Mapping Exome-seq data

Elodie Girard

U900 INSERM - Mines ParisTech - Institut Curie

ITMO Roscoff – 11/19/2013

institutCurie

Instituts thématiques Inserm Institut national de la santé et de la recherche médicale

MINES ParisTech

Dataset

- Public data: exome sequenced by the International HapMap Project
- Single-end reads of 100bp, Illumina Genome Analyzer IIx
- RNA-seq data of this exome available (Pickrell *et al.*, Nature, 2010)

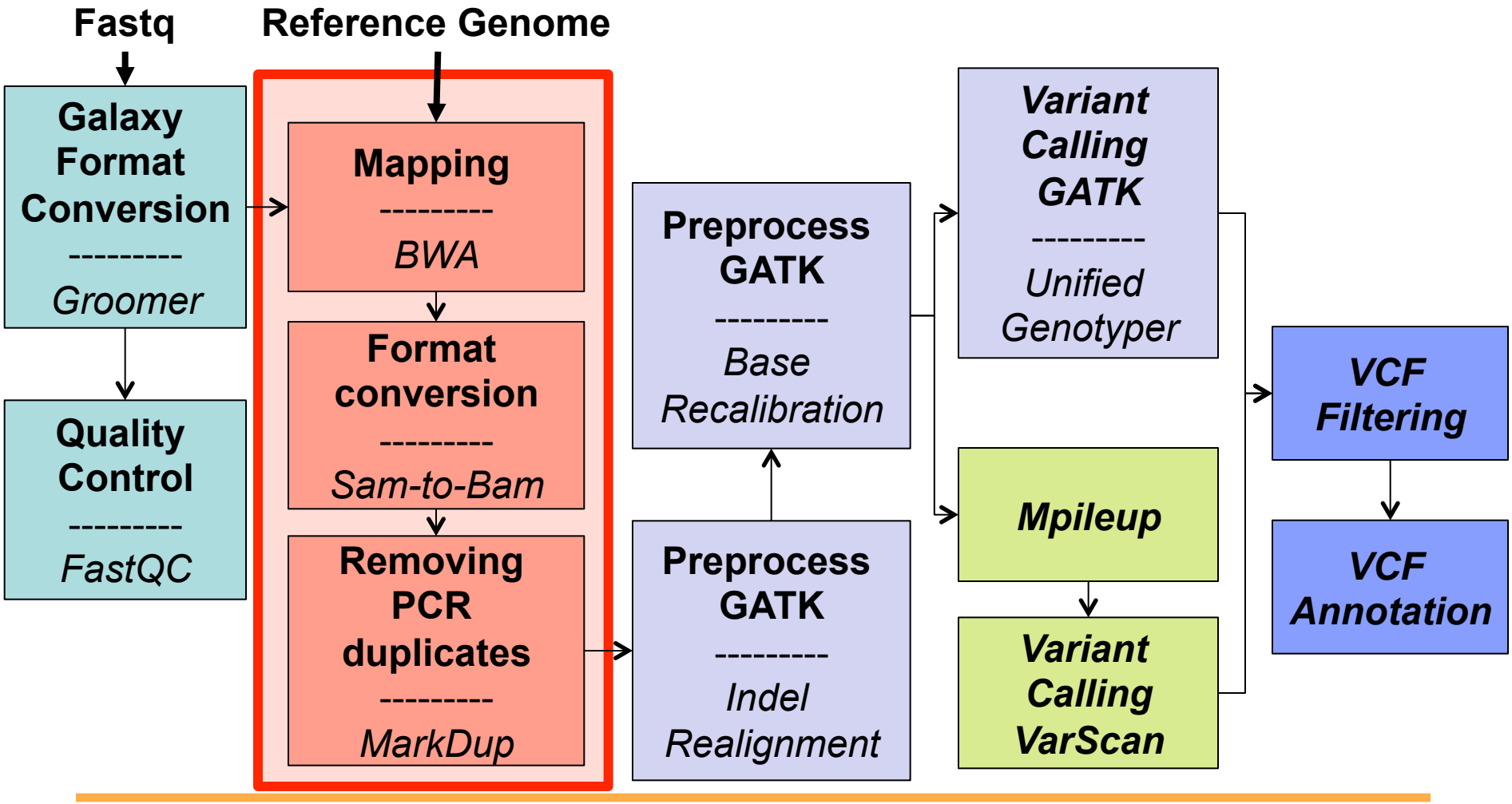
Objectives of the workshop:

- Variant calling, filtering and annotation in exome-seq data
- Observing the potential impact of these variants by looking at the corresponding RNA-seq data

Objectif of this session :

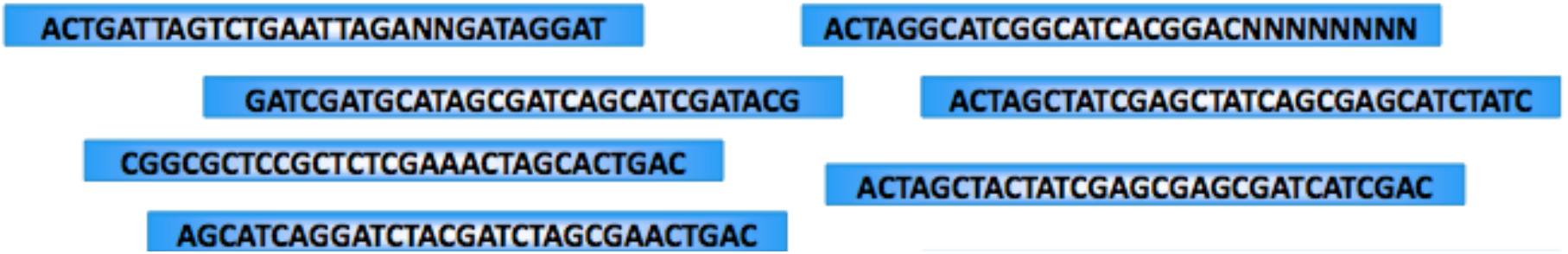
➢ Mapping the exome-seq data to the reference genome

Workflow


Fastq
Reference Genome
Galaxy Format Conversion
Groomer
Quality Control
FastQC
Mapping
BWA
Format conversion
Sam-to-Bam
Removing PCR duplicates
MarkDup
Preprocess GATK
Base Recalibration
Preprocess GATK
Indel Realignment
Variant Calling GATK
Unified Genotyper
Mpileup
Variant Calling VarScan
VCF Filtering
VCF Annotation

Raw data

- **Raw data**: millions of short reads of the same size (Illumina) or different sizes (Ion PGM), single-end or paired-end

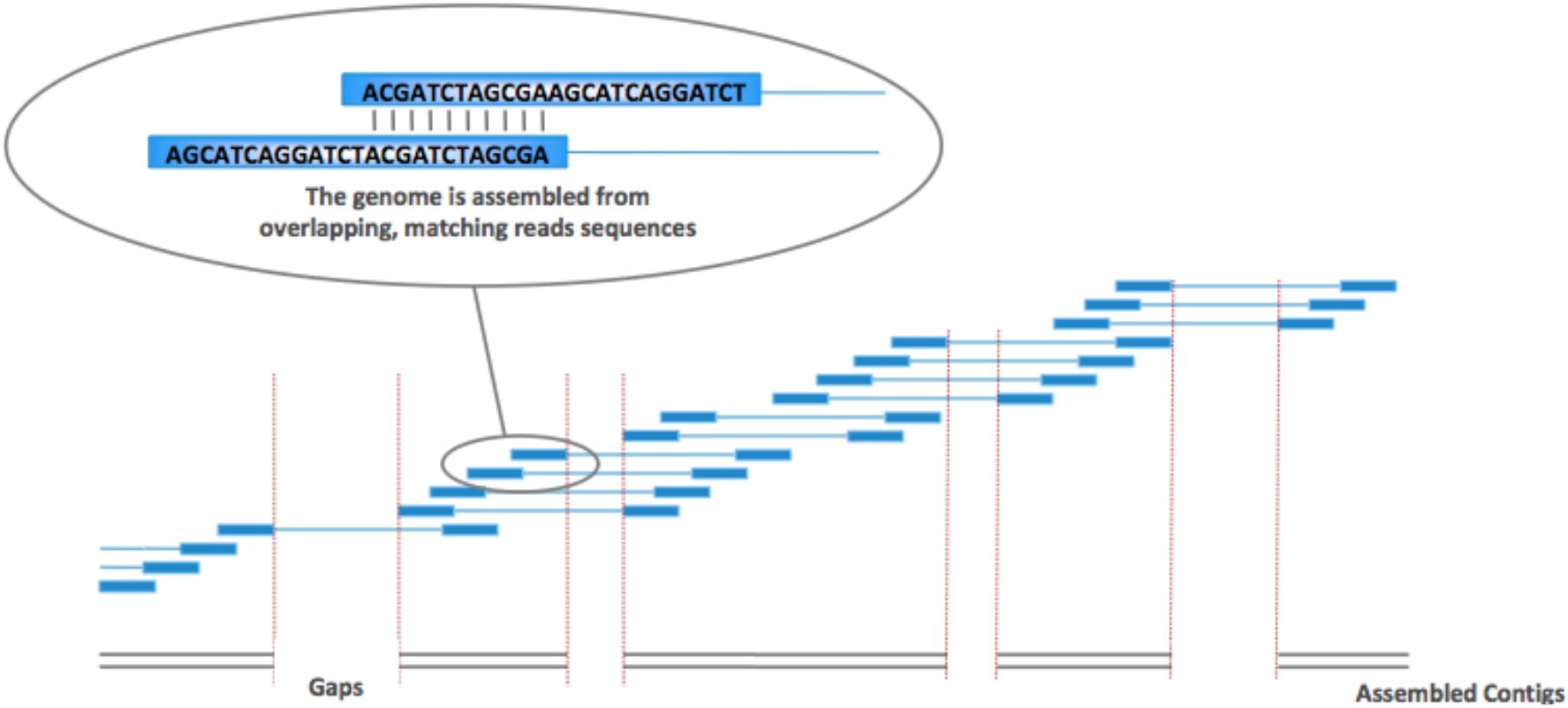


Example: fastq format

@SRR081222.573928
CATTCCCTAGAAGGCTGGCCTGCTCTTGCCTCTCTGTCCATTCCCTAGAAGGCTGGCTGGCCC
+
5@@-BFFGDFHHIJIJIJIIIIIHHJGIIIJHJHJJGHIIHGIHIJGAIIJIJJFJI%%%%%%%%%%%%%%%%%%%%%

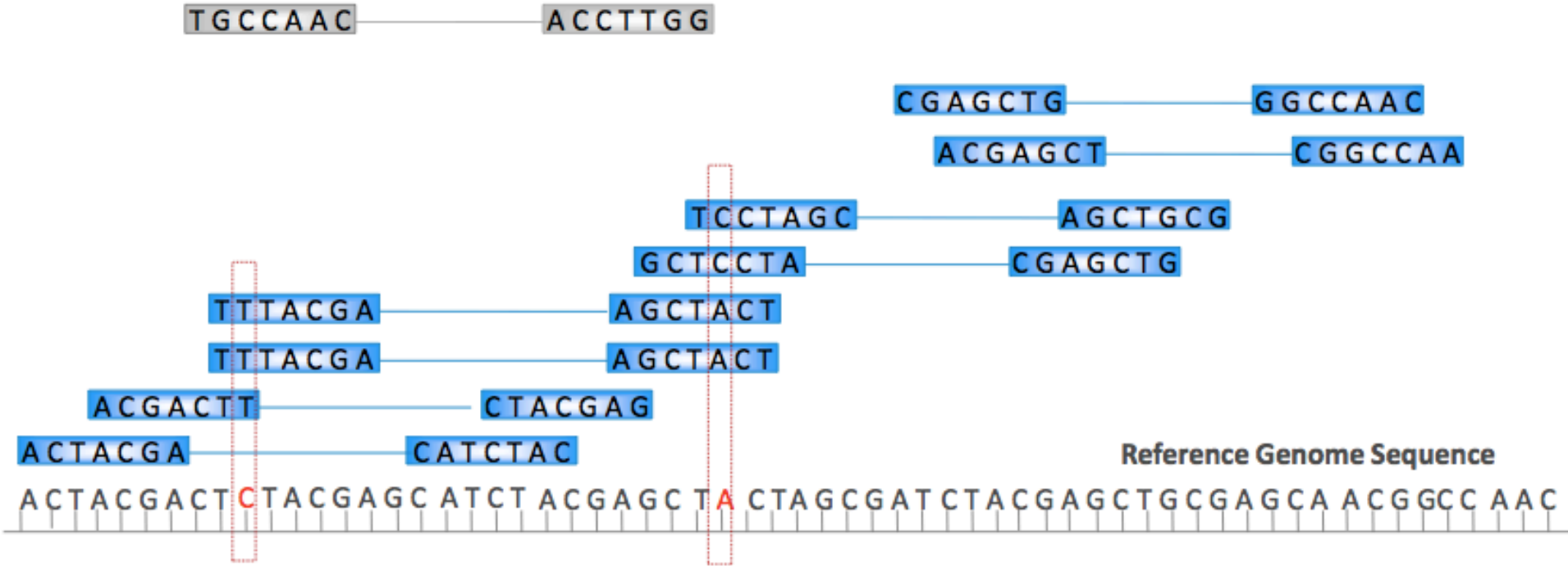
De Novo Reads Assembly

- **De Novo Reads Assembly**: used when there's no reference genome ; aims at reconstructing long scaffolds from single reads



Mapping on a reference Genome

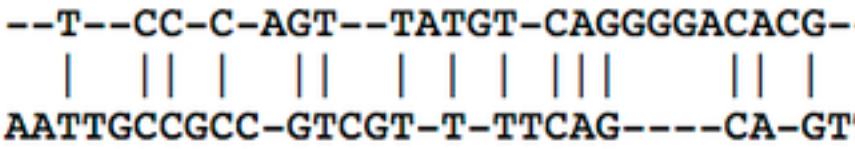
- **Reference genome:** known sequence supposed to be **as close as possible** to the input genome and which is used as an **anchor** to organize the reads information



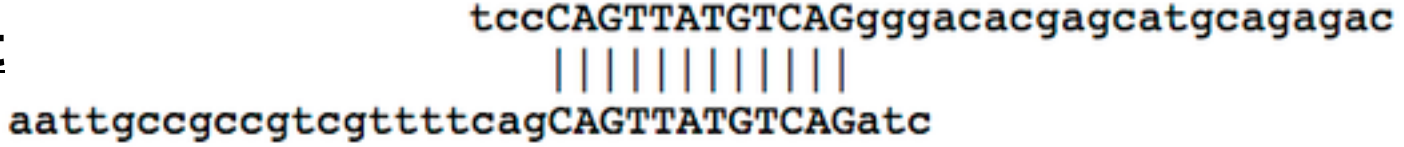
Reads Alignment - Vocabulary

- **Mismatch**: Incoherence between two nucleotides
- **Indels:** Insertion/Deletion into the reference genome
- **Gap**: Bridge within the read alignment (i.e. small indels)
- **Mappability:** Uniqueness of a region
  - ➢ repeated region = low mappability
  - ➢ unique region = good mappability

Global Alignment

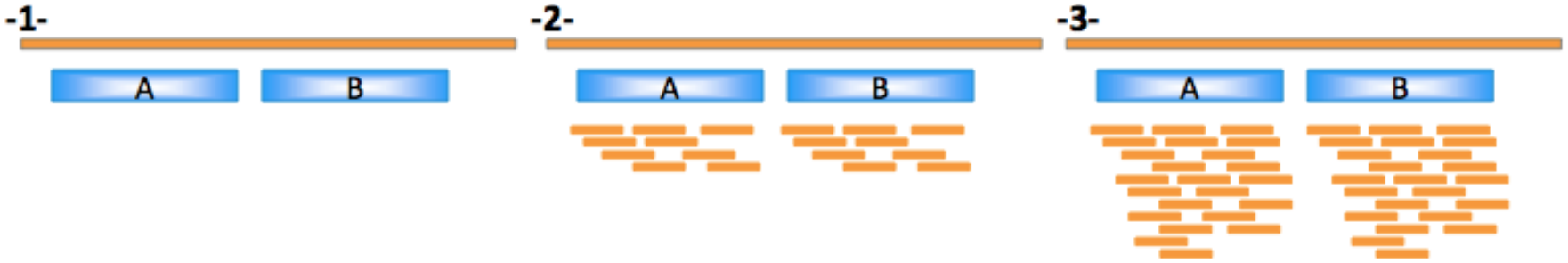


Local Alignment



Multiple Alignments

- A read can align **multiple times** on the genome (repeated elements…)
- How to deal with these multiple alignments reads ?
- Three strategies:
  - **-1-** Report only unique alignment
  - **-2-** Report best alignments & randomly assign reads across equally good loci
  - **-3-** Report all (best) alignments



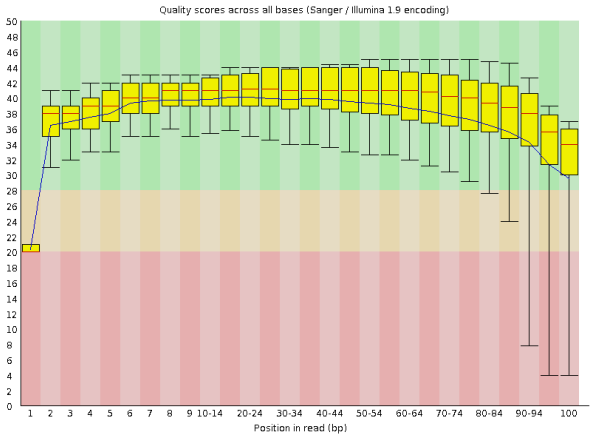
- **Mapping Quality:** quantify the probability that a read is misplaced. Low if a read has multiple alignments

Alignment tools

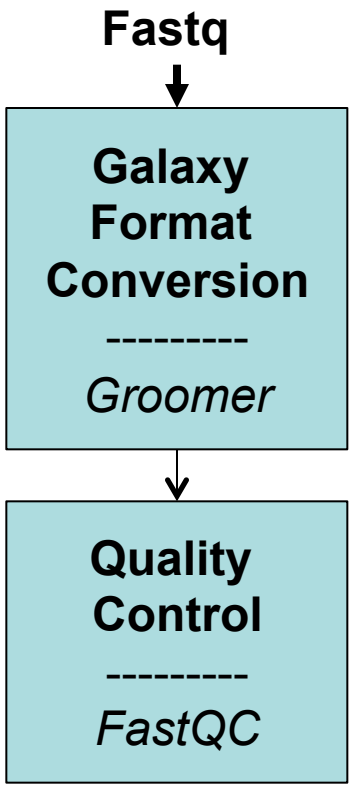
- Multitude of alignment tools: BWA, Bowtie, Bowtie2, Bfast…
- How to choose the best tool ?
  - Is my sequencing technology supported ?
  - Do I have short or long reads ? Reads of different sizes ?
  - Do I want to allow gapped alignment ? Multiple alignments ?
  - Does it support single/paired-end reads ?
  - On which alignment algorithm is it based ?
  - Computational issues ? Is it used by the community ?

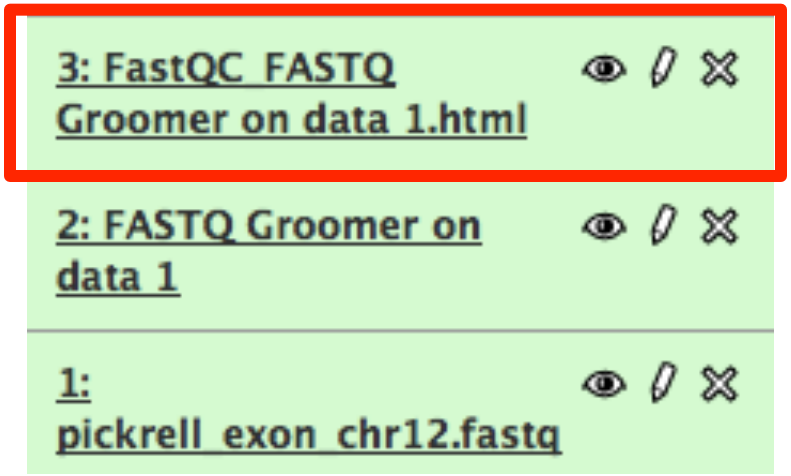
➢ A classical and performant tool for Illumina sequencing: BWA (Burrows-Wheeler Aligner)

Galaxy: summary of the previous steps



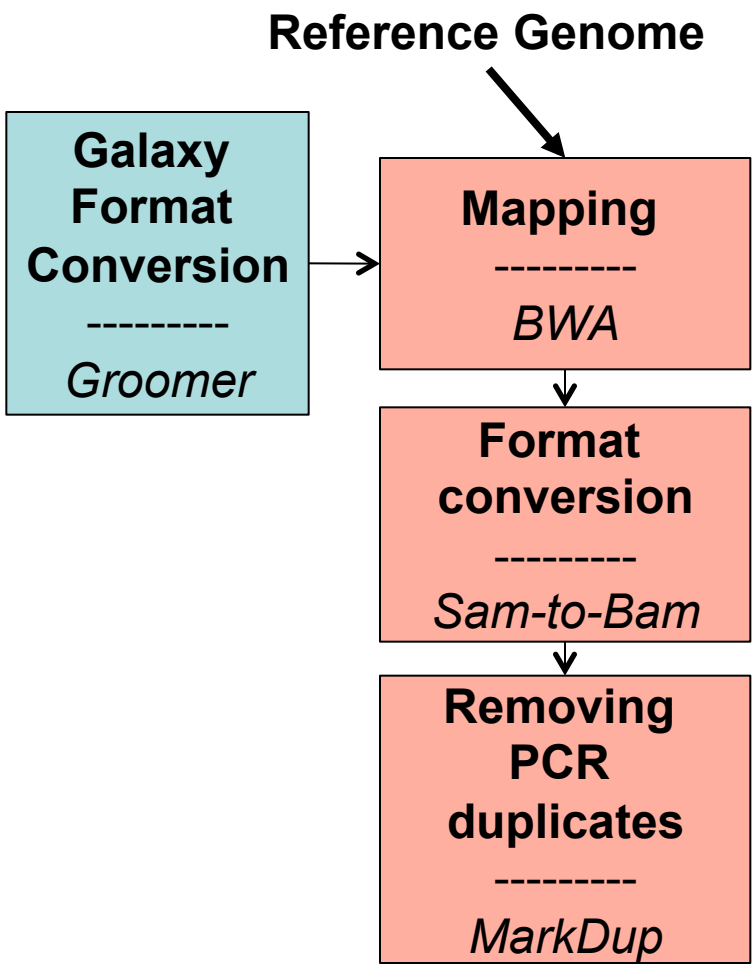
Number of reads: 5,664,374





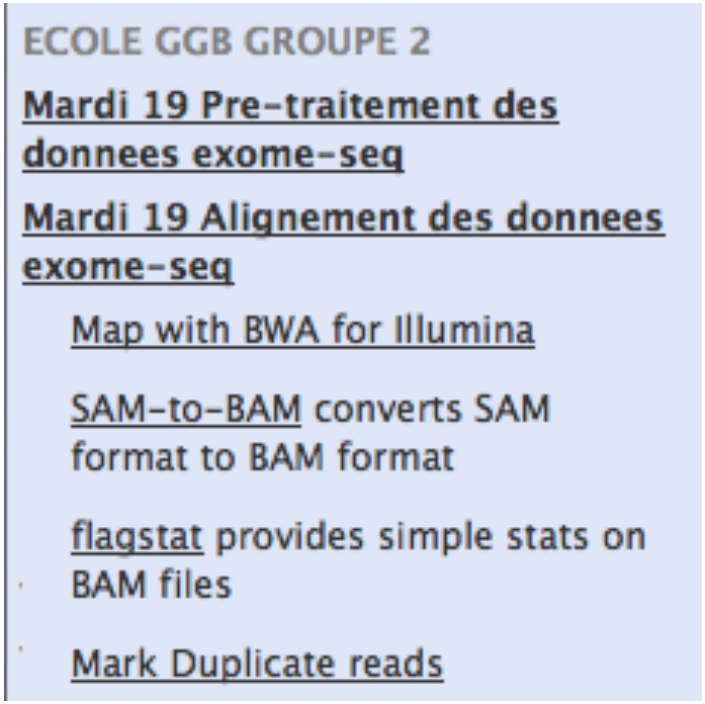
Input file of this session: FASTQ Groomer on data 1

Tools



All the tools are in the left panel:

Mardi 19 Alignement des données exome-seq

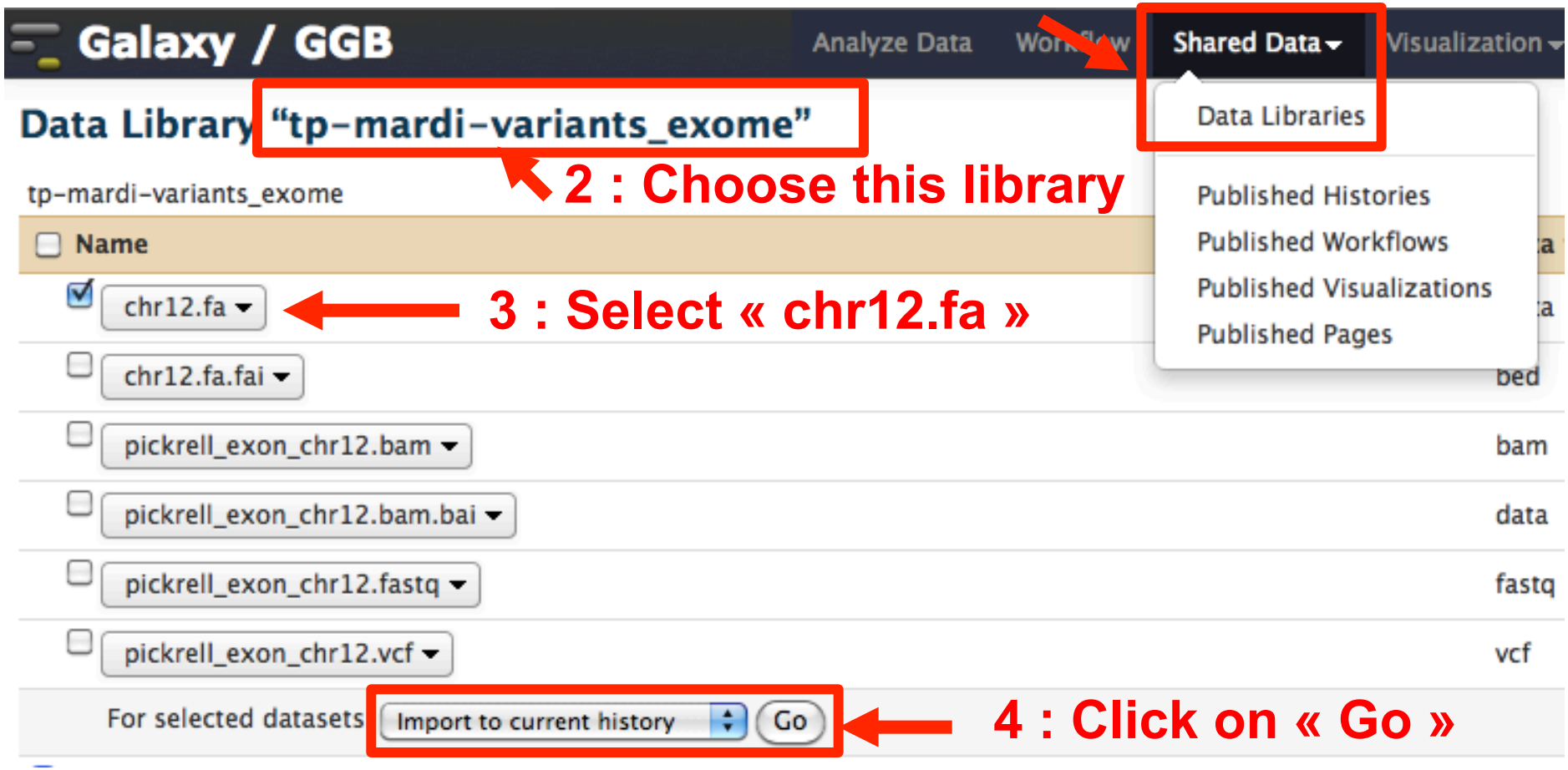


institutCurie

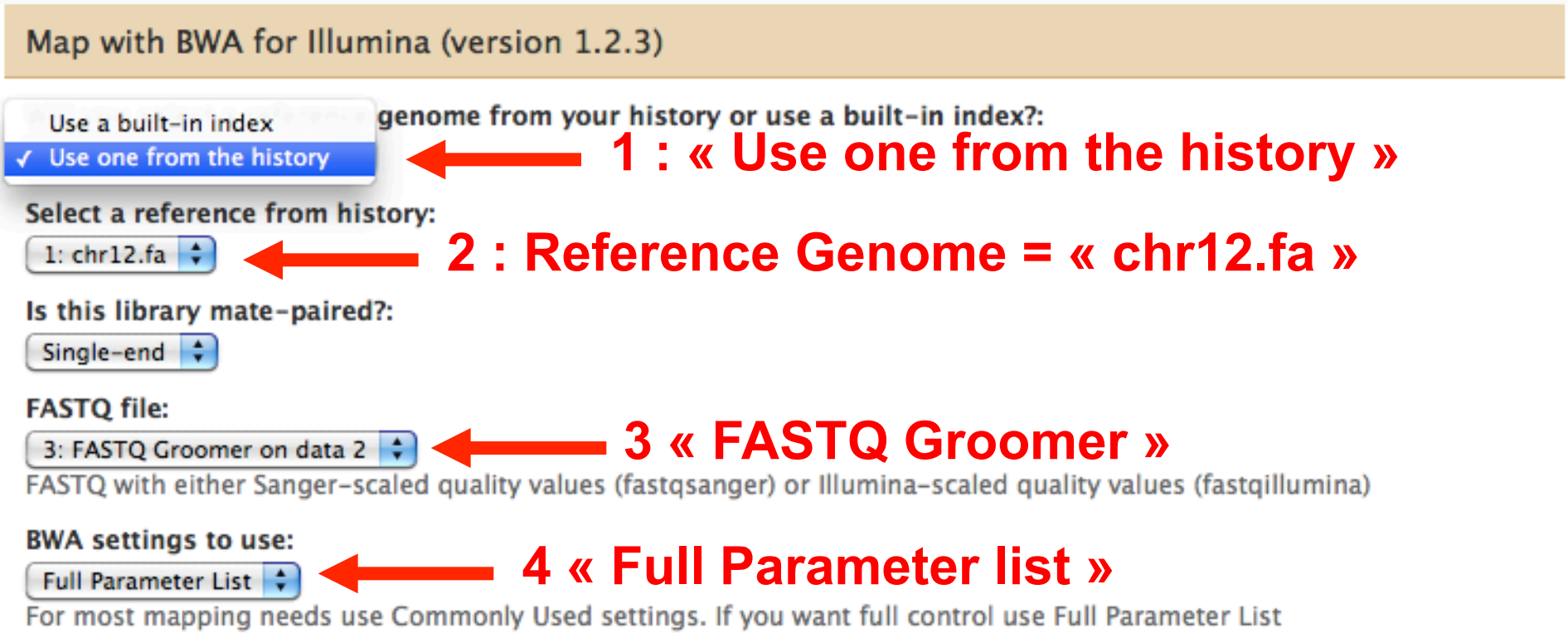
Instituts thématiques Inserm
Institut national de la santé et de la recherche médicale

MINES ParisTech

Loading the Reference Genome



Mapping with BWA



Mapping with BWA

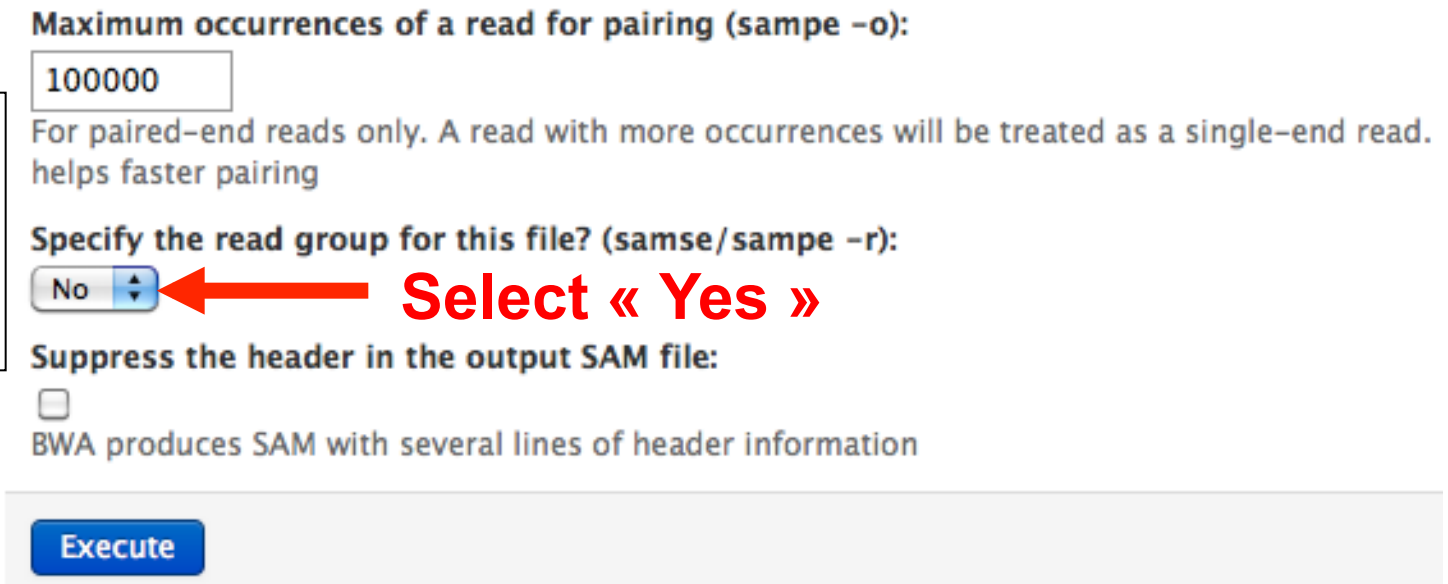
Adding a **Read Group** is required for the Variant Calling with GATK

→ Informations on the sample, library used, platform…

→ Attribute a read to a sample (useful in multi-samples study)

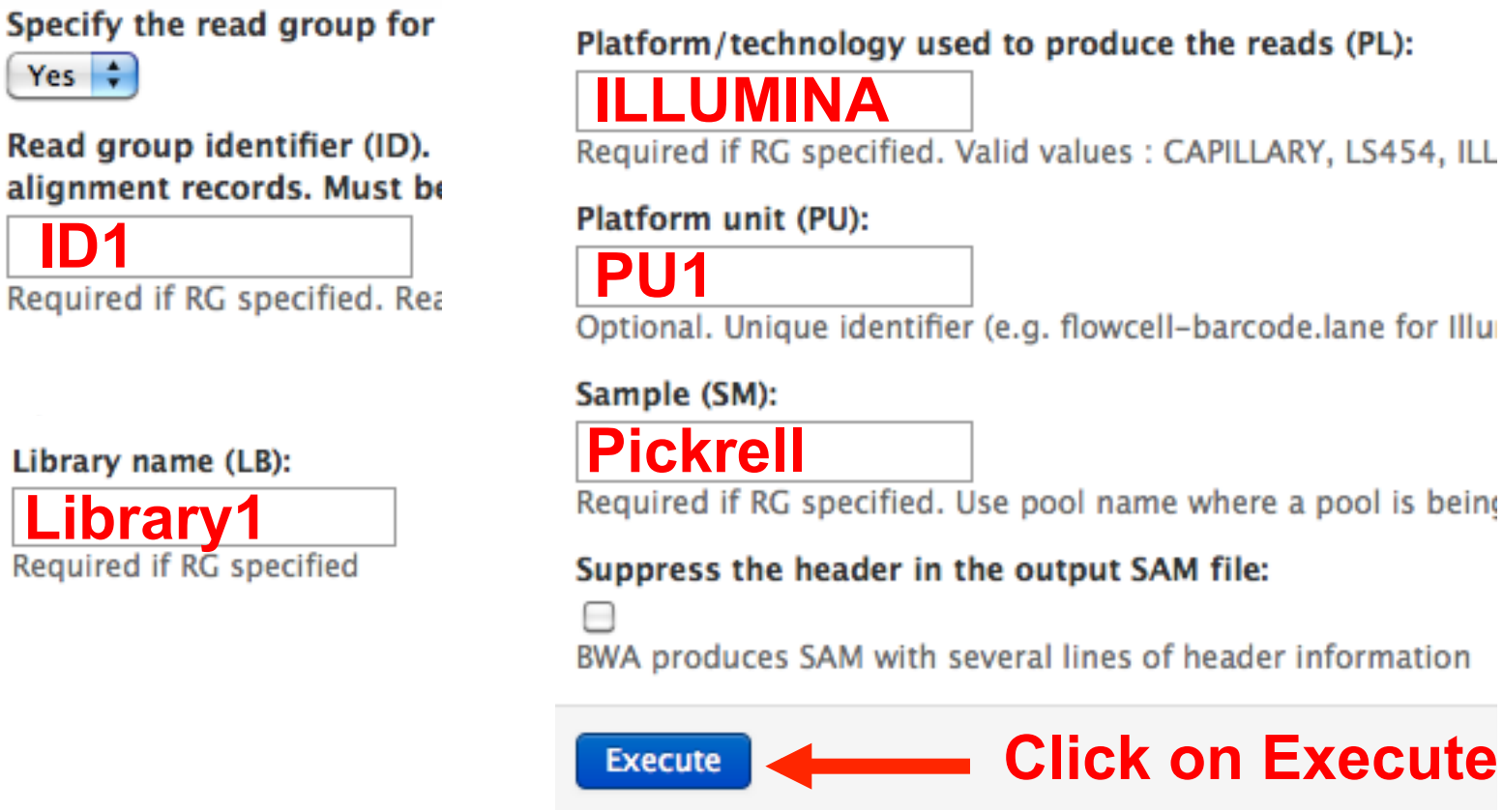
→ Other tools exist to add Read Groups

At the bottom of the parameters list



Mapping with BWA

Fill every box noted as « Required »



SAM-to-BAM

- **SAM Format:**

@SQ SN:chr12 LN:133851895

@RG ID:1 LB:sample PL:ILLUMINA SM:sample

Read Name / Flag / chr / Pos / MapQ / Cigar

SRR081222.573928 16 chr12 60124 23 100M * 0 0

GCCCCTGGGGATGTTTTGCACCAAGCCACTGTCTCCAGCTGG

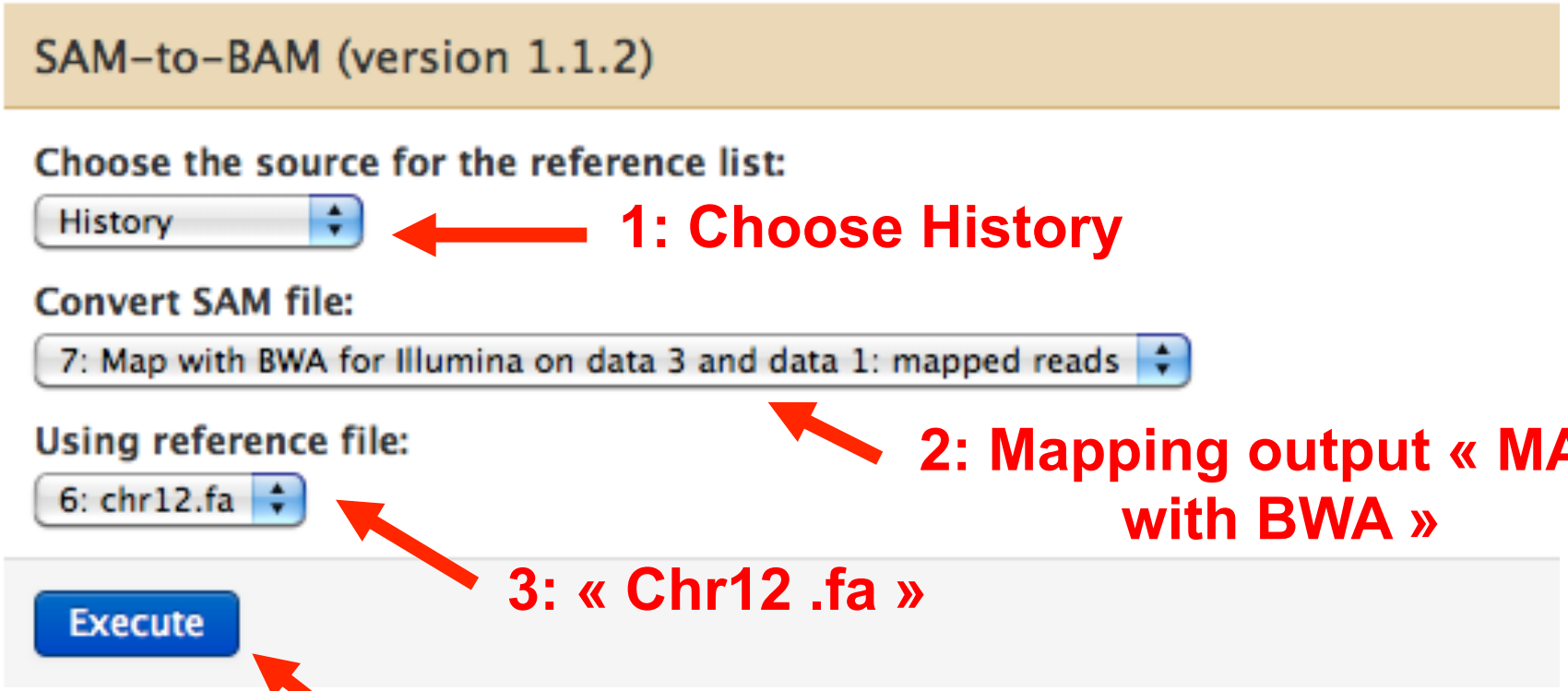
BBC@GIIHGCFCIEHEAIEIFFGEONDNJFINIONHNGJNNNNKNJN

RG:Z:1 XT:A:U NM:i:0 X0:i:1 X1:i:1 XM:i:0 XO:i:0 XG:i:0 MD:Z:100 XA:Z

Read group affiliation

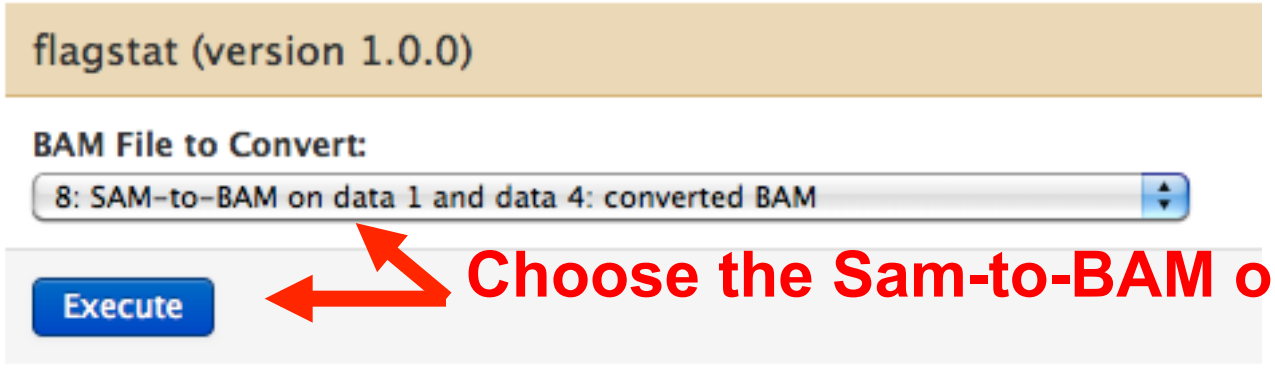
- **BAM Format:** Binary SAM Format (compressed = smaller size)

SAM-to-BAM


SAM-to-BAM (version 1.1.2)
Choose the source for the reference list:
History
1: Choose History
Convert SAM file:
7: Map with BWA for Illumina on data 3 and data 1: mapped reads
2: Mapping output « MAP with BWA »
Using reference file:
6: chr12.fa
3: « Chr12 .fa »
Execute
4: Click on Execute

Mapping Statistics

Flagstat: a tool showing some mapping statistics (from the SAMtools software)



Choose the Sam-to-BAM output and execute

```
5664374 + 0 in total (QC-passed reads + QC-failed reads)
0 + 0 duplicates
5523578 + 0 mapped (97.51%)
0 + 0 paired in sequencing
0 + 0 read1
0 + 0 read2
0 + 0 properly paired (-nan%:-nan%)
0 + 0 with itself and mate mapped
0 + 0 singletons (-nan%:-nan%)
0 + 0 with mate mapped to a different chr
0 + 0 with mate mapped to a different chr (mapQ>=5)
```

% of mapped reads

Informations on pairs when available

institutCurie

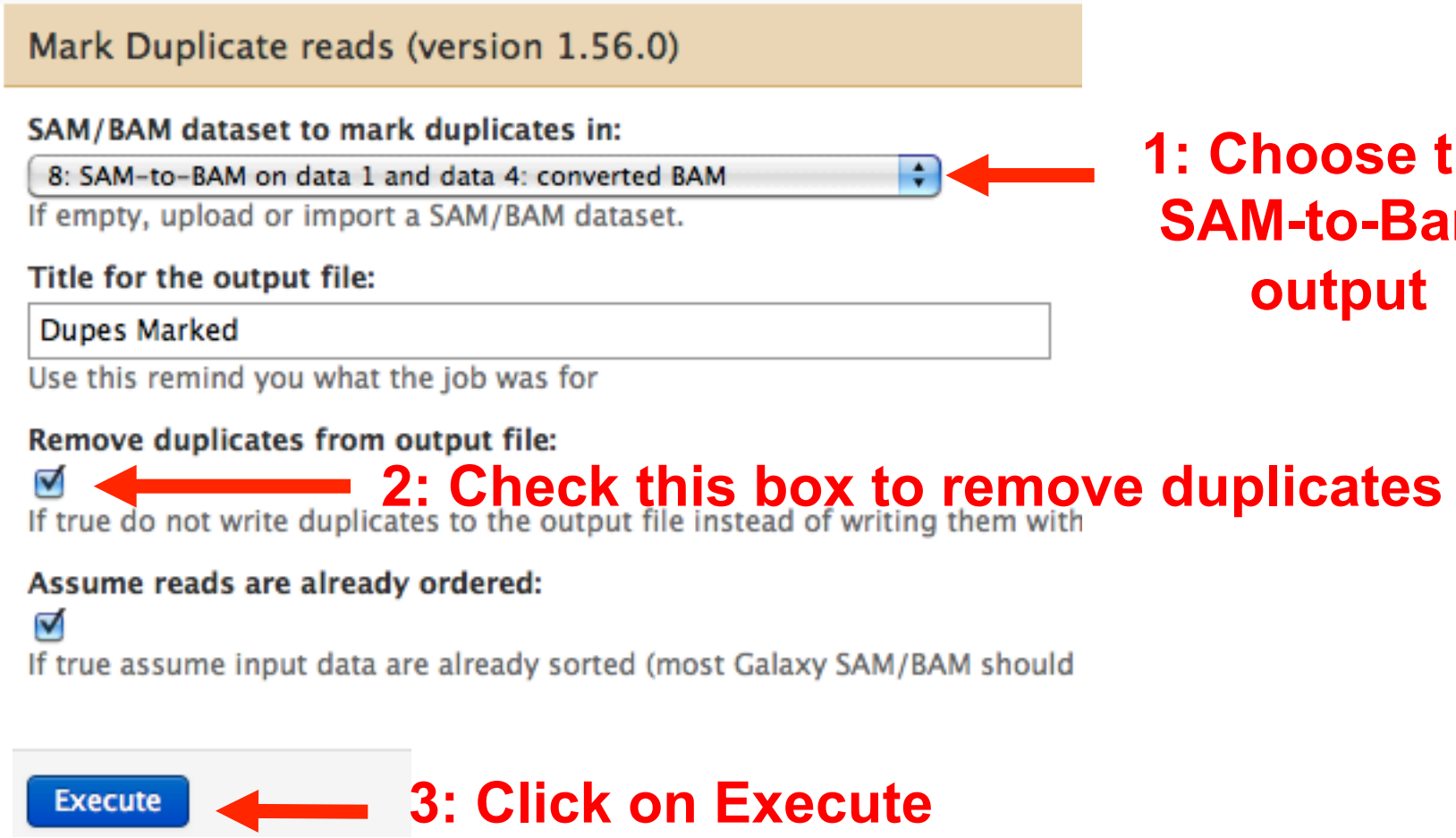
Instituts thématiques Inserm Institut national de la santé et de la recherche médicale

MINES ParisTech

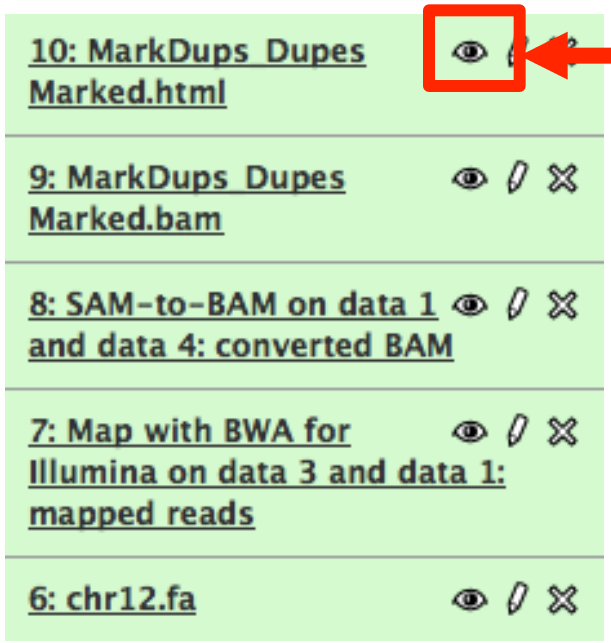
Removing Duplicates

- **Duplicates reads**: different reads having the same sequence
- PCR amplification causes **molecular duplicates** and sequencing artifacts cause **optical duplicates** (*e.g.:* same cluster read twice)
- 2 Tools to **mark** (only a flag is added) or **remove** duplicates:
  - **Rmdup** (SAMtools): efficient with single-end reads
  - **MarkDuplicates** (Picard): efficient with single/paired-end reads (takes into account the 5' coordinates, the mapping orientation and all gaps in the alignment)
- The removal of the duplicates depends on the application (not appropriate for ion PGM targeted sequencing by example)

Removing Duplicates

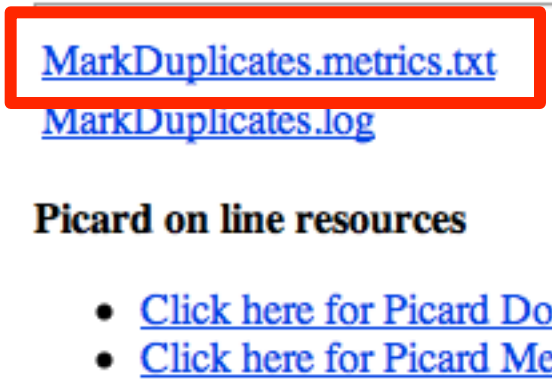


Removing Duplicates



1: Click on the eye of the « html » output

2: Click on MarkDuplicates.metrics.txt




```
## METRICS CLASS        net.sf.picard.sam.DuplicationMetrics
LIBRARY UNPAIRED_READS_EXAMINED READ_P          UNMAPPED_READS
UNPAIRED_READ_DUPLICATES        READ_P          READ_PAIR_OPTICAL_DUPLICATES
PERCENT_DUPLICATION     ESTIMATED_LIBRARY_SIZE
Unknown Library 0       0       140796  0       0       0       ?
sample  5523578 0       0       1835477 0       0       0.332299
```

#Unmapped

#Mapped

#Duplicates

% Duplicates

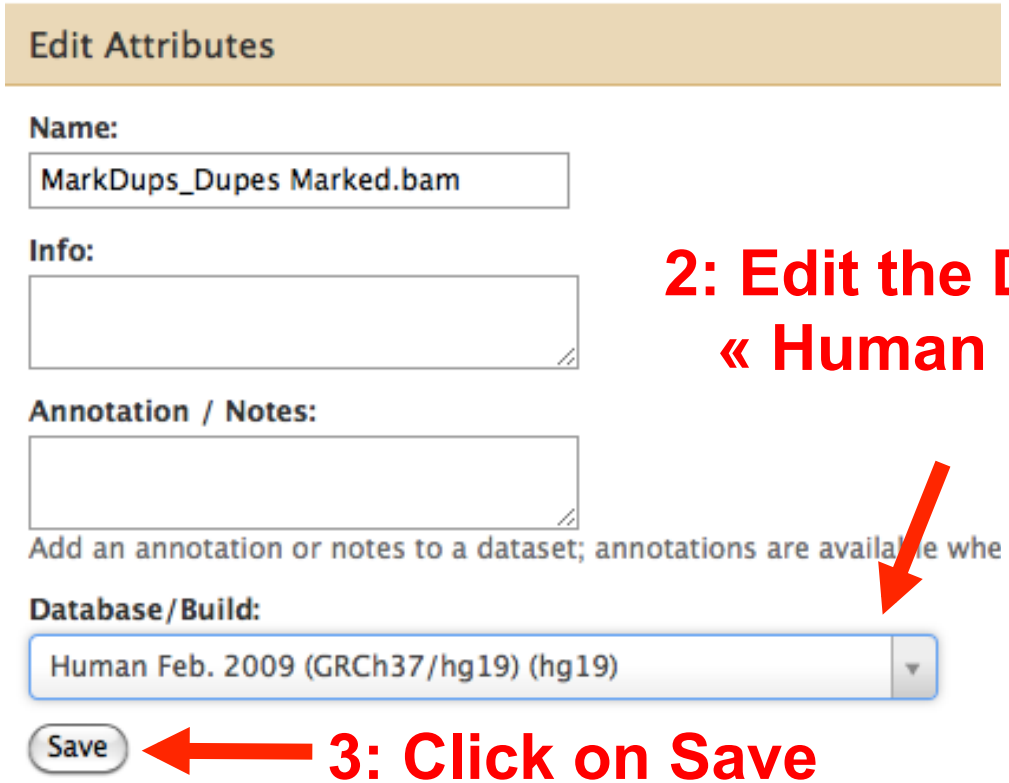
institutCurie

Instituts thématiques Inserm Institut national de la santé et de la recherche médicale

MINES ParisTech

Visualization with IGV (**I**ntegrative **G**enomics **V**iewer)

1: Click on the pencil to edit the attributes of this bam



2: Edit the Database/build by choosing « Human Feb.2009 (hg19) » genome

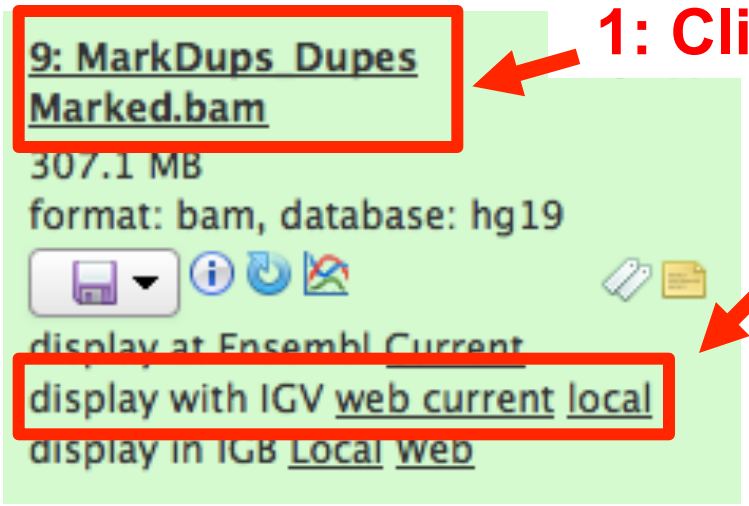
3: Click on Save

institutCurie

Instituts thématiques Inserm Institut national de la santé et de la recherche médicale

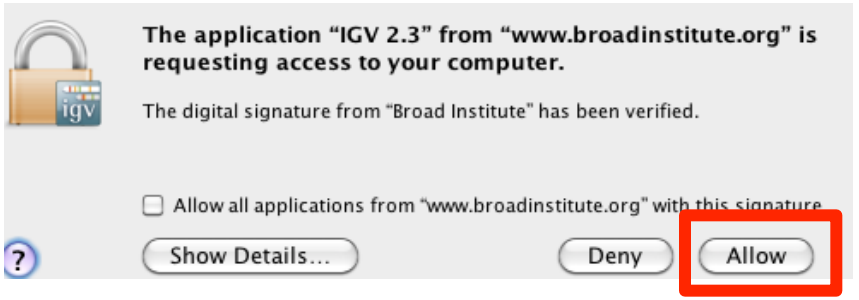
MINES ParisTech

Visualization with IGV



1: Click on the name to open the options

2: Click on « web current » to download and open IGV



3: Allow IGV

Once IGV is opened, don't click on « web current » but on « local » to visualize an other BAM

institutCurie

Instituts thématiques Inserm
Institut national de la santé et de la recherche médicale

MINES ParisTech

Visualization with IGV

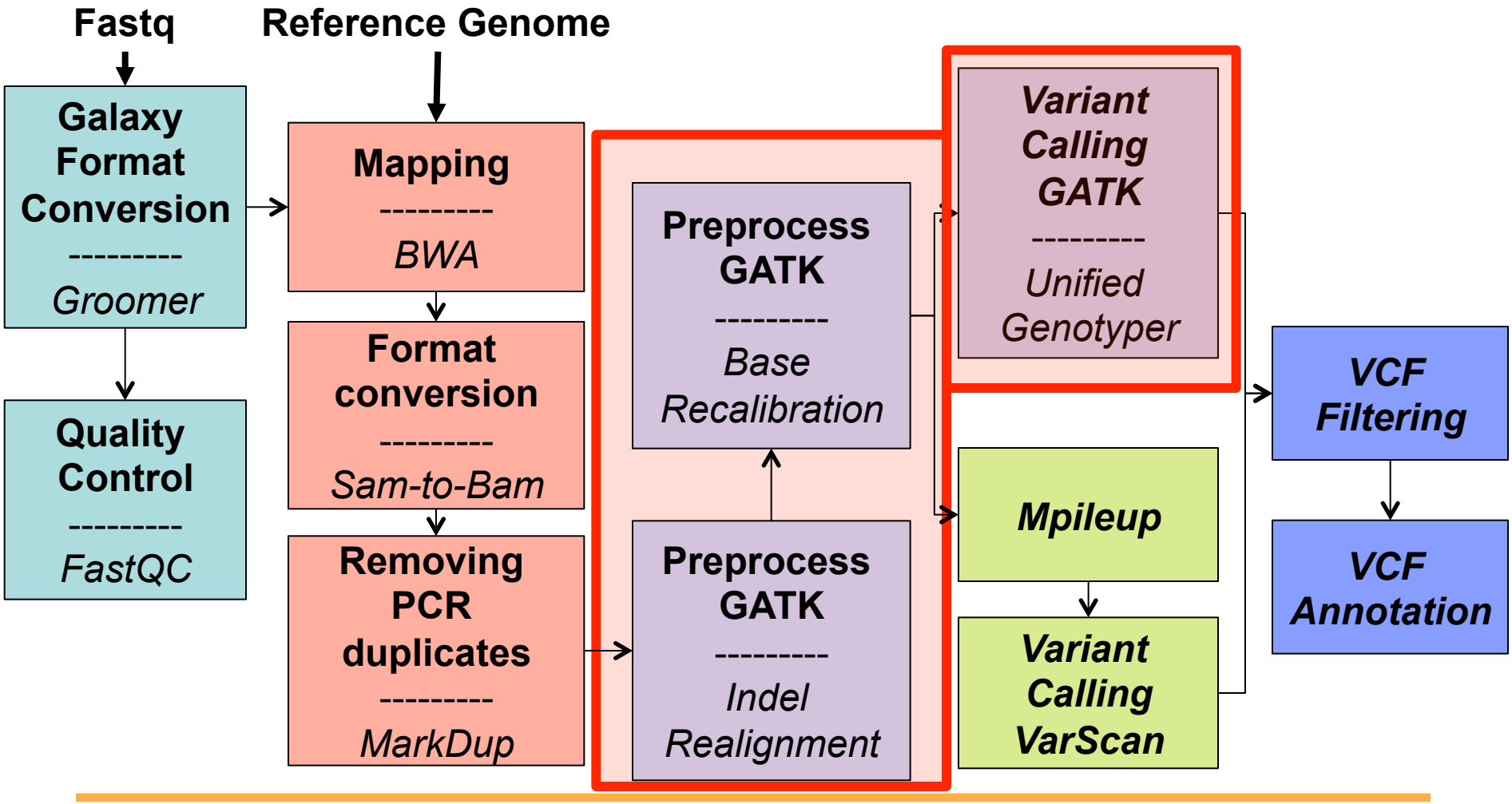
Enter « chr12:113,342,800-113,359,800 » and click on « Go »

institutCurie

Instituts thématiques Inserm Institut national de la santé et de la recherche médicale

MINES ParisTech

Next Step


Fastq
Reference Genome
Galaxy Format Conversion

Groomer
Quality Control

FastQC
Mapping

BWA
Format conversion

Sam-to-Bam
Removing PCR duplicates

MarkDup
Preprocess GATK

Base Recalibration
Preprocess GATK

Indel Realignment
Variant Calling GATK

Unified Genotyper
Mpileup
Variant Calling VarScan
VCF Filtering
VCF Annotation