

# Module 1/5: 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 1/5: Analyses ADN

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

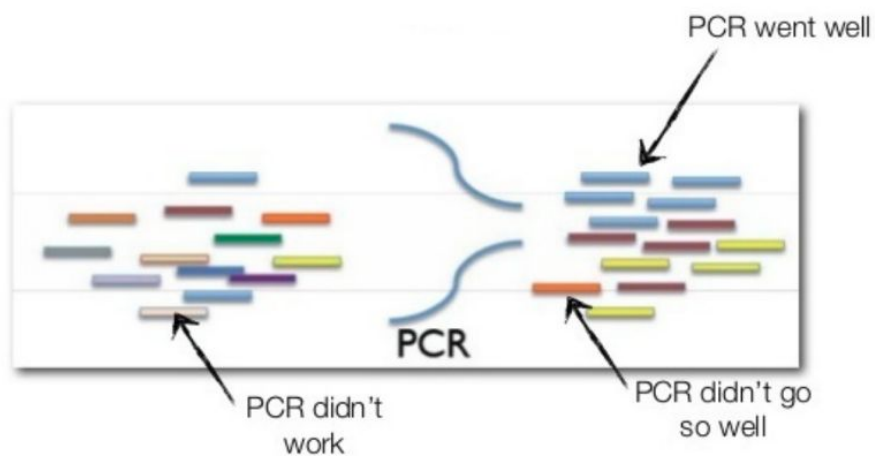
# Reads alignment: reads duplicates

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

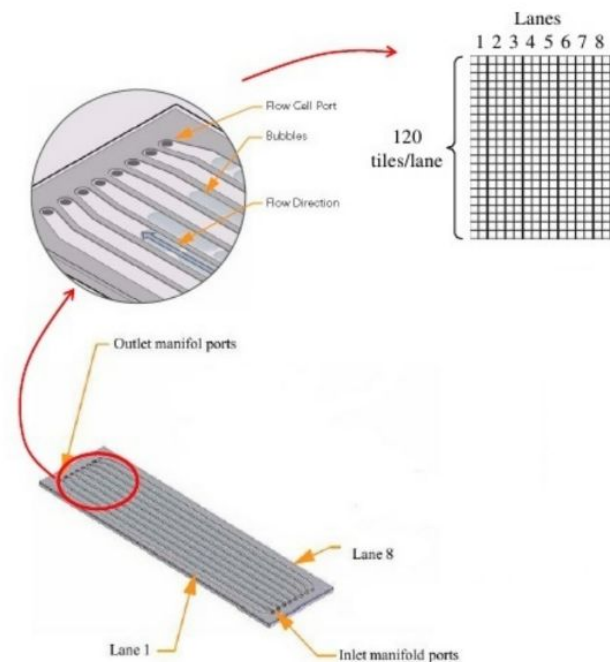
# Cleaning duplicated reads

How do duplication events arise?

## PCR DUPLICATES



## OPTICAL DUPLICATES

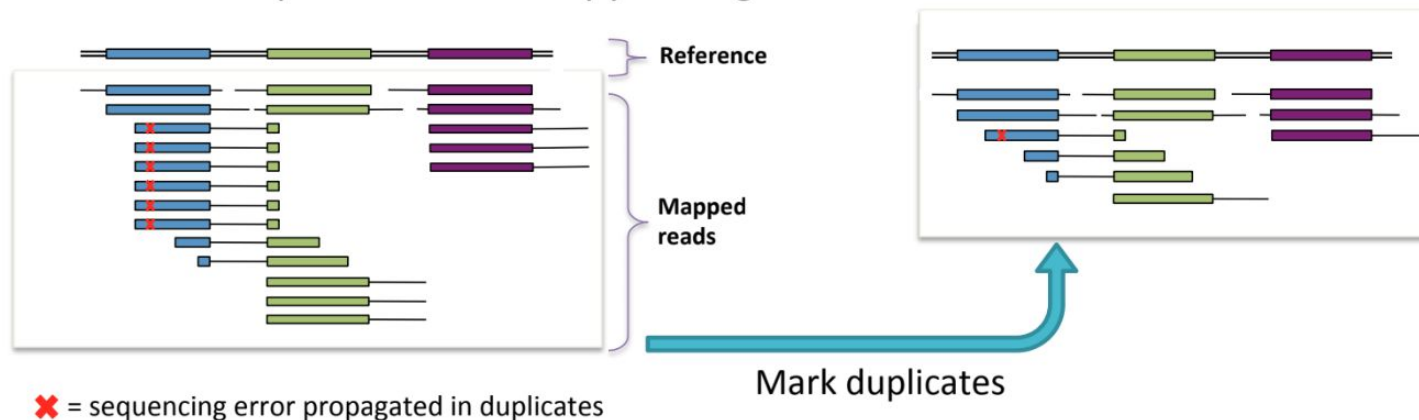


<http://www.slideshare.net/andot/next-generation-sequencing-course-part-2-sequence-mapping>  
<http://www.slideshare.net/rosentia/illumina-galix-for-high-throughput-sequencing>

# Cleaning duplicated reads

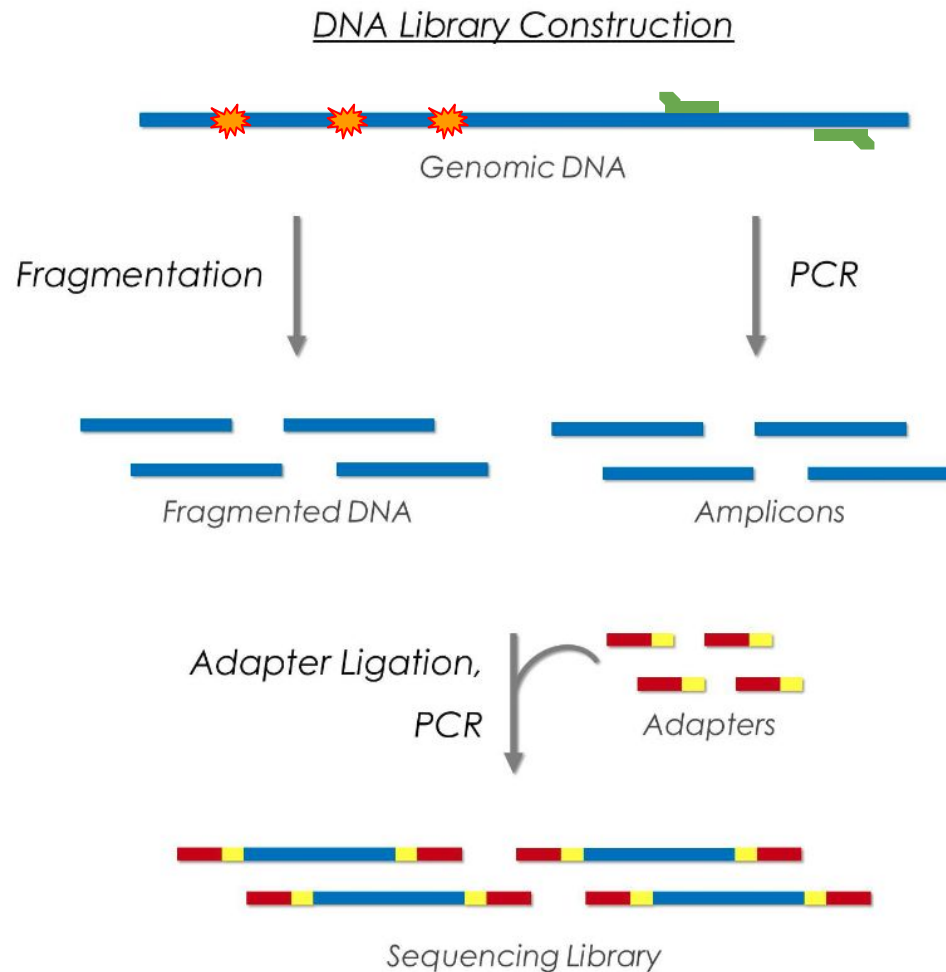
Why mark duplicates?

- 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



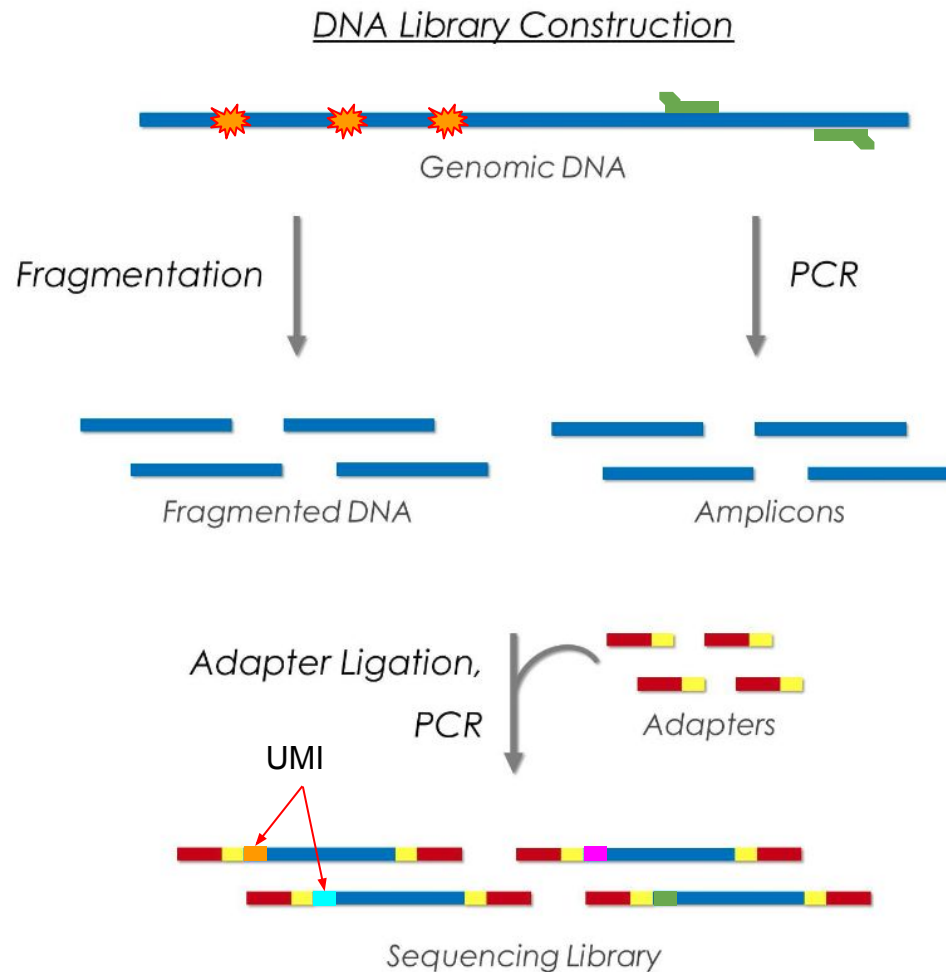
# Cleaning duplicated reads

Molecular Barcoding (UMI, *unique molecular identifiers*)



# Cleaning duplicated reads

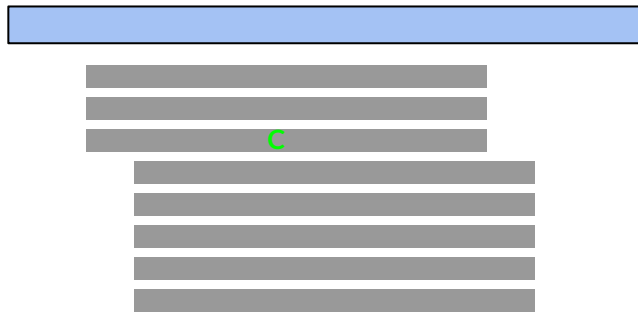
Molecular Barcoding (UMI, *unique molecular identifiers*)



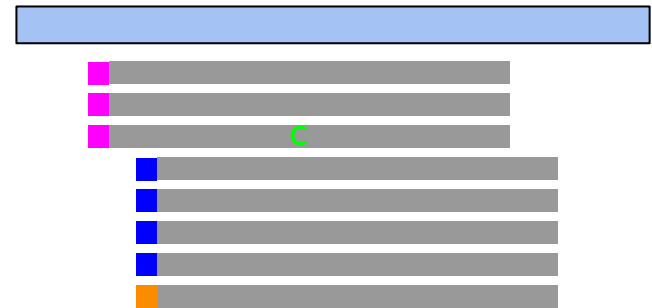
# Cleaning duplicated reads

Molecular Barcoding (UMI, *unique molecular identifiers*)

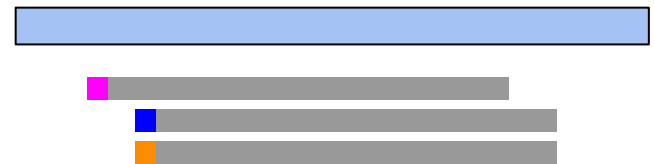
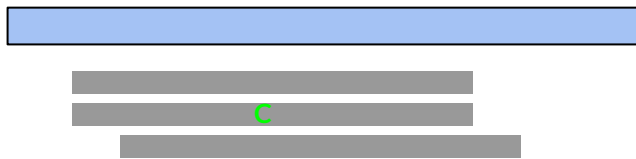
Without UMI



With UMI

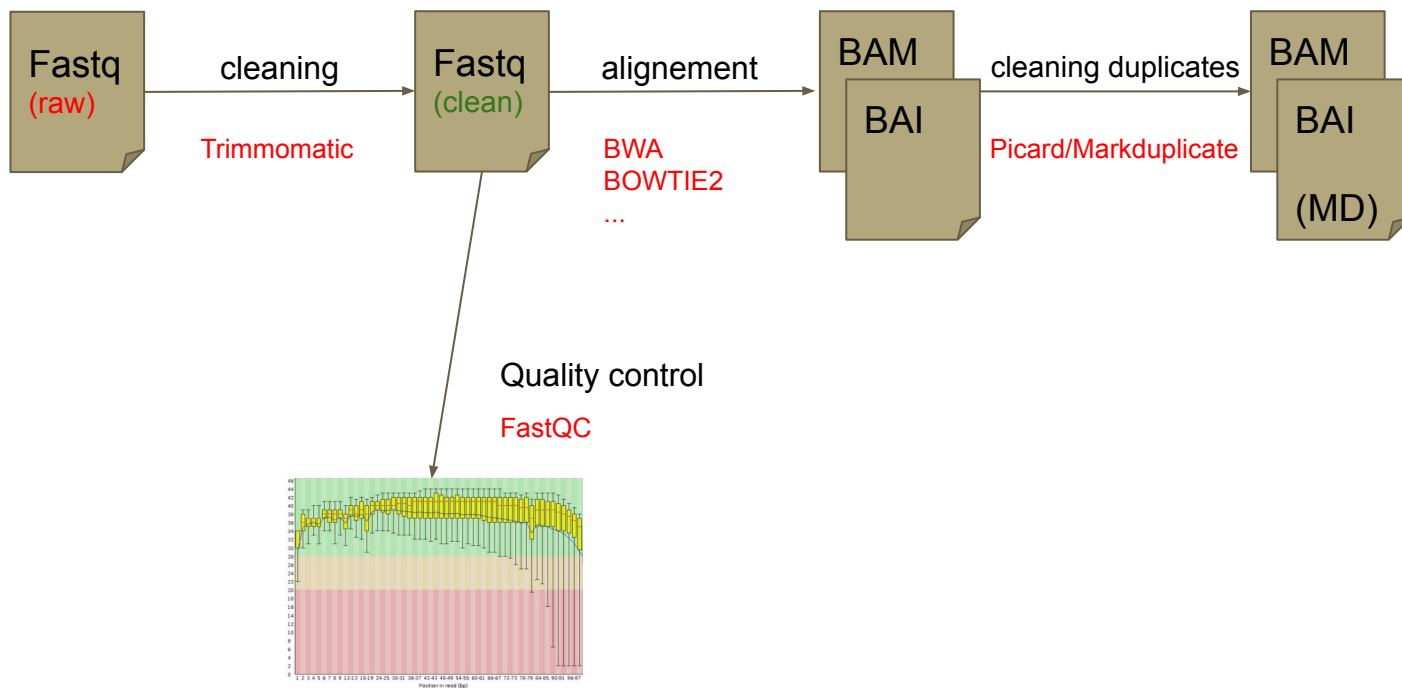


MarkDuplicate + Déduplication (consensus reads)



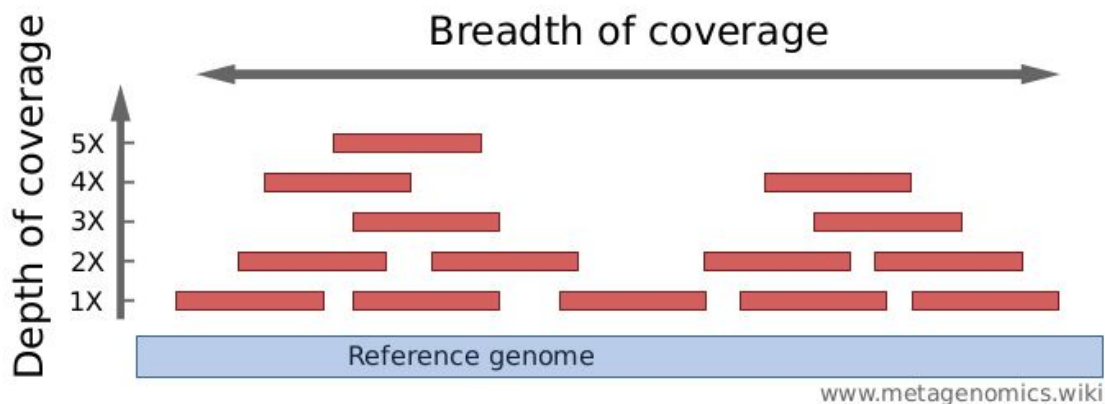


# Workflow



# 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 :

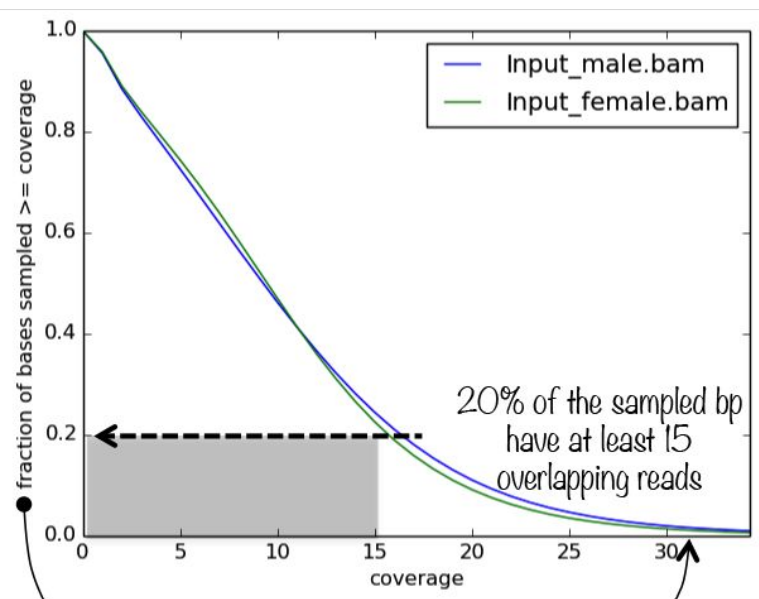
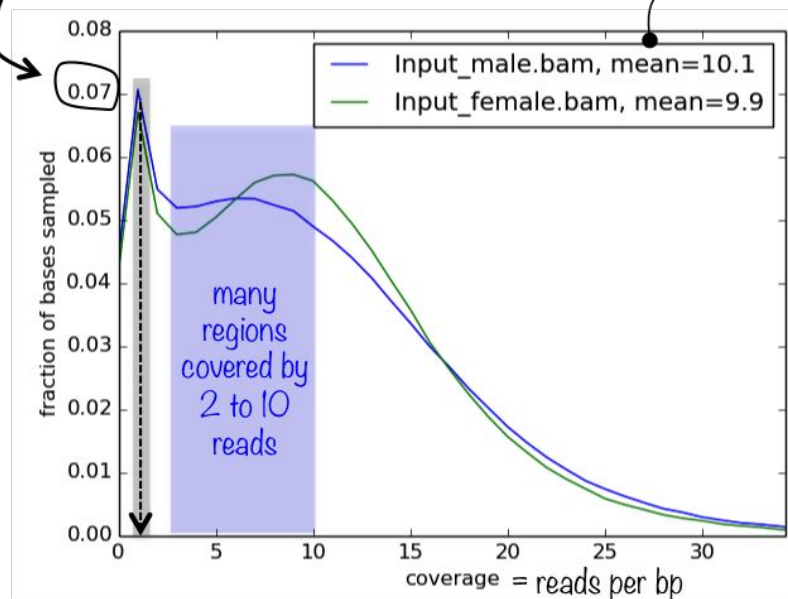
- Élodie Girard , 5ème Ecole de bioinformatique AVIESAN-IFB 2016 , [http://www.france-bioinformatique.fr/sites/default/files/V01\\_ITMO\\_2016\\_EG\\_from\\_fastq\\_to\\_mapping\\_1.pdf](http://www.france-bioinformatique.fr/sites/default/files/V01_ITMO_2016_EG_from_fastq_to_mapping_1.pdf)  
- <http://www.metagenomics.wiki/pdf/definition/coverage-read-depth>

# Computing coverage and depth of coverage

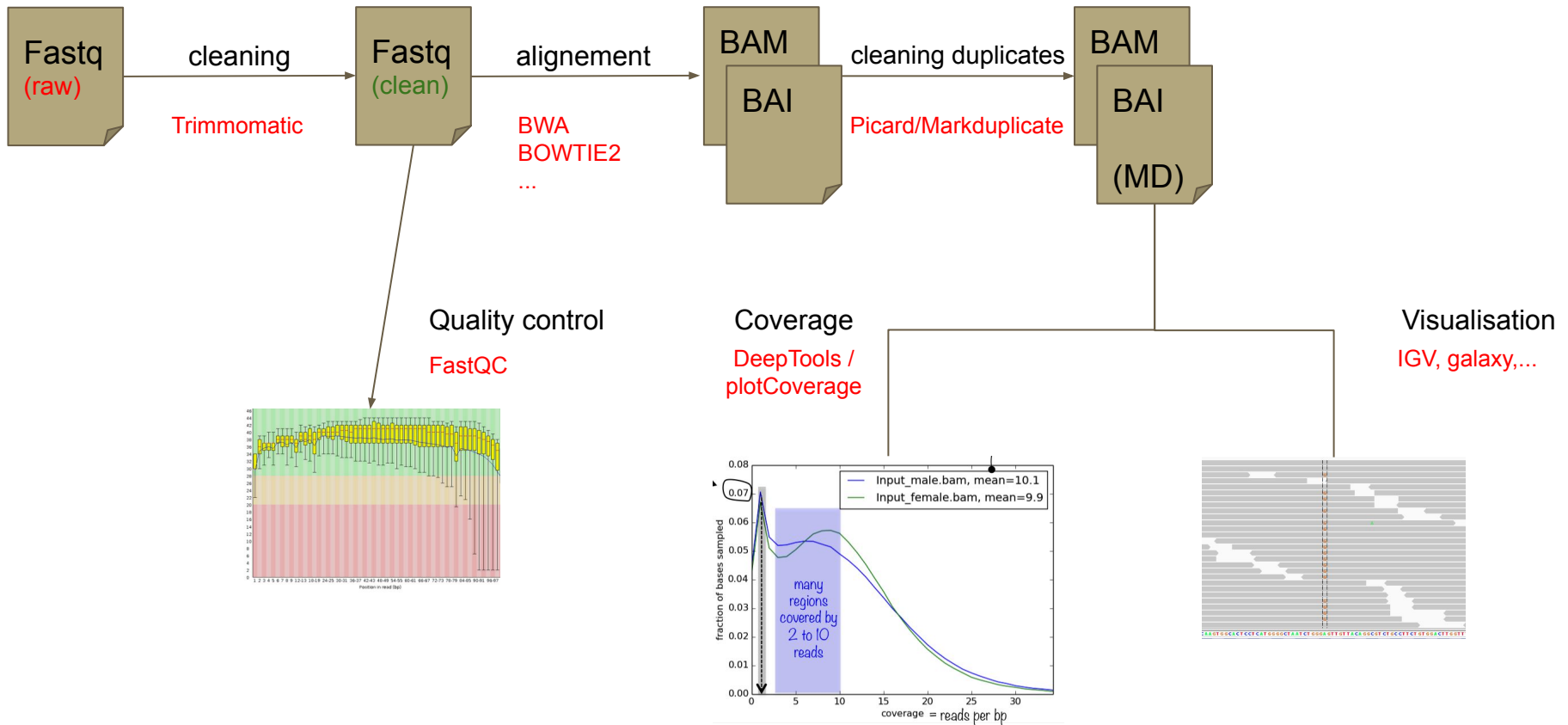
Tool: DeepTools / plotCoverage

7% of the sampled bp are covered once

mean coverage per bp



# Workflow



# Workflow

