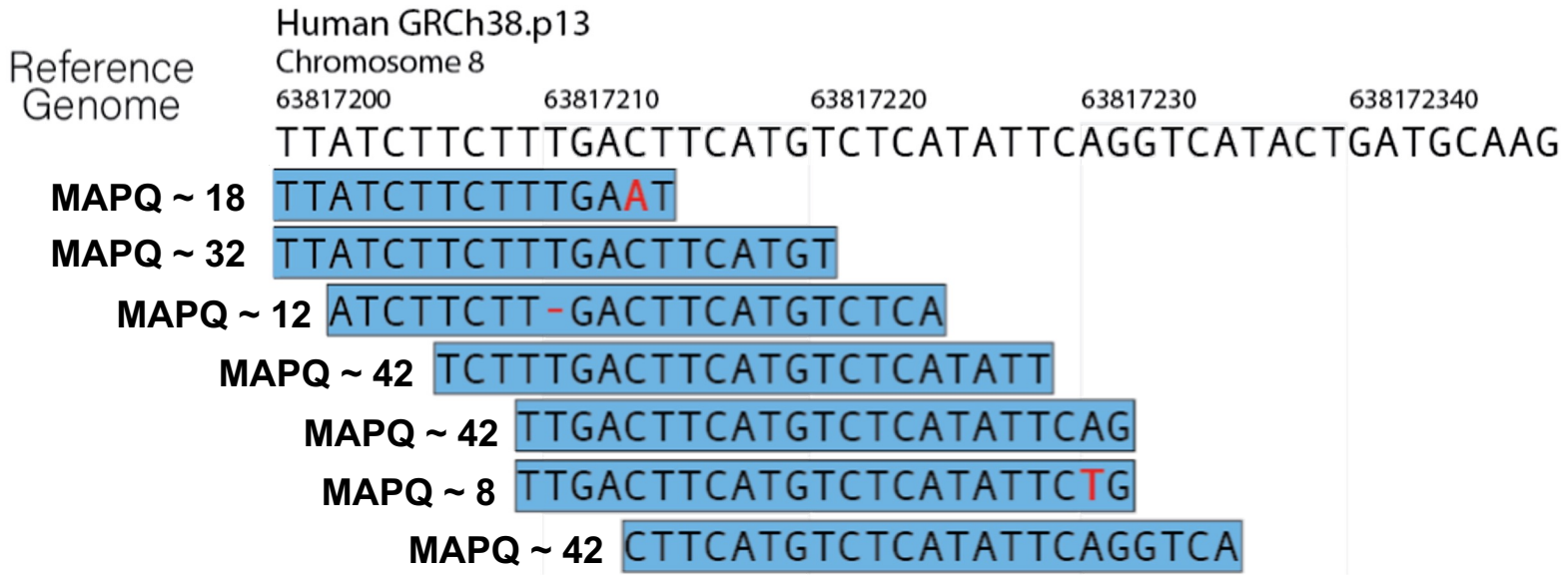
Filtering duplicates

Multiple reads (fragments) with same mapping position (start & end) can be viewed as PCR duplicates.

Filtering out low-quality mapping reads

Mapping Quality Scores (**MAPQ**) quantify the probability that a read is misplaced.

$$MAPQ = -10 * log_{10}\big(P(read\ is\ wrongly\ mapped)\big)$$

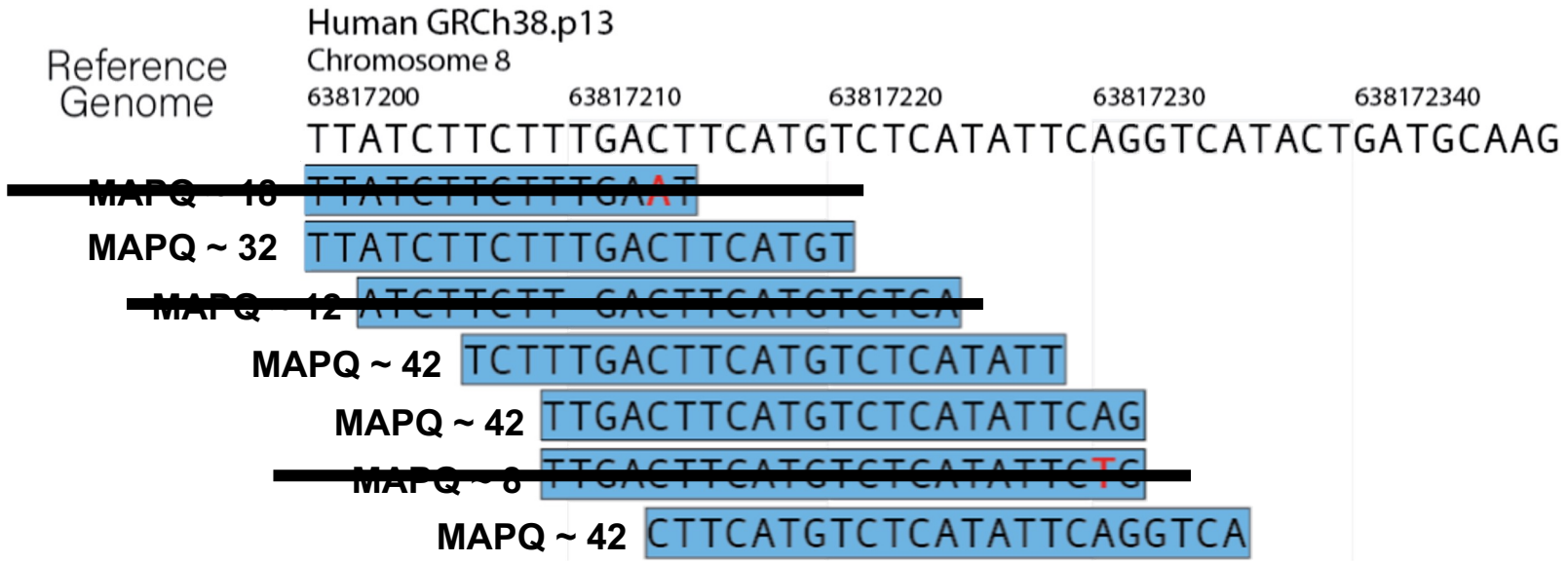


Filtering out low-quality mapping reads

Mapping Quality Scores (**MAPQ**) quantify the probability that a read is misplaced.

$$MAPQ = -10 * log_{10}\big(P(read\ is\ wrongly\ mapped)\big)$$

For example, a MAPQ score of 20 indicates that the probability for the read to be map at the indicated position is 0.01.

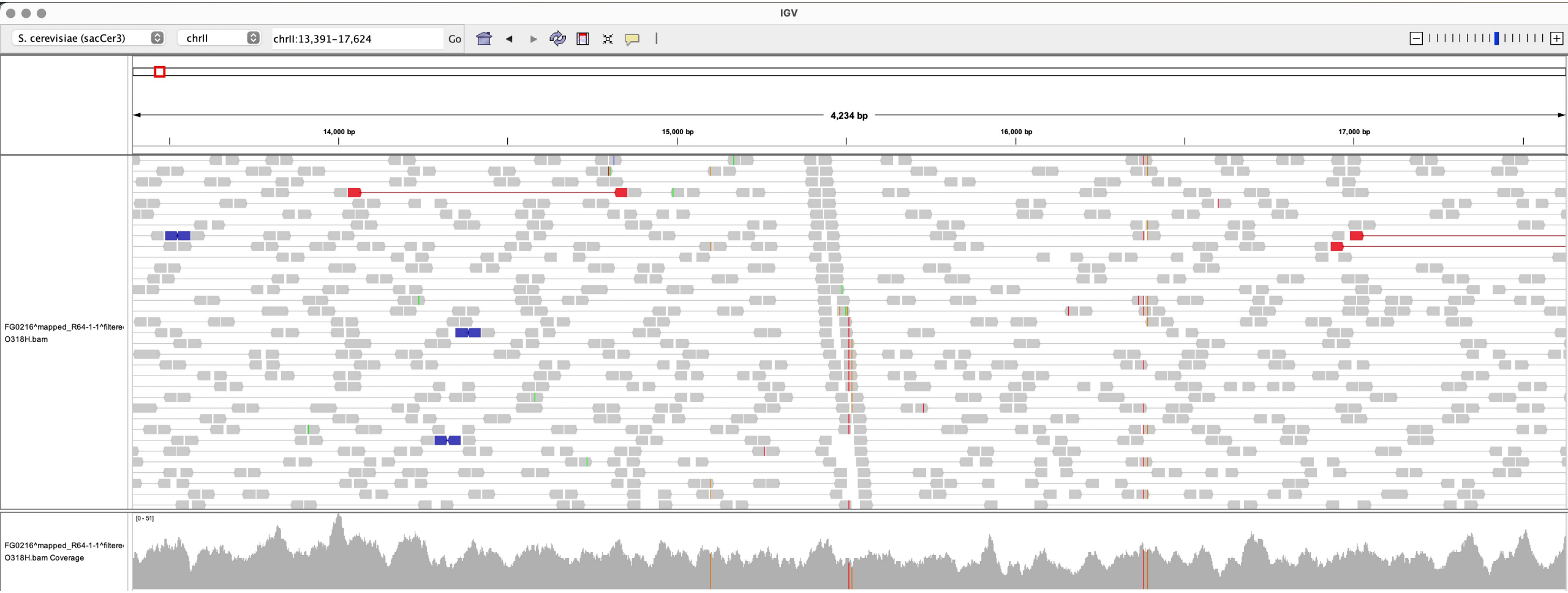


NGS processing workflow

- Get .bcl files 
- Create fastq files  Or **`bcl2fastq`**
- QC: remove/trim low quality reads E.g. **`cutadapt`**
- Align fastq to BAM E.g. **`bowtie2`**
- Filter duplicates, artifacts, … E.g. **`samtools`**
- Generate tracks
- Assay-specific downstream analysis

Generating tracks from mapped reads

Basic approach: ”pile-up” of all the fragments in `bam` files to generate a **coverage track**.



deepTools: a software suite to manage/produce genomic tracks

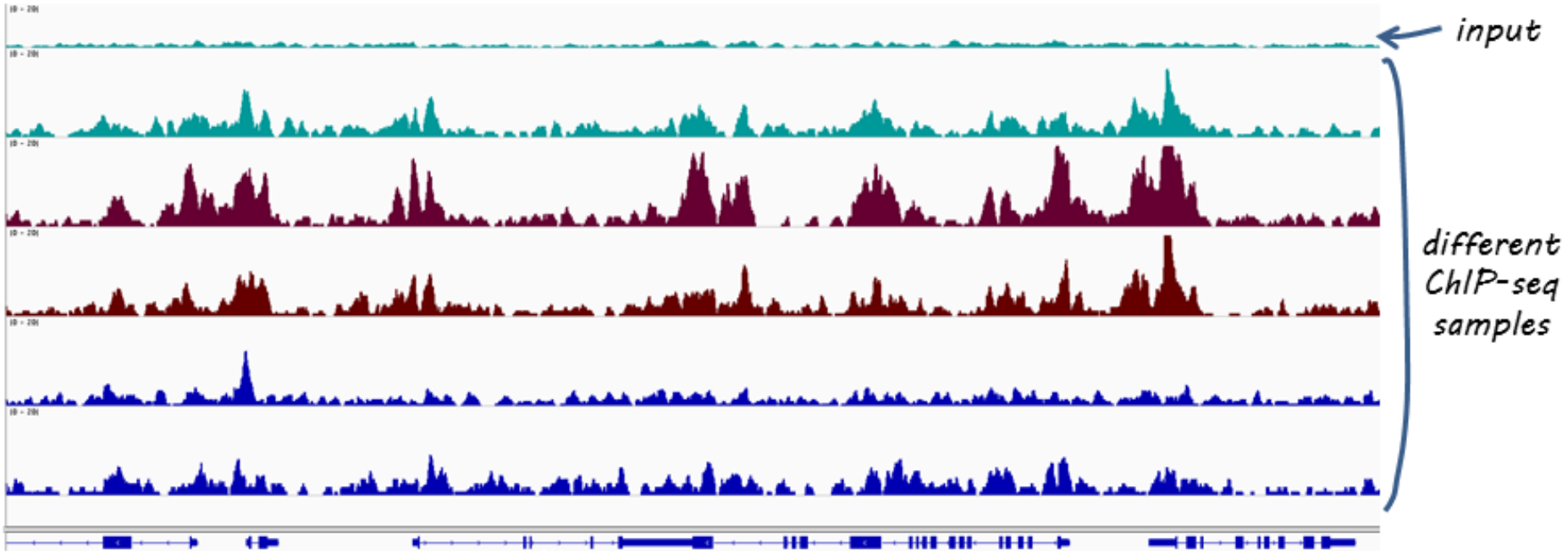
https://deeptools.readthedocs.io/en/latest/content/list_of_tools.html

- Tools for BAM and bigWig file processing
  - multiBamSummary
  - multiBigwigSummary
  - correctGCBias
  - bamCoverage
  - bamCompare
  - bigwigCompare
  - bigwigAverage
  - computeMatrix
  - alignmentSieve

- Tools for QC
  - plotCorrelation
  - plotPCA
  - plotFingerprint
  - bamPEFragmentSize
  - computeGCBias
  - plotCoverage
- Heatmaps and summary plots
  - plotHeatmap
  - plotProfile
  - plotEnrichment
- Miscellaneous
  - computeMatrixOperations
  - estimateReadFiltering

deepTools: a software suite to manage/produce genomic tracks

for visualizing continuous data, e·g· in the UCSC Genome Browser or IGV, bigWig files come in really handy



remember that there are 2 deepTools for bam → bigWig conversion:

- ***bamCoverage:*** *for individual files (like those shown here)*
- ***bamCompare:*** *to normalize two files to each other*

NGS processing workflow

- Get .bcl files 
- Create fastq files  Or **`bcl2fastq`**
- QC: remove/trim low quality reads — E.g. **`cutadapt`**
- Align fastq to BAM — E.g. **`bowtie2`**
- Filter duplicates, artifacts, … — E.g. **`samtools`**
- Generate tracks — E.g. **`deepTools`**
- Assay-specific downstream analysis

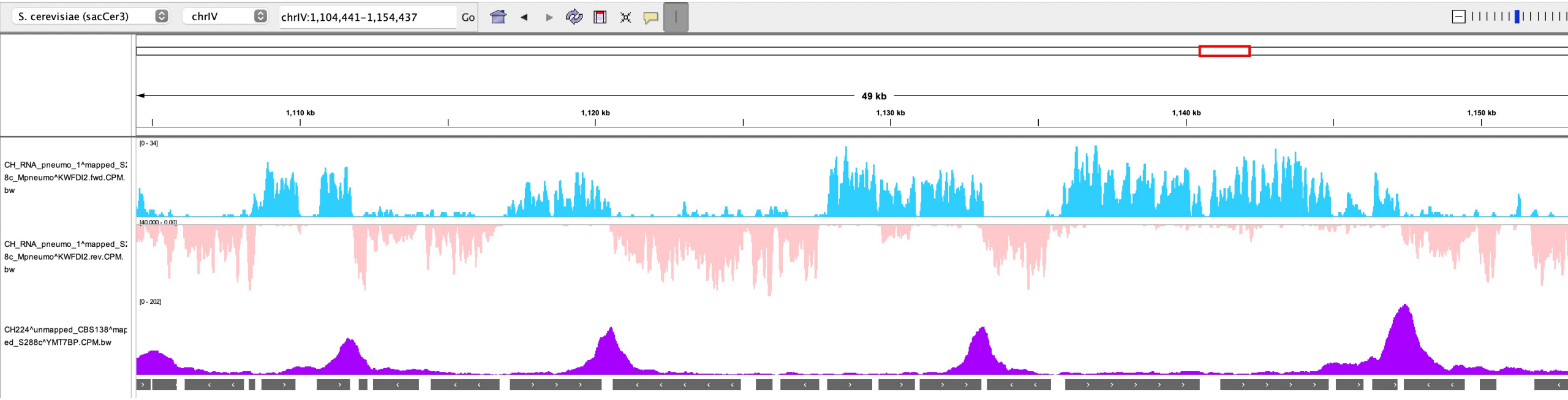
NGS processing workflow

- Get .bcl files 
- Create fastq files  Or **`bcl2fastq`**
- QC: remove/trim low quality reads — E.g. **`cutadapt`**
- Align fastq to BAM — E.g. **`bowtie2`**
- Filter duplicates, artifacts, … — E.g. **`samtools`**
- Generate tracks — E.g. **`deepTools`**

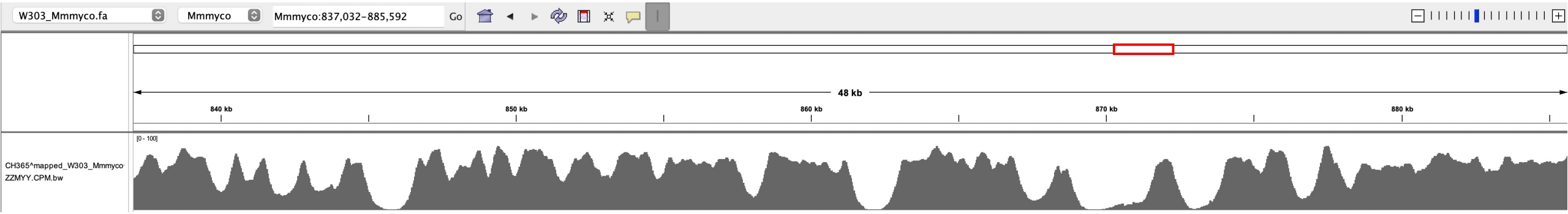
CHECK YOUR DATA

- Assay-specific downstream analysis

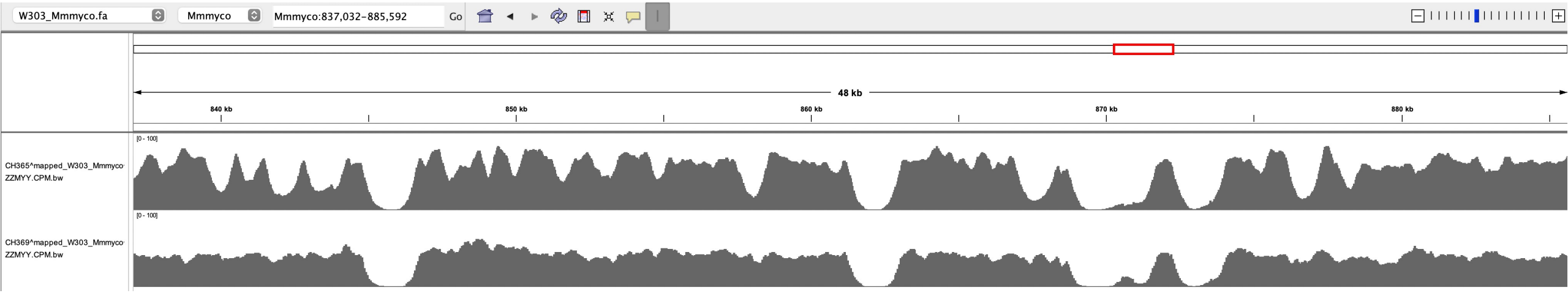
Watch out for artefacts



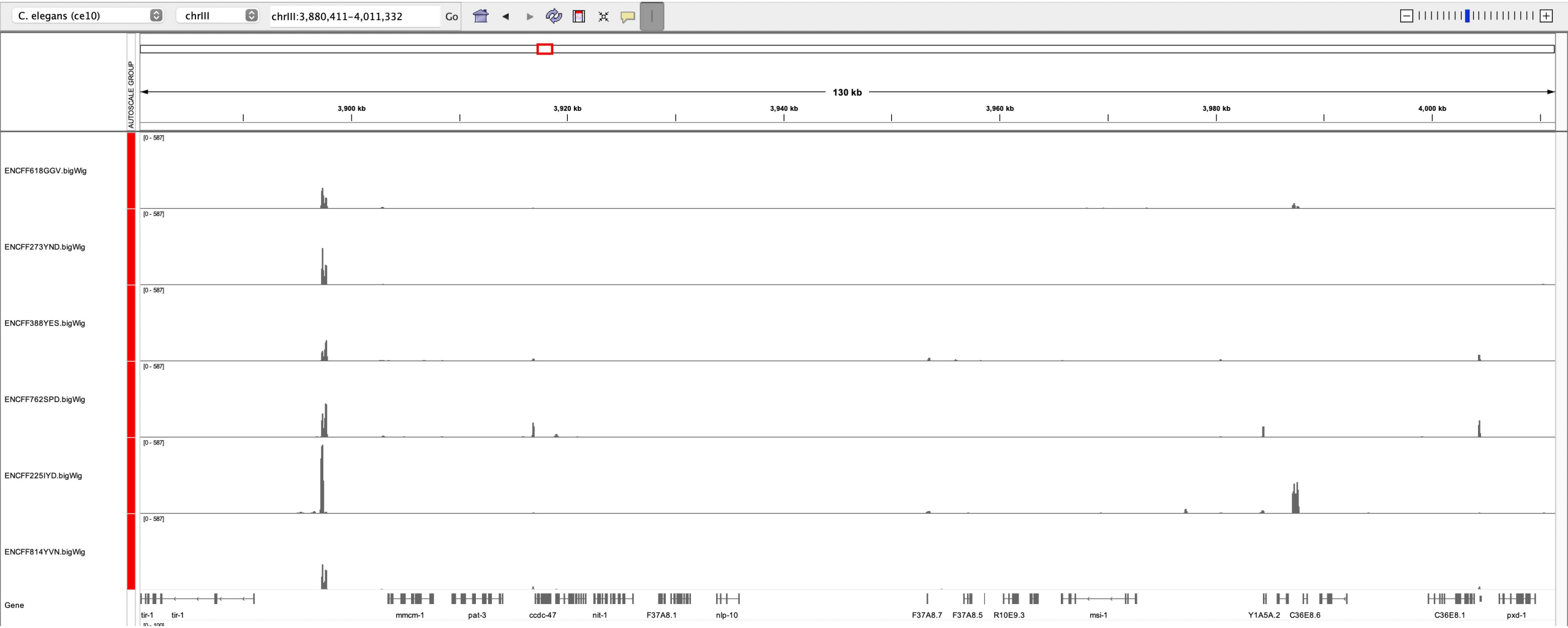
Watch out for artefacts



Watch out for artefacts



Watch out for artefacts



Watch out for artefacts

Genome-wide identification and characterisation of HOT regions in the human genome

Hao Li, Feng Liu, Chao Ren, Xiaochen Bo & Wenjie Shu

BMC Genomics **17**, Article number: 733 (2016) | Cite this article

Regulatory analysis of the *C. elegans* genome with spatiotemporal resolution

Carlos L. Araya, Trupti Kawli, Anshul Kundaje, Lixia Jiang, Beijing Wu, Dionne Vafeados, Robert Terrell, Peter Weissdepp, Louis Gevirtzman, Daniel Mace, Wei Niu, Alan P. Boyle, Dan Xie, Lijia Ma, John I. Murray, Valerie Reinke, Robert H. Waterston & Michael Snyder

Nature **512**, 400–405 (2014) | Cite this article

HOT regions are dynamic in development

JOURNAL ARTICLE

HOT or not: examining the basis of high-occupancy target regions

Katarzyna Wreczycka, Vedran Franke, Bora Uyar, Ricardo Wurmus, Selman Bulut, Baris Tursun, Altuna Akalin ✉ Author Notes

Nucleic Acids Research, Volume 47, Issue 11, 20 June 2019, Pages 5735–5745, https://doi.org/10.1093/nar/gkz460

Extreme HOT regions are CpG-dense promoters in *C. elegans* and humans

Ron A.-J. Chen, Przemyslaw Stempor, Thomas A. Down, Eva Zeiser, Sky K. Feuer, and Julie Ahringer[1]

The Gurdon Institute and Department of Genetics, University of Cambridge, Cambridge CB3 0DH, United Kingdom