Module 1/6: Analyses ADN

- NGS Introduction
- Reads Quality Control
- Reads Cleaning
- Aligning reads on reference → *Hélène Touzet*
- Alignment parameters → *Hélène Touzet*
- Reads duplicates
- Assembly → *Hélène Touzet*

Module 1/6: Analyses ADN

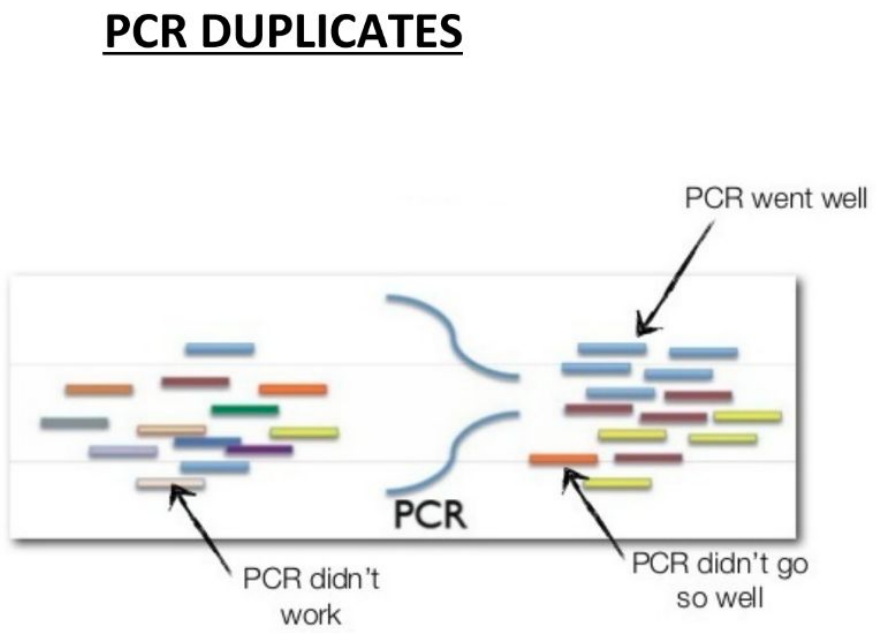
- NGS Introduction
- Reads Quality Control
- Reads Cleaning
- Aligning reads on reference → *Hélène Touzet*
- Alignment parameters → *Hélène Touzet*
- Reads duplicates
- Assembly → *Hélène Touzet*

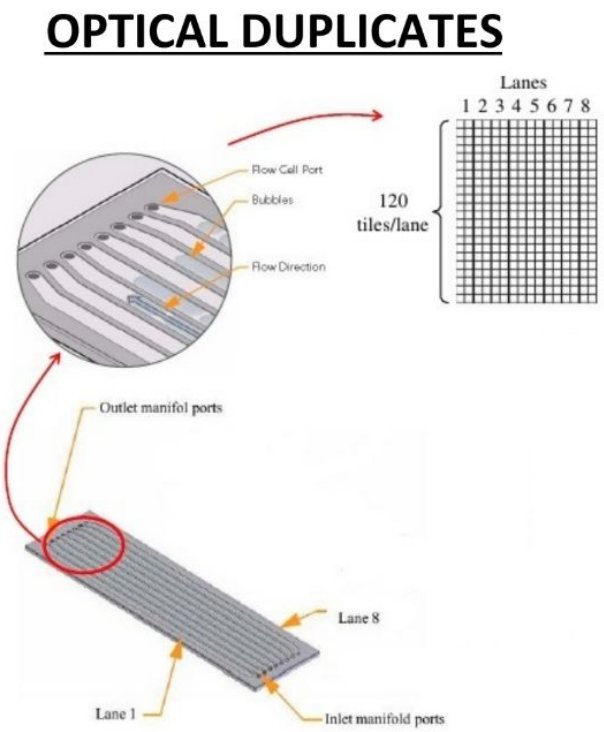
Reads alignment: reads duplicates

- Manage duplicated reads *(picard / MarkDuplicates)*
- Count alignments *(samtools / flagstat)*
- Compute depth and coverage *(Deeptools/PlotCoverage)*

Cleaning duplicated reads

How do duplication events arise?





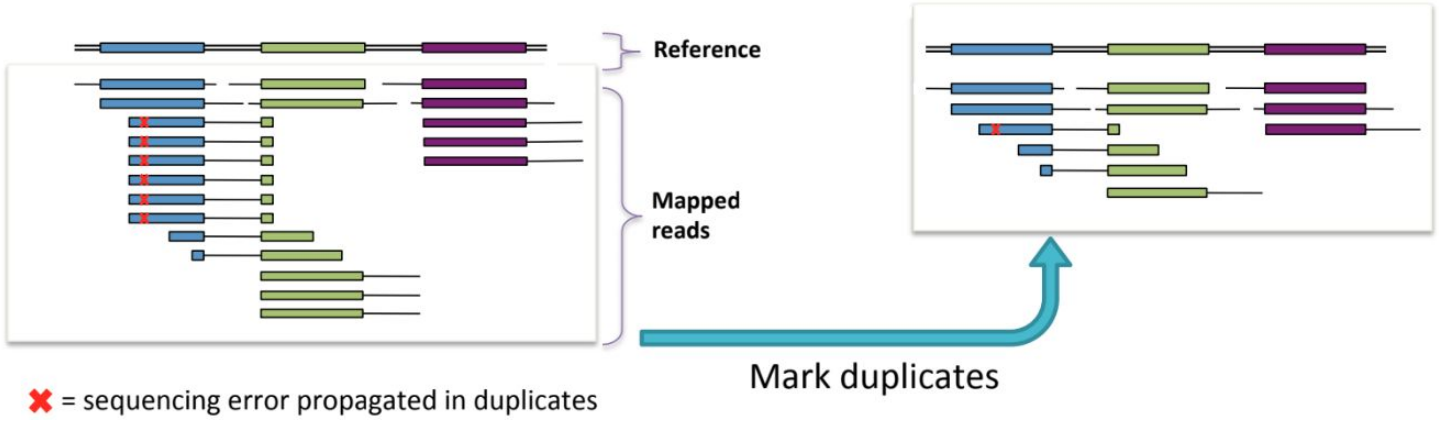
http://www.slideshare.net/jandot/next-generation-sequencing-course-part-2-sequence-mapping
http://www.slideshare.net/cosentia/illumina-gaiix-for-high-throughput-sequencing

source: GATK workshops, https://drive.google.com/drive/folders/0BwTg3aXzGxEDR2JKTUIxVGxCT1k

Cleaning duplicated reads

Why mark duplicates?

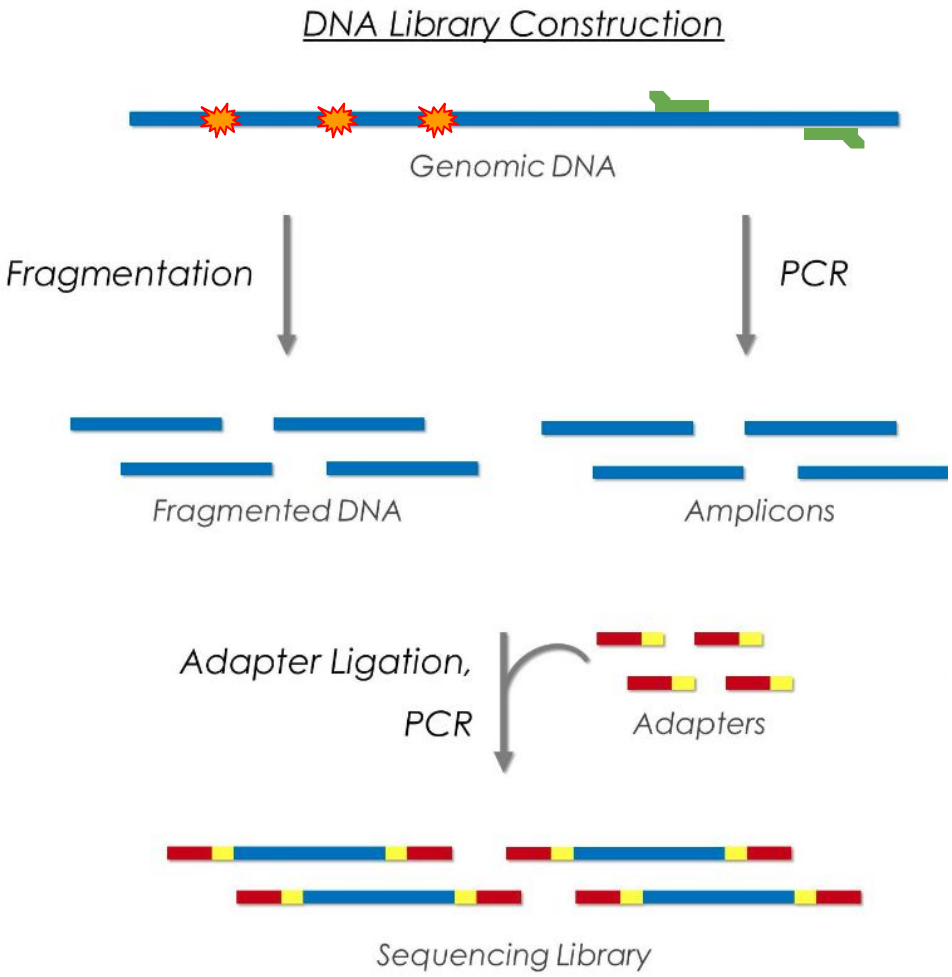
- Duplicates are sets of reads pairs that have the same unclipped alignment start and unclipped alignment end
- They're suspected to be **non-independent measurements** of a sequence
  - Sampled from the exact same template of DNA
  - Violates assumptions of variant calling
- What's more, errors in sample/library prep will get propagated to *all* the duplicates
  - Just pick the "best" copy – mitigates the effects of errors



source: GATK workshops, https://drive.google.com/drive/folders/0BwTg3aXzGxEDR2JKTUIxVGxCT1k

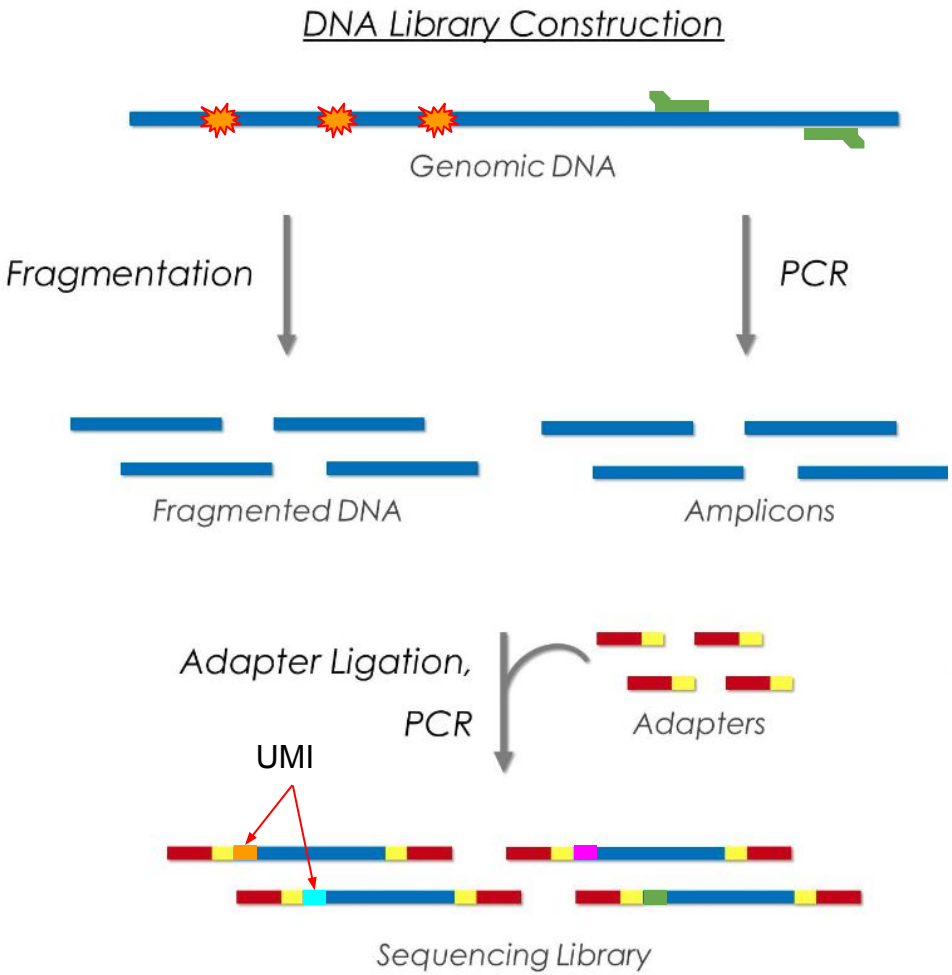
Cleaning duplicated reads

Molecular Barcoding (UMI, *unique molecular identifiers*)



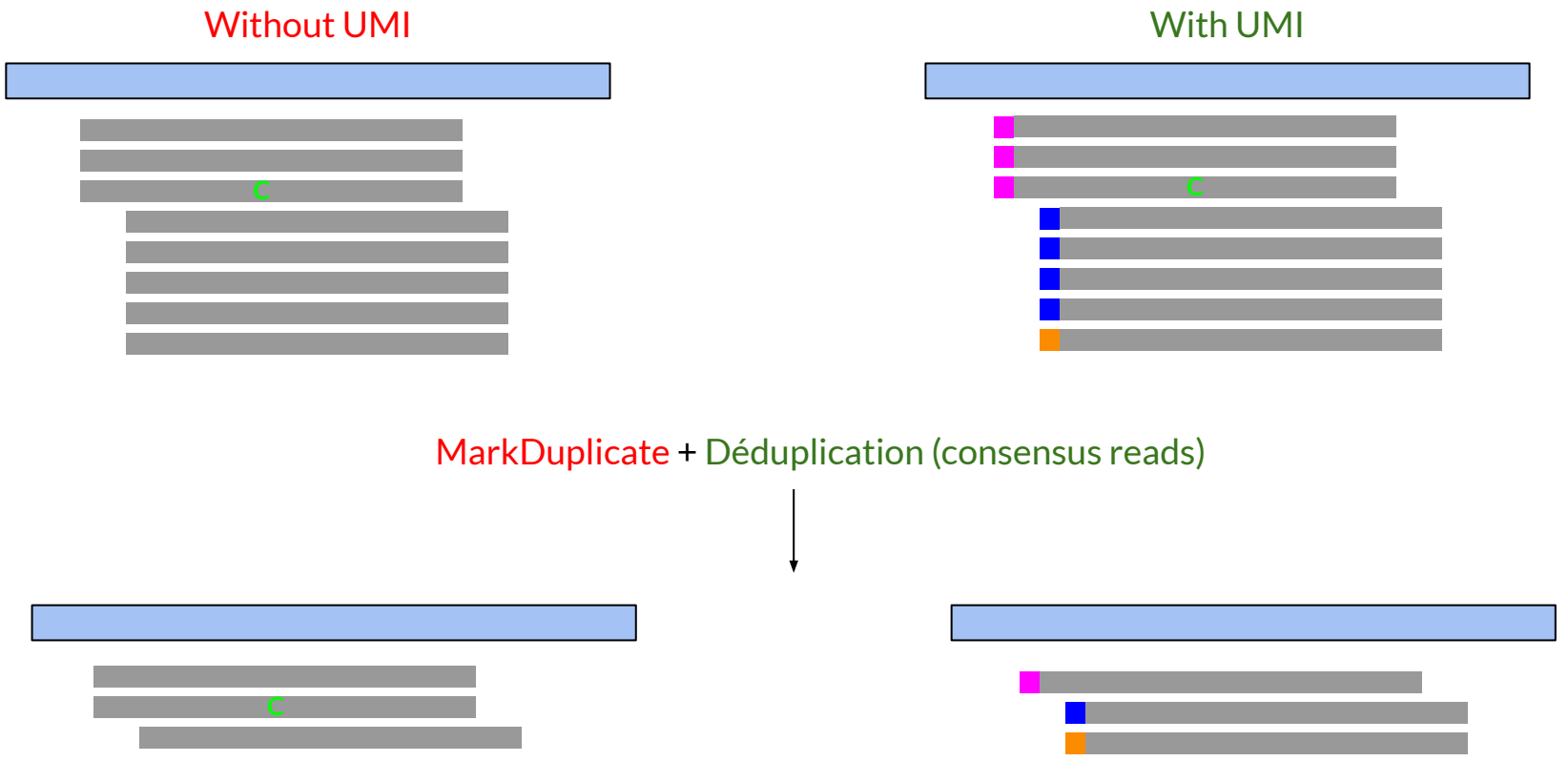
Cleaning duplicated reads

Molecular Barcoding (UMI, *unique molecular identifiers*)

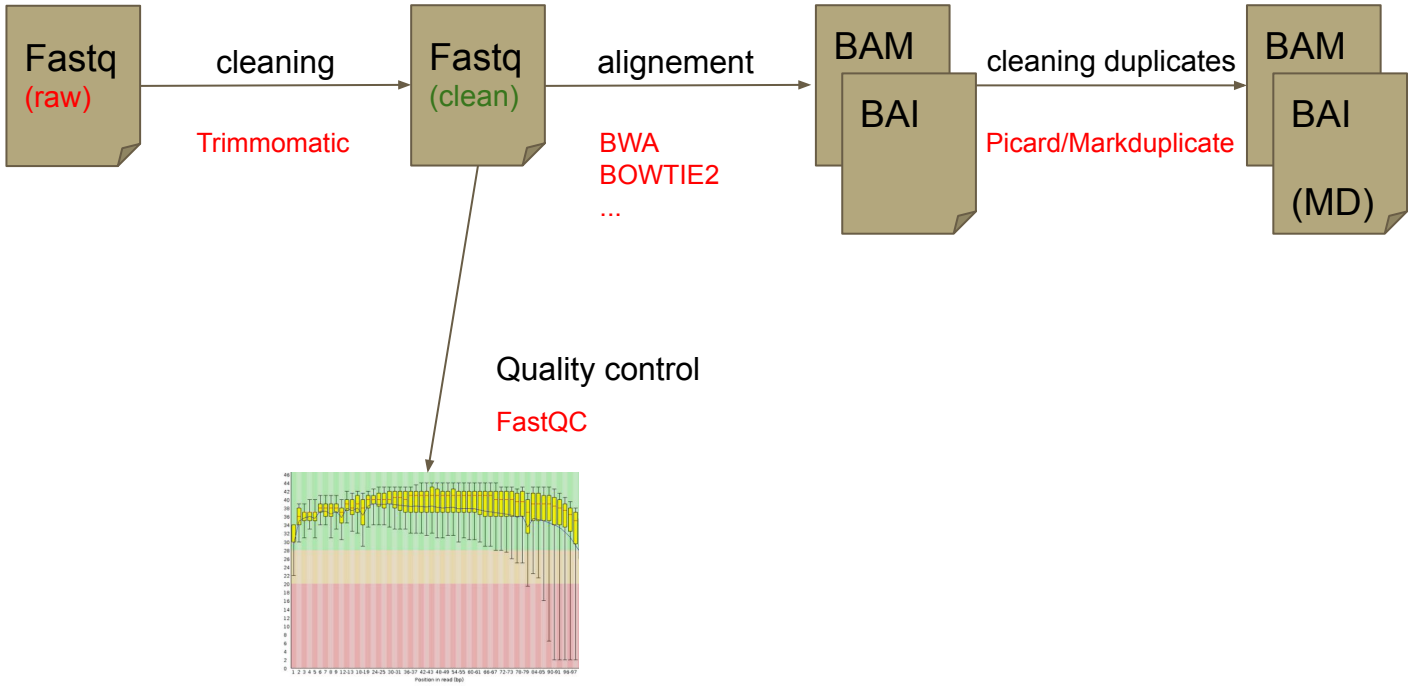


Cleaning duplicated reads

Molecular Barcoding (UMI, *unique molecular identifiers*)



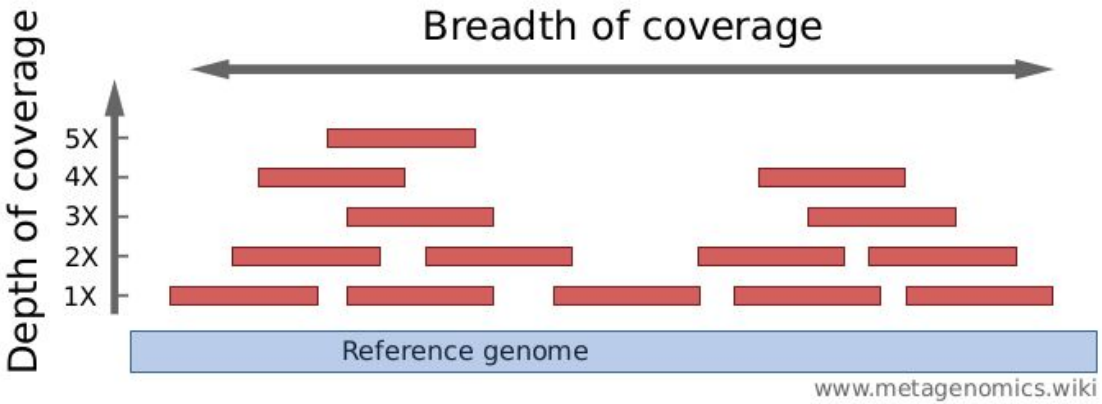
Workflow


Fastq
(raw)
cleaning
Trimmomatic
Fastq
(clean)
alignement
BWA
BOWTIE2
...
BAM
BAI
cleaning duplicates
Picard/Markduplicate
BAM
BAI
(MD)
Quality control
FastQC

Coverage and depth of coverage

- **Depth of coverage =** average number of reads covering a base (X)
  - Example: 30X for normal sample, 100X for tumor sample

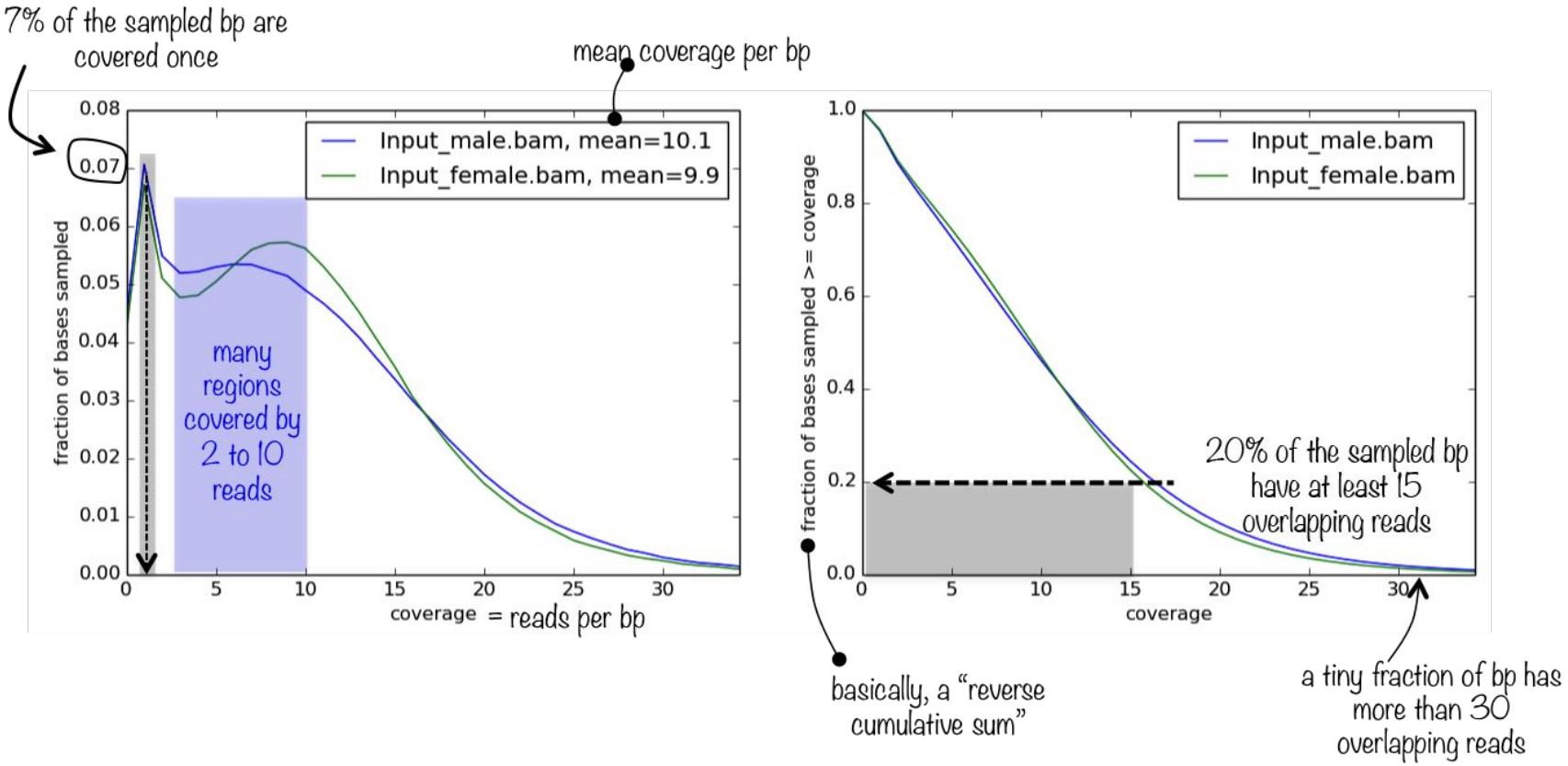
- **(Breadth of) Coverage =** percentage of the targeted regions covered by at least X read
  - For example: 90% of a genome is covered at 1X depth; and still 40% is covered at 4X depth.



Source :
- Élodie Girard , 5ème Ecole de bioinformatique AVIESAN-IFB 2016 ,http://www.france-bioinformatique.fr/sites/default/files/V01_ITMO_2016_EG_from_fastq_to_mapping_1.pdf
- http://www.metagenomics.wiki/pdf/definition/coverage-read-depth

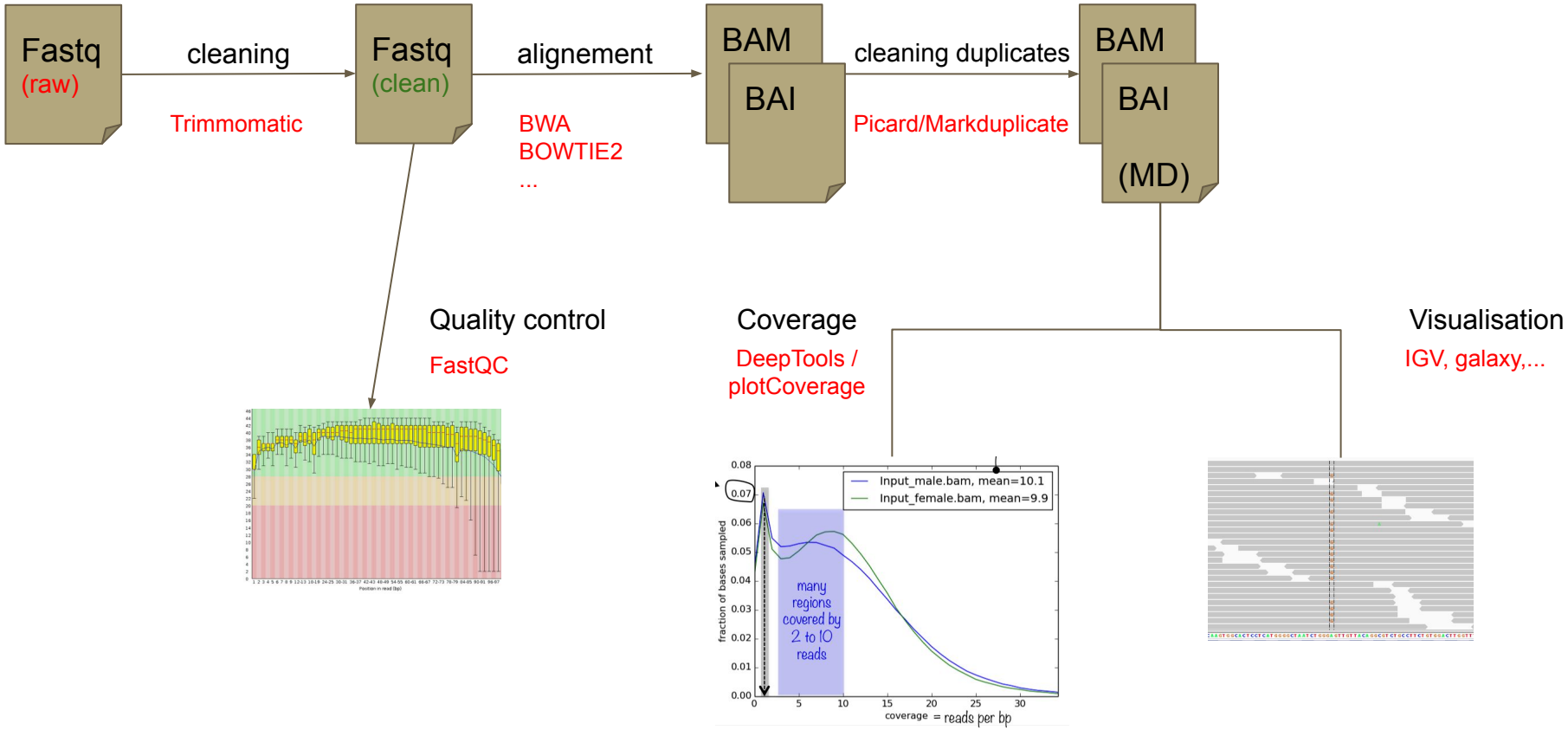
Computing coverage and depth of coverage

Tool: DeepTools / plotCoverage

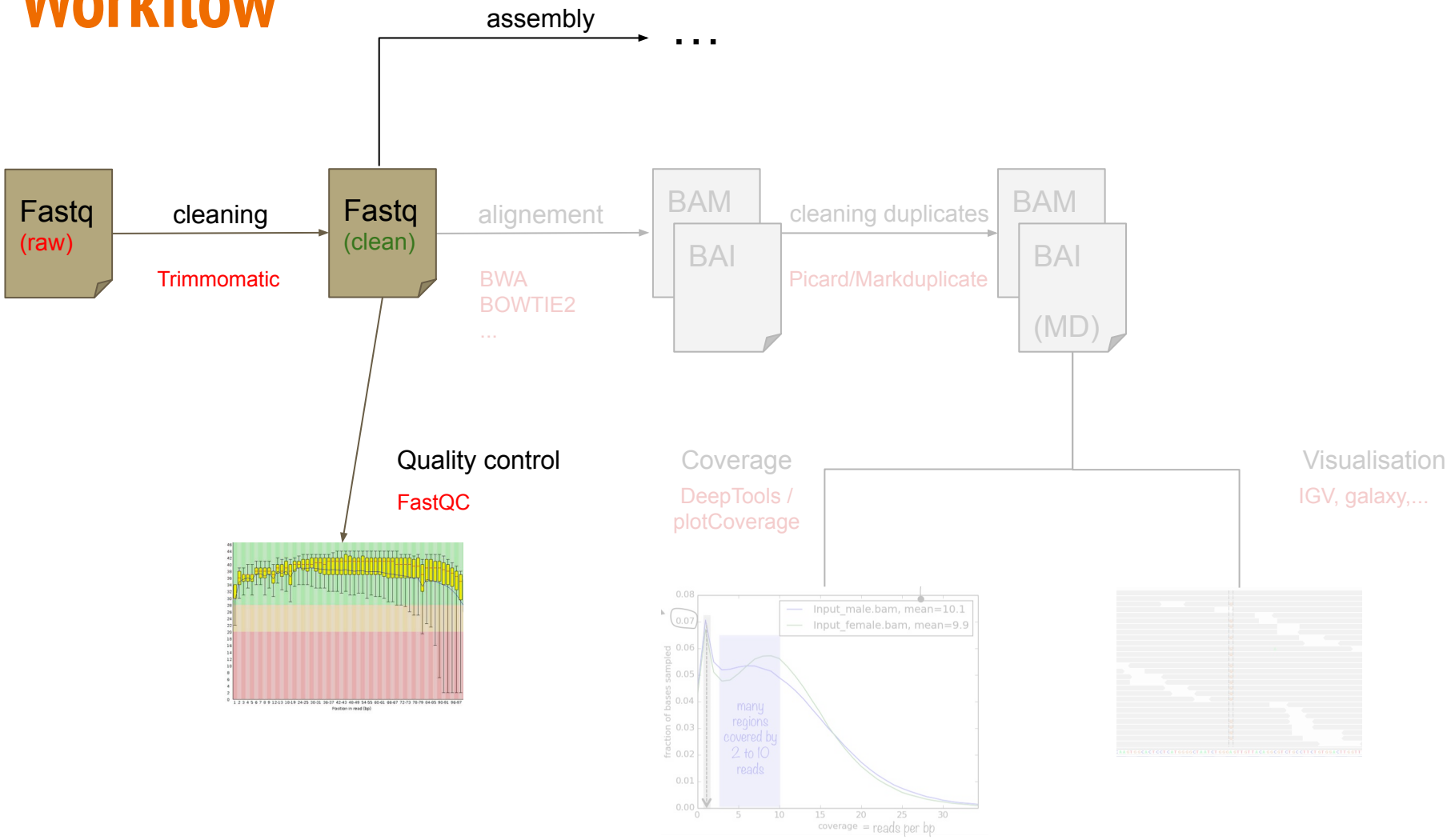


Ramírez, Fidel, Devon P. Ryan, Björn Grüning, Vivek Bhardwaj, Fabian Kilpert, Andreas S. Richter, Steffen Heyne, Friederike Dündar, and Thomas Manke. deepTools2: A next Generation Web Server for Deep-Sequencing Data Analysis. Nucleic Acids Research (2016). doi:10.1093/nar/gkw257.

Workflow


Fastq
(raw)
cleaning
Trimmomatic
Fastq
(clean)
alignement
BWA
BOWTIE2
...
BAM
BAI
cleaning duplicates
Picard/Markduplicate
BAM
BAI
(MD)
Quality control
FastQC
Coverage
DeepTools /
plotCoverage
Visualisation
IGV, galaxy,...
Input_male.bam, mean=10.1
Input_female.bam, mean=9.9
many regions covered by 2 to 10 reads
fraction of bases sampled
coverage = reads per bp

Workflow


assembly
...
Fastq
(raw)
cleaning
Trimmomatic
Fastq
(clean)
alignement
BWA
BOWTIE2
...
BAM
BAI
cleaning duplicates
Picard/Markduplicate
BAM
BAI
(MD)
Quality control
FastQC
Coverage
DeepTools /
plotCoverage
Visualisation
IGV, galaxy,...
Input_male.bam, mean=10.1
Input_female.bam, mean=9.9
many
regions
covered by
2 to 10
reads
fraction of bases sampled
coverage = reads per bp