Mapping and visualization of ChIP-seq data

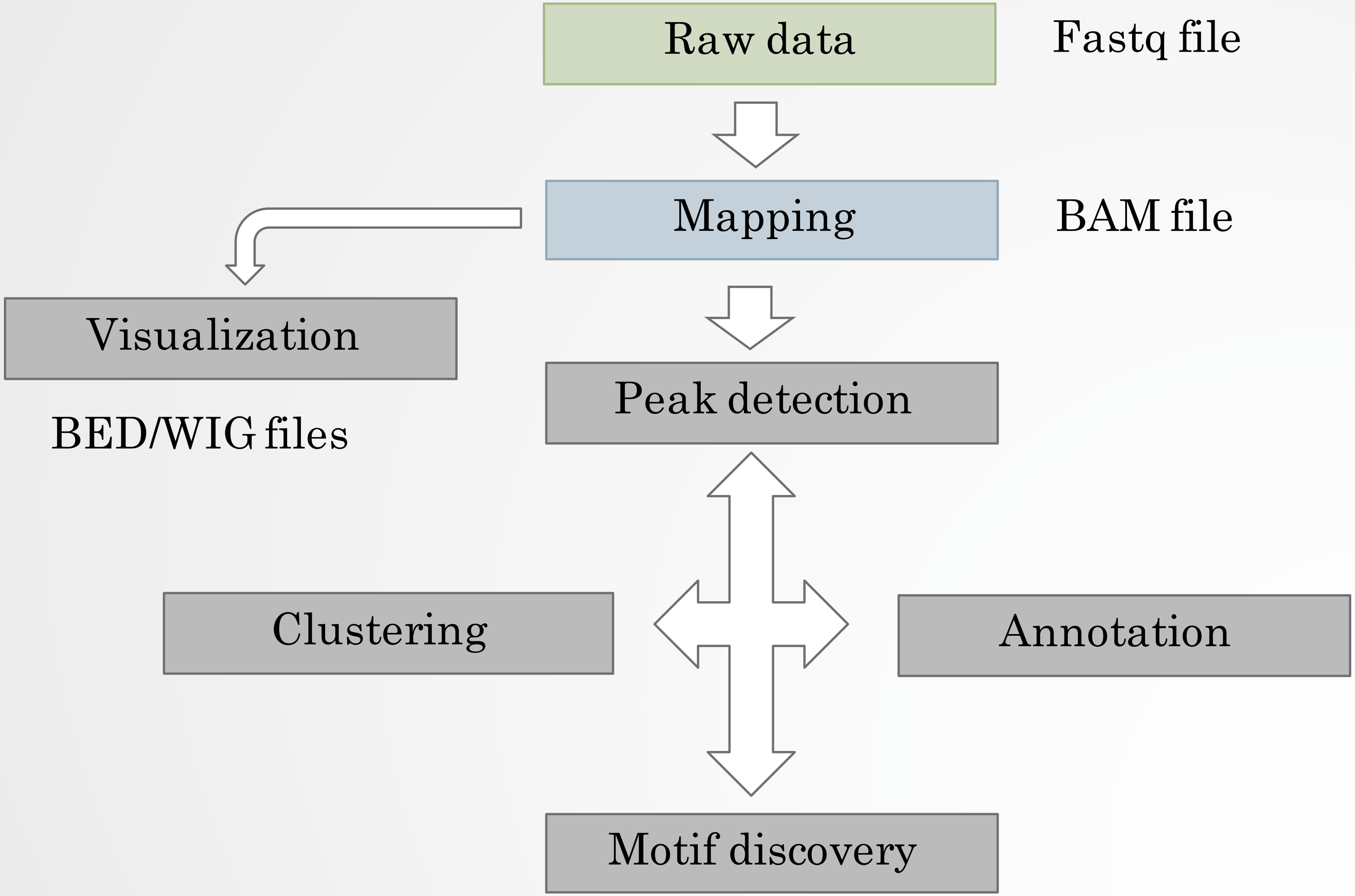
Stéphanie Le Gras
(slegras@igbmc.fr)

Exercise 1: Upload the data in a Genome Browser

We want to check that the IP worked i.e some regions are enriched in reads compared to the control sample

- 1. Upload the wig files (mitf.wig.gz, ctrl.wig.gz) from chipseq > visualization to UCSC

Guidelines



Raw data

Fastq file

Mapping

BAM file

Visualization

BED/WIG files

Peak detection

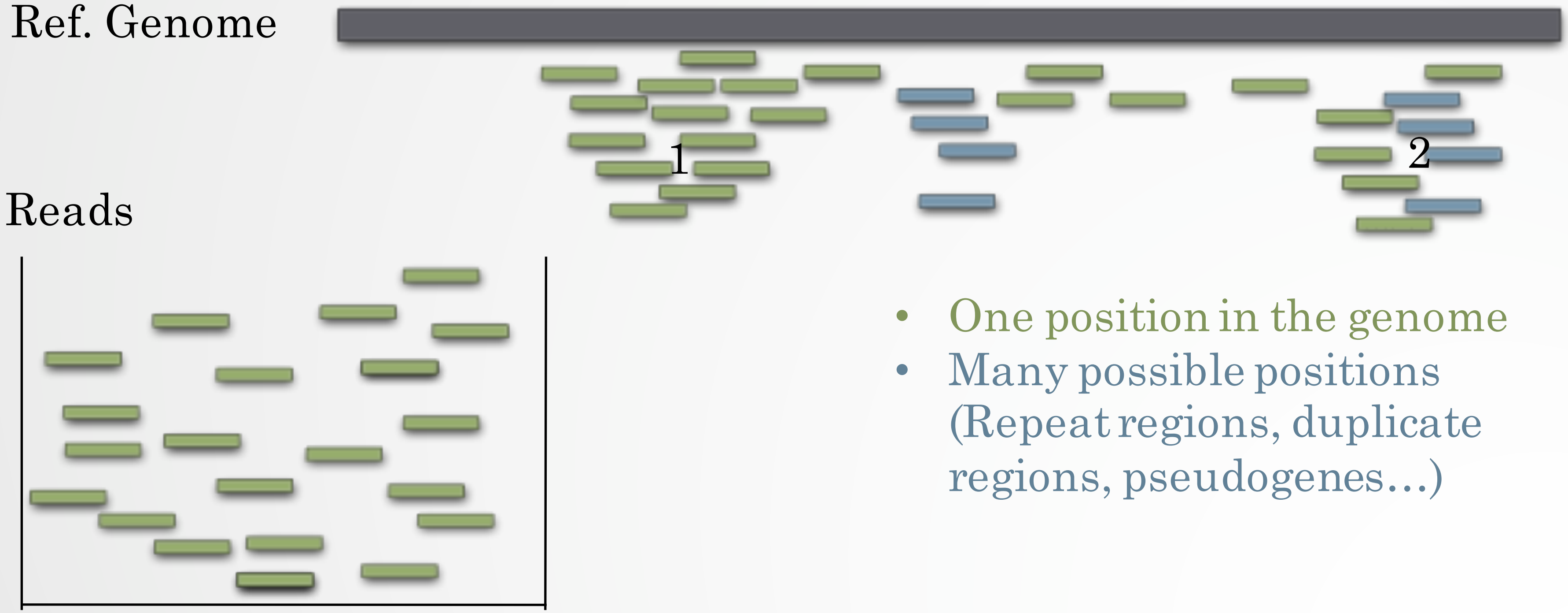
Clustering

Annotation

Motif discovery

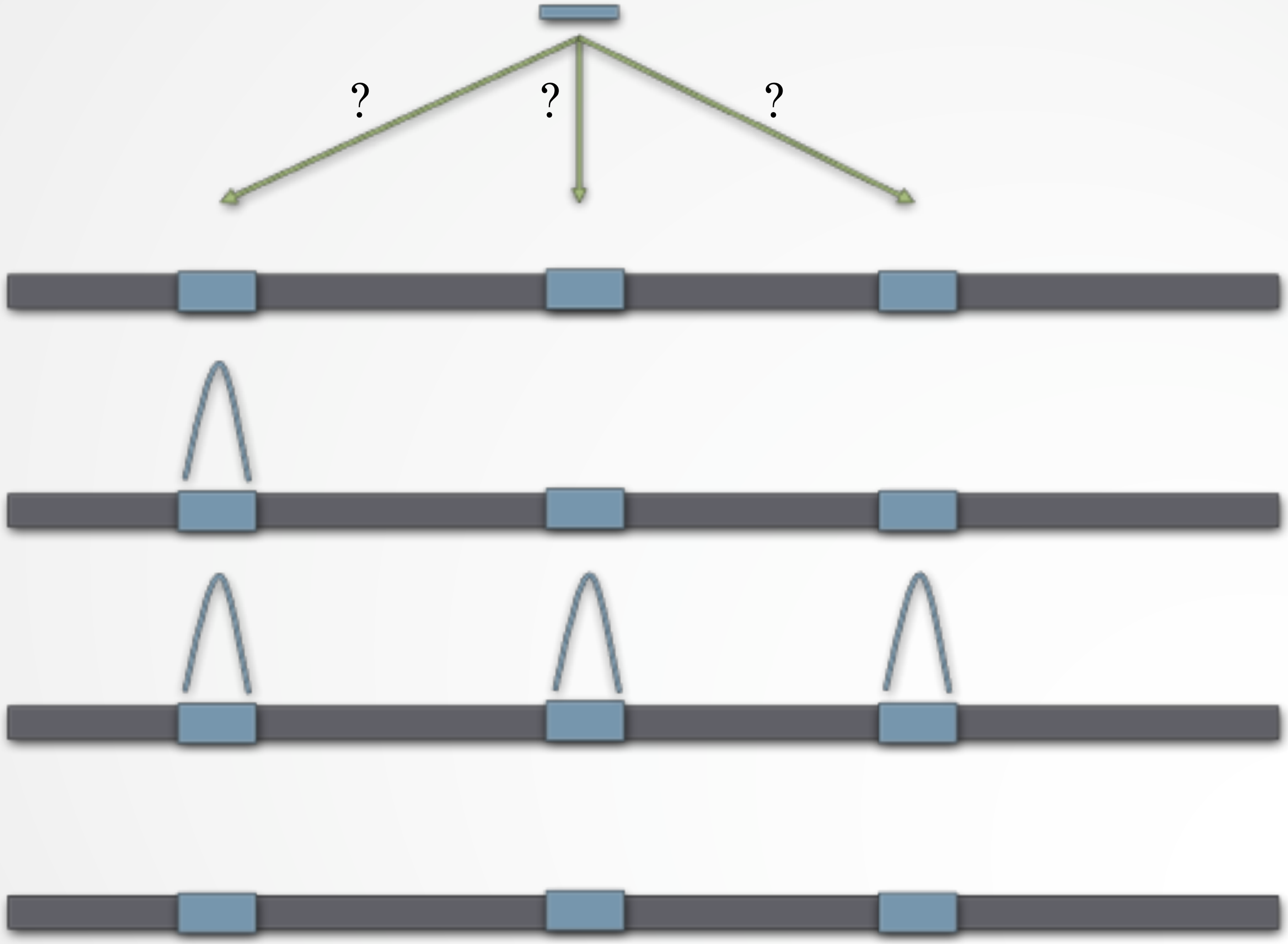
Mapping

- Find out the position of the reads within the genome



- One position in the genome
- Many possible positions (Repeat regions, duplicate regions, pseudogenes…)

Duplicated genomic regions



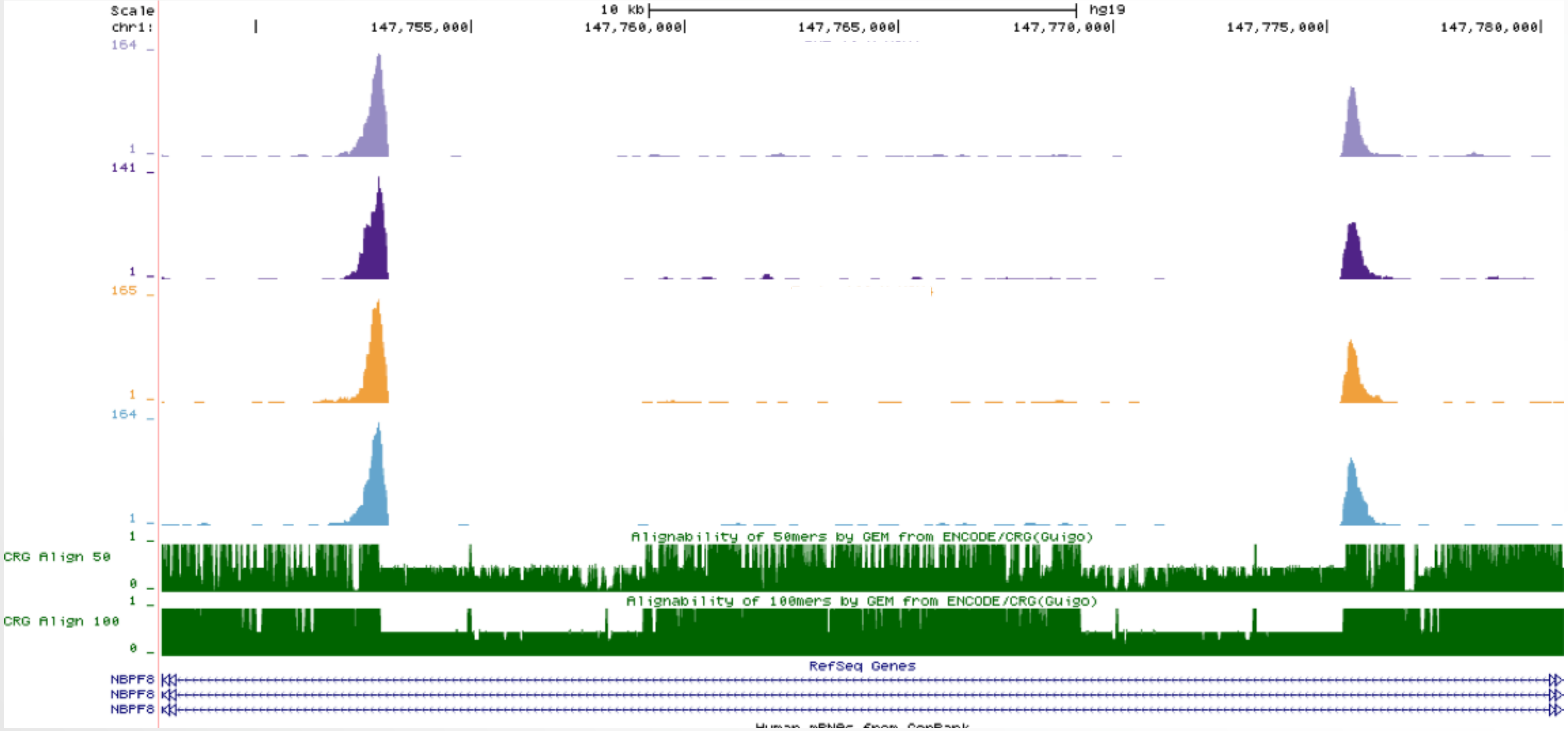
Keep 1 position randomly

Keep all possible position

Keep none

Mappability

- Mappability (a): how many times a read of a given length can align at a given position in the genome
  - a=1 (read align once)
  - a=1/n (read align n times)
  - Regions are empty or poorly covered if the mappability is low



Exercise 2: mapping statistics

We used Bowtie 1 with the following parameters "-m 1 --strata --best" to align the reads. How many reads are aligned for each of the samples?

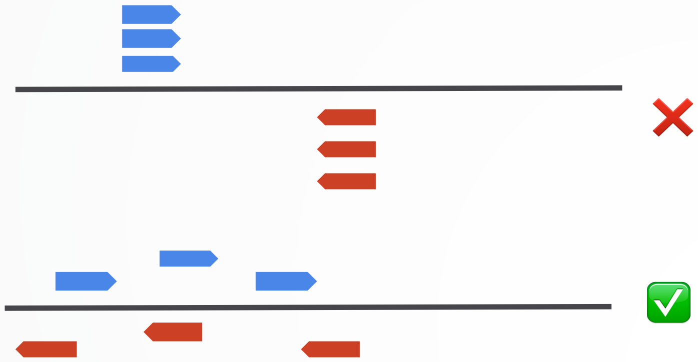
- 1. go to GalaxEast (http://use.galaxeast.fr/)
- 2. create a new history named "ChIP-seq data analysis"
- 3. import 2 BAM files (mitf.bam and ctrl.bam) from the data library CNRS training > ChIPseq > mapping
- 4. use the tool **Flagstat** from the "NGS: Sam Tools" section to compute the number of aligned reads in the samples. The tools gives alignment statistics on a BAM file.

PCR duplicates

- PCR duplicates
    - Related to poor library complexity
    - The same set of fragments are amplified
        - Indicates that Immuno-precipitation failed
    - Tools to check for
        - FastQC report (duplicate diagram)
        - PCR bottleneck metric (ENCODE)

QC : PBC (PCR bottleneck coefficient)

- An approximate measure of library complexity
- PBC = N1/Nd
  - N1= Genomic position with 1 read aligned
  - Nd = Genomic position with ≧ 1 read aligned
- Value :
  - 0-0.5: severe bottlenecking (PCR bias, or a biological finding, such as a very rare genomic feature)
  - 0.5-0.8: moderate bottlenecking
  - 0.8-0.9: mild bottlenecking
  - 0.9-1.0: no bottlenecking (Control or IP with a good library complexity)



https://genome.ucsc.edu/ENCODE/qualityMetrics.html

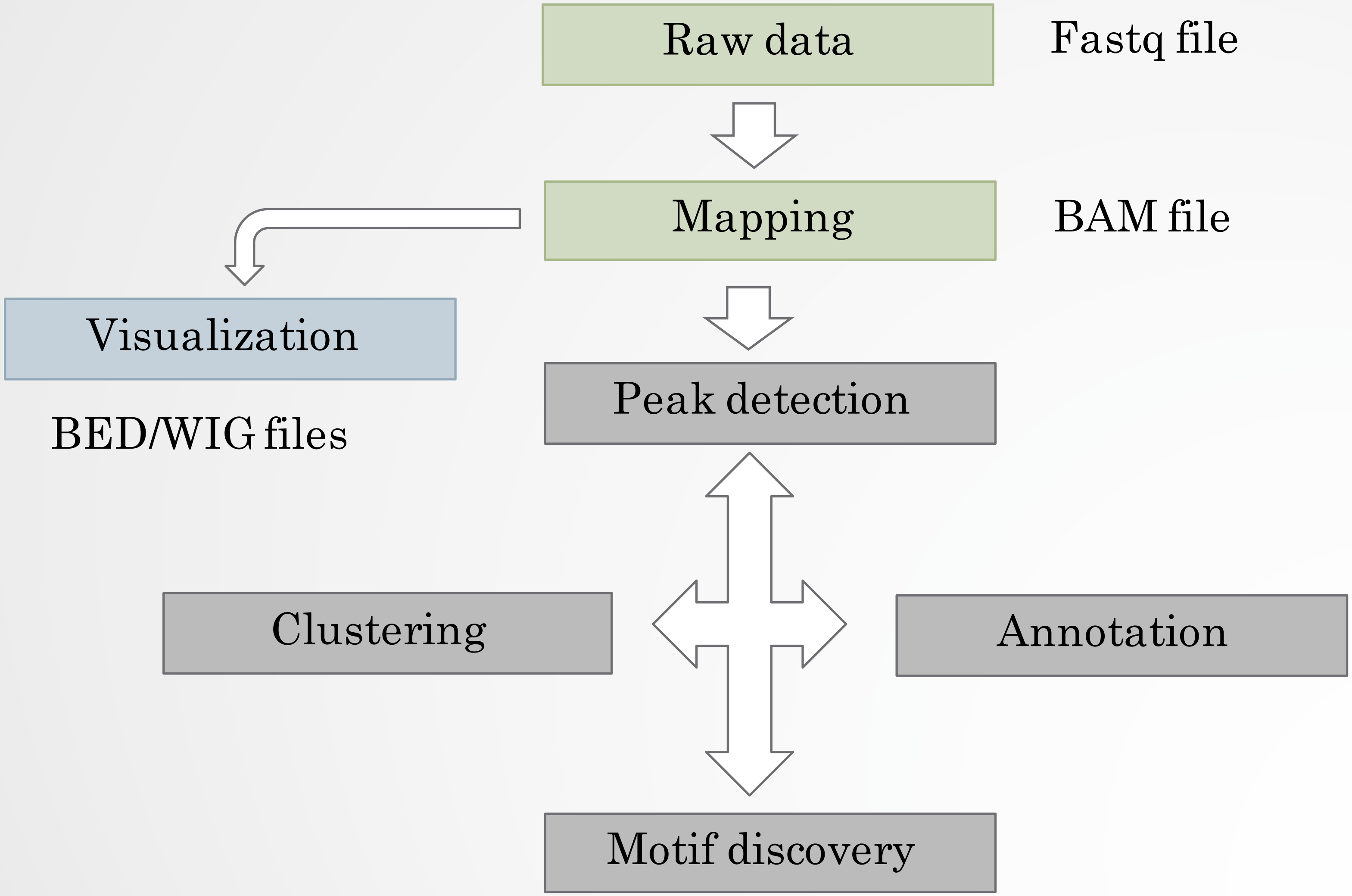
Exercise 3: duplicate reads estimate

We want to assess the number of duplicate reads

1. Use the tool **MarkDuplicates** to assess the complexity of the libraries (i.e the number of unique sequences). Use default parameters except for:
   - Select validation stringency: Silent (The picard tools validation strategy of BAM file is very stringent. So we turn off validation stringency)

   - The tool generates two datasets:
     - A log/metric file that contains statistics on the tool processing (number of input reads, number of duplicate reads)
     - A BAM file in which duplicated reads are flagged

Guidelines



Raw data

Fastq file

Mapping

BAM file

Visualization

BED/WIG files

Peak detection

Clustering

Annotation

Motif discovery

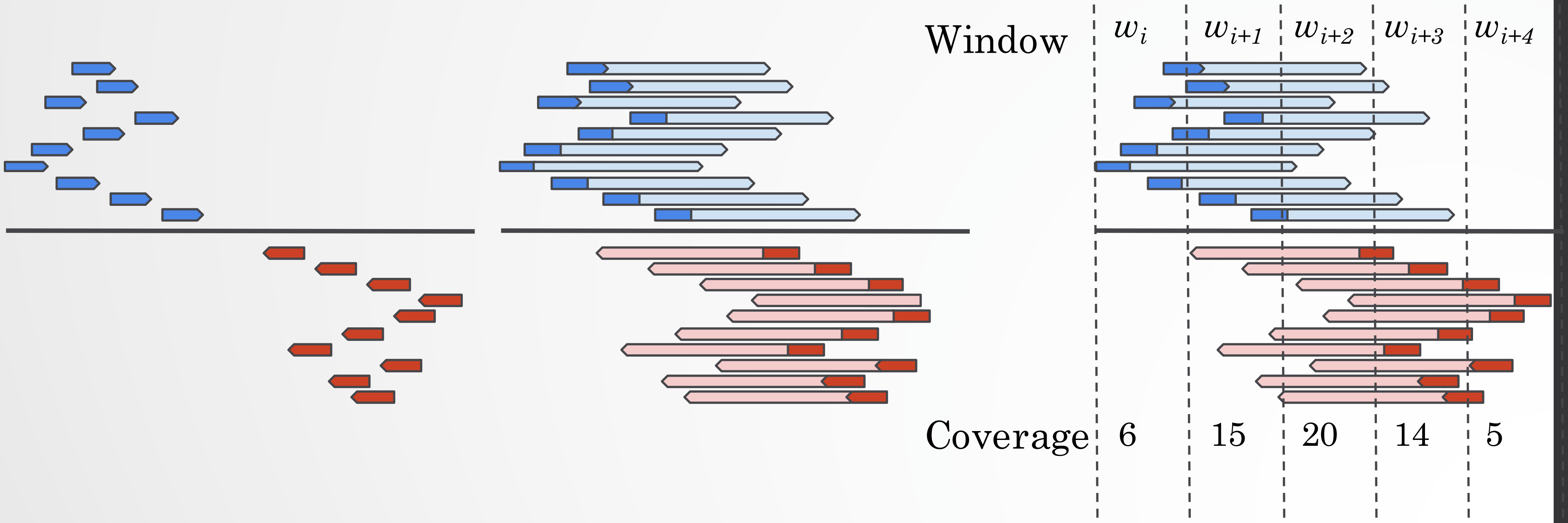
Bam files are fat

- **BAM files are fat** as they do contain exhaustive information about read alignments.
  - Memory issues (can only visualize fraction of the BAM).

- Need a more **lightweight file format containing only genomic coverage information:**
  - ❌ **Wig (not compressed, not indexed)**
  - ✅ **TDF (compressed, indexed)**
  - ✅ **BigWig (compressed, indexed)**

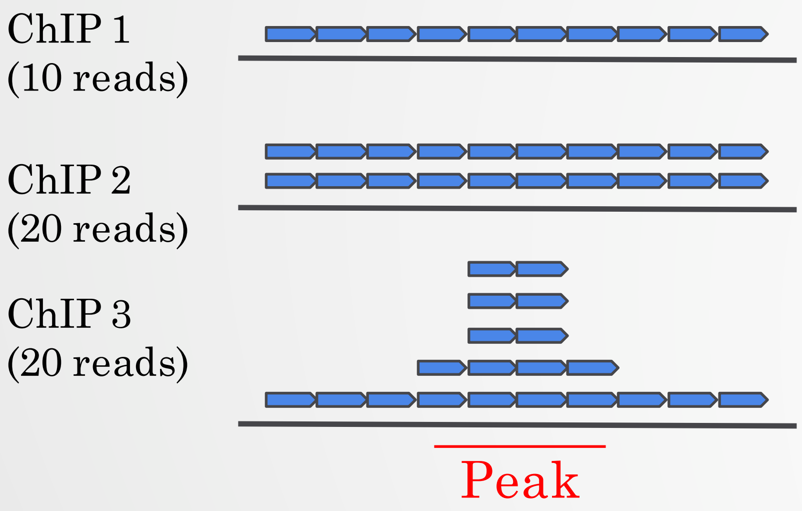
Coverage file and read extension

- BAM files **do not contain fragment location** but read location
- We need to extend reads to compute fragments coordinates before coverage analysis
- Not required for PE



Library size normalization

- **Signal need to be normalized**
  - E.g. Normalize coverage to 1x
    - Popular but not optimal



✅ **Already normalized to 1x coverage**

✅ **Should be decreased by 2 fold to get 1x coverage**

❌ **Decreasing by 2 fold would underestimate peak signal. Problem …**

UCSC

- https://genome.ucsc.edu/
- Online Genome Browser
- Hosted by the University of California, Santa Cruz
- Offers access to genome sequence data from a variety of vertebrate and invertebrate species and major model organisms
- Easy browsing
- Easy to display/hide tracks
- Easy upload of your data
- Lot of external data available (ENCODE, Ensembl…)
- Linked to many external tools (Galaxy, GREAT…)
- Useful tools (BLAT, table browser, « get DNA »,…)
- Best for chIP-seq data

Exercise 4: Visualization of the data

Go to UCSC and look at the datasets to check whether the IP worked.

- 1. Go to check the genes:
  - ANKRD30BL
  - CFAP221
  - DBI

  Do you see peaks at these locations?