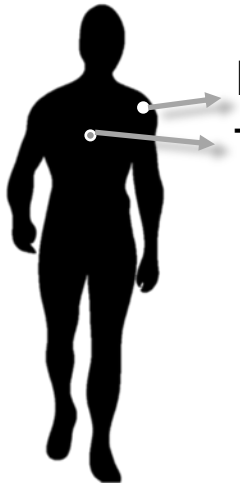
Réalisation d’un Pipeline d’analyse d’exome

UNIVERSITÉ PARIS SUD

GUSTAVE ROUSSY
CANCER CAMPUS GRAND PARIS

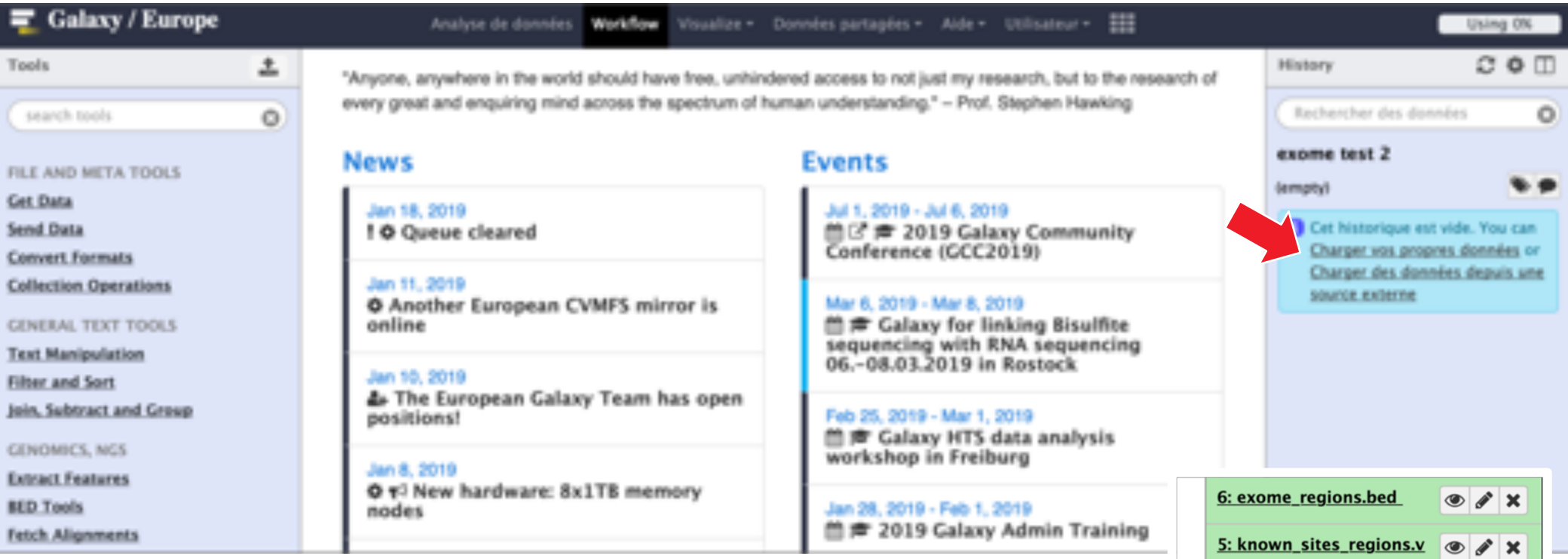
Les données



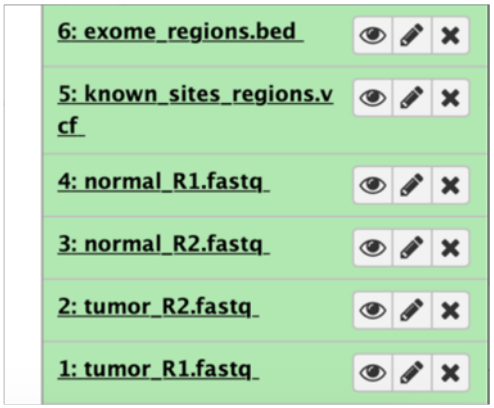
Ju et al. Genome Res. 22:436–445, 2012
100bp paired-end reads, Illumina HiSeq 2000
SRA (Sequence Read Archive): ERA148528

- Mean depth higher for the tumor sample (~100X) than for the normal sample (~30X) to detect somatic variant with a low allelic frequency
- Aligned Exome size: ~15 Go tumor ; ~7 Go blood
  Complete analysis processing Hme: ~20h
- **Fastq files restricted to a few regions (~112kb)**
- **to limit processing time**

Chargez vos données



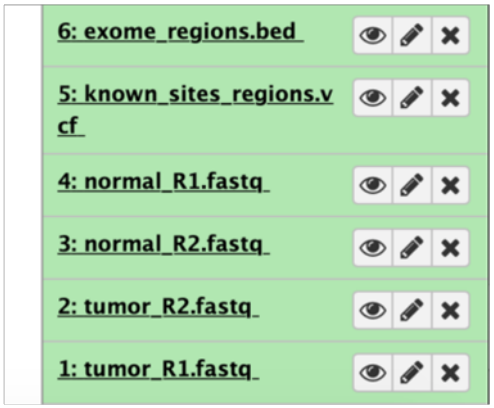
Fera apparaître:



(alternativement: à partir de données partagées)

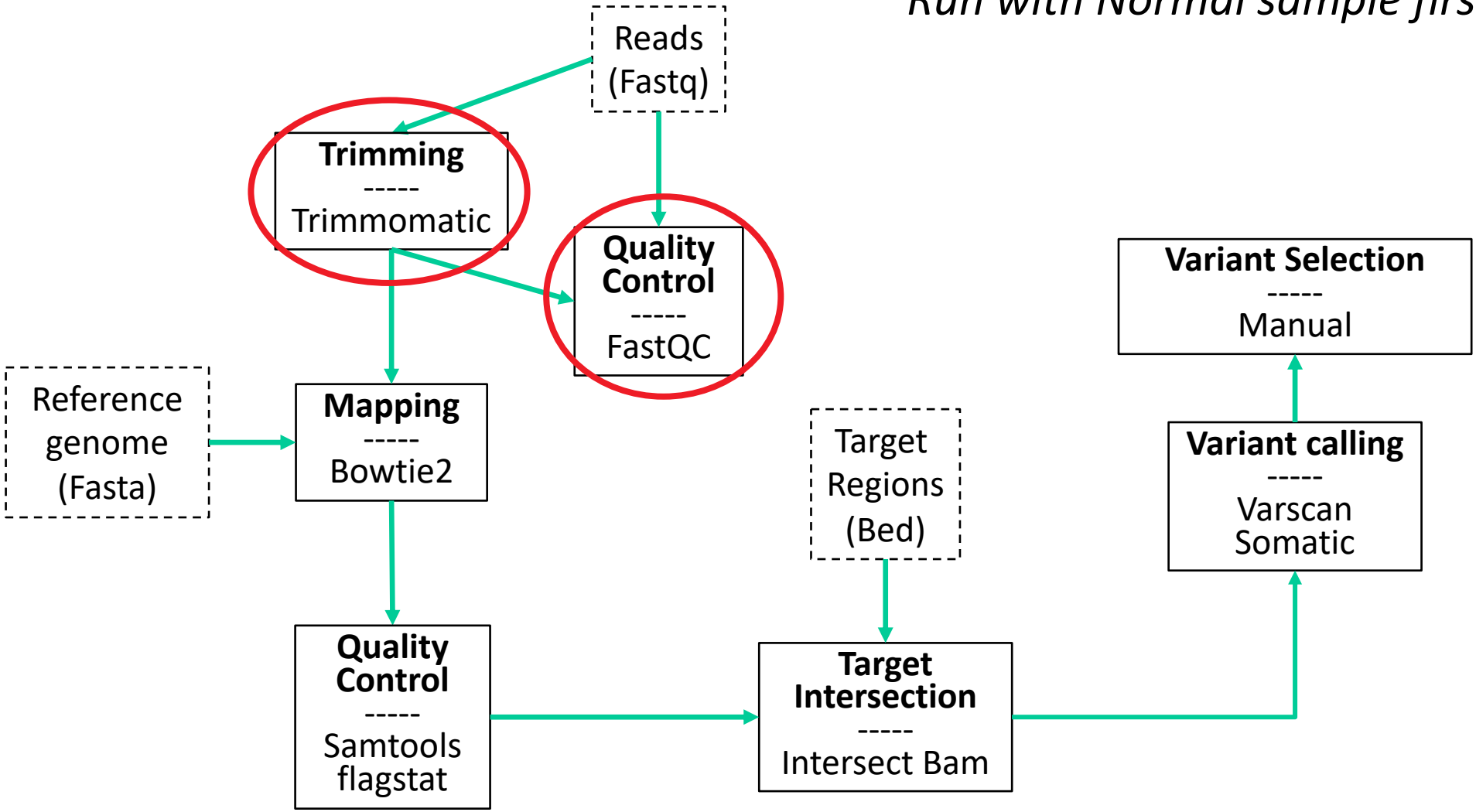
- Menu « Données partagées »
- Histories
- Choisir History « ... IFSBM ...»
- Import History

Fera apparaître:

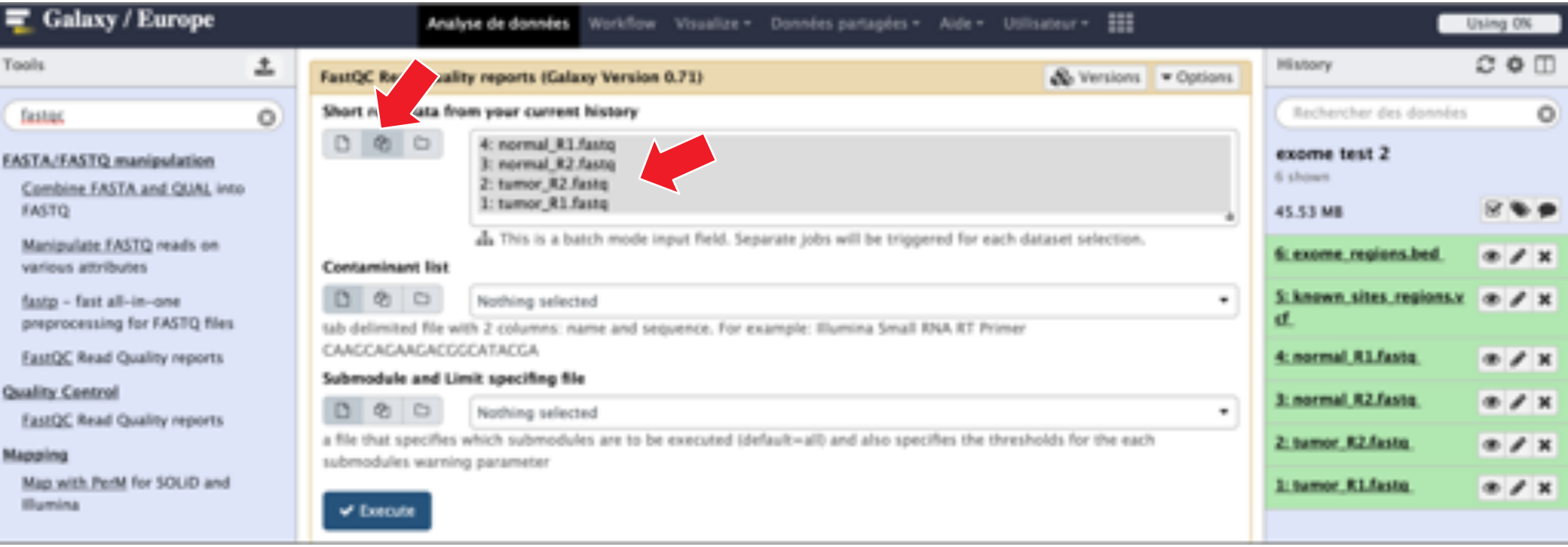


A simplified Variant Pipeline

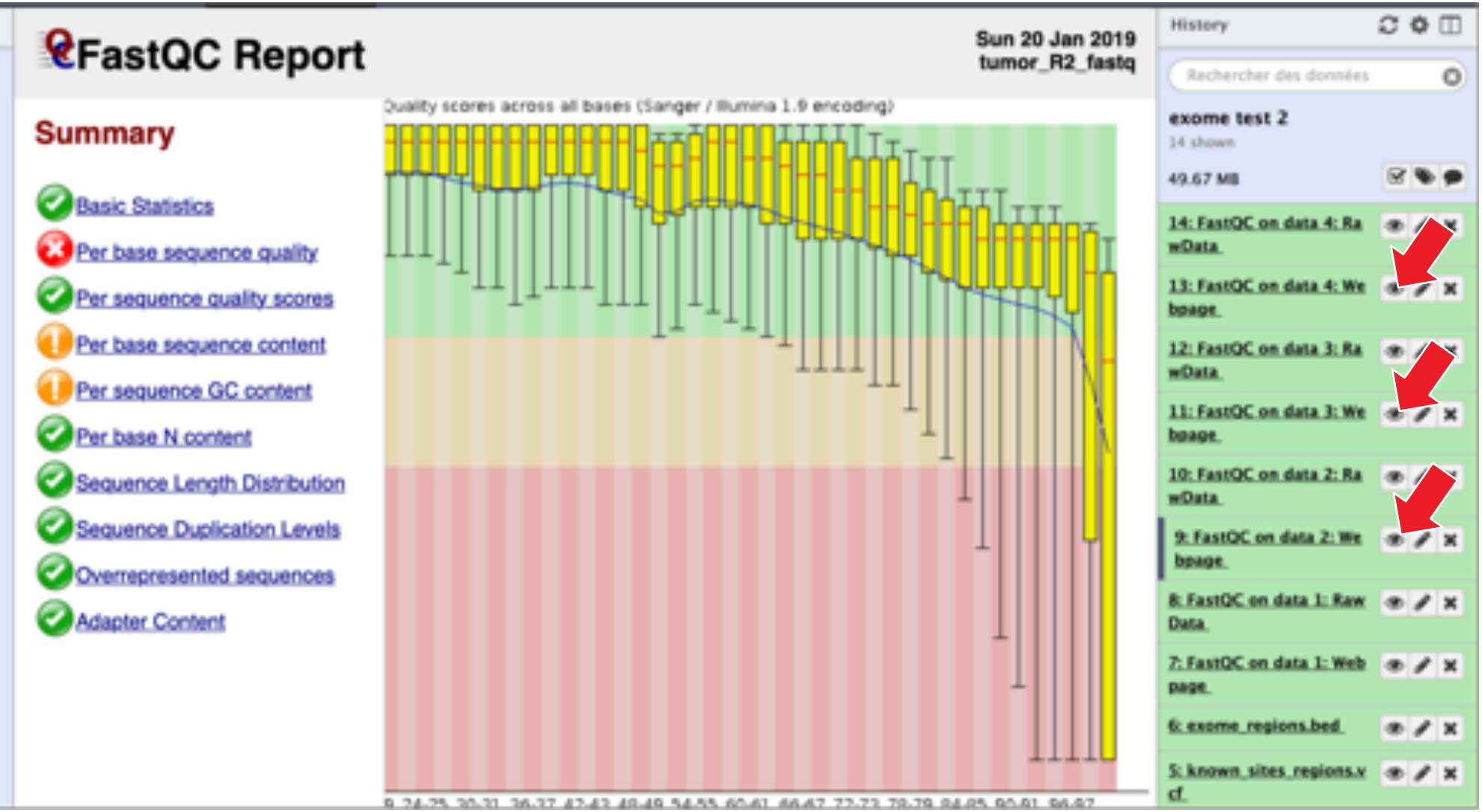
Run with Normal sample first



fastqc

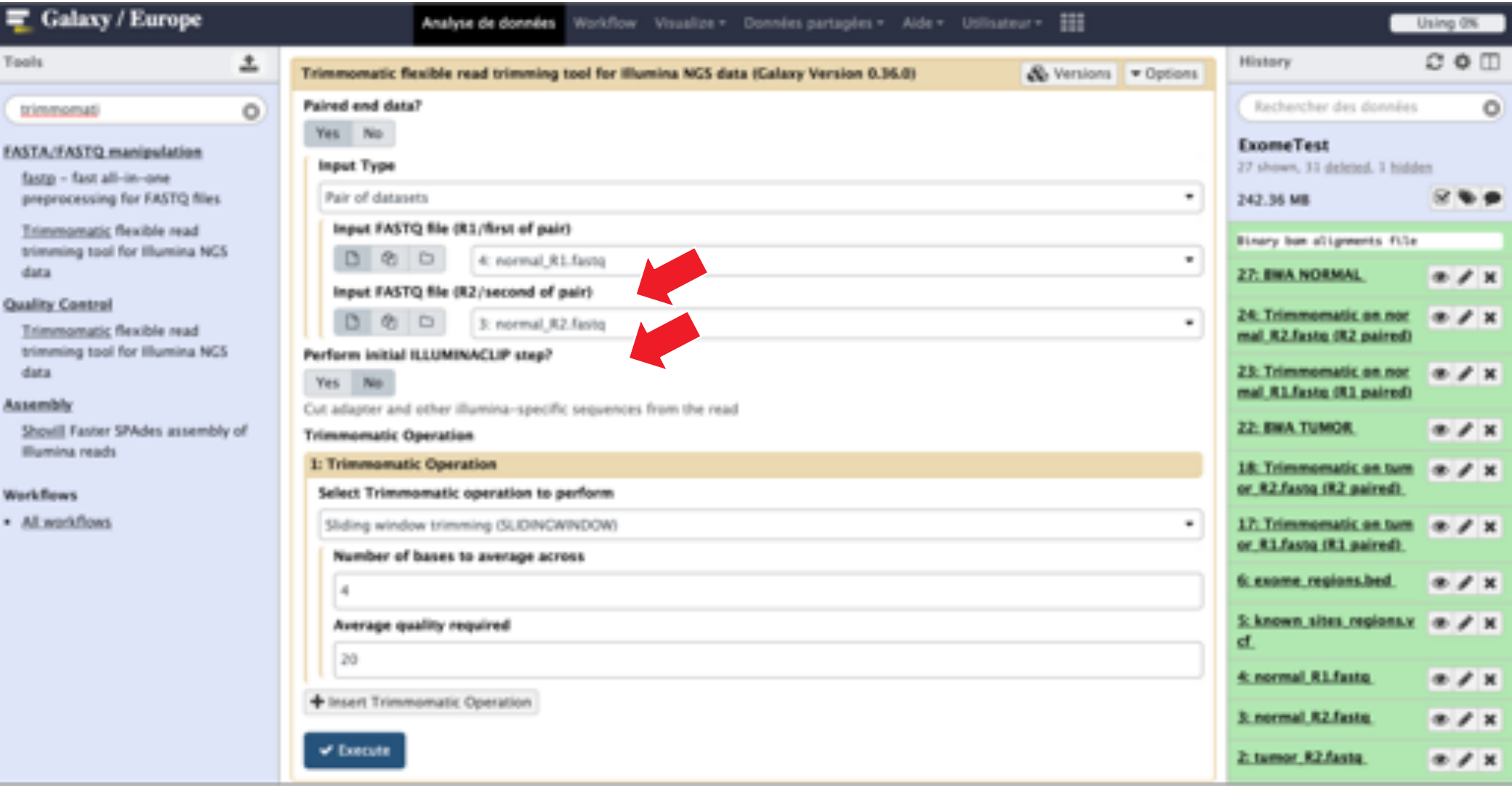

Galaxy / Europe
Analyse de données
Workflow
Visualize
Données partagées
Aide
Utilisateur
Using 0%
Tools
fastqc
FASTA/FASTQ manipulation
Combine FASTA and QUAL into FASTQ
Manipulate FASTQ reads on various attributes
fastp - fast all-in-one preprocessing for FASTQ files
FastQC Read Quality reports
Quality Control
FastQC Read Quality reports
Mapping
Map with PerM for SOLiD and Illumina
FastQC Re ality reports (Galaxy Version 0.71)
Versions
Options
Short r ata from your current history
4: normal_R1.fastq
3: normal_R2.fastq
2: tumor_R2.fastq
1: tumor_R1.fastq
This is a batch mode input field. Separate jobs will be triggered for each dataset selection.
Contaminant list
Nothing selected
tab delimited file with 2 columns: name and sequence. For example: Illumina Small RNA RT Primer
CAACCACAACACCGCATACGA
Submodule and Limit specifing file
Nothing selected
a file that specifies which submodules are to be executed (default=all) and also specifies the thresholds for the each submodules warning parameter
Execute
History
Rechercher des données
exome test 2
6 shown
45.53 MB
6: exome_regions.bed
5: known_sites_regions.vcf
4: normal_R1.fastq
3: normal_R2.fastq
2: tumor_R2.fastq
1: tumor_R1.fastq

Fastqc results



- Look at the different metrics for both reads
- **Problem**: the per base sequence quality of the Read2 are quite low towards the end

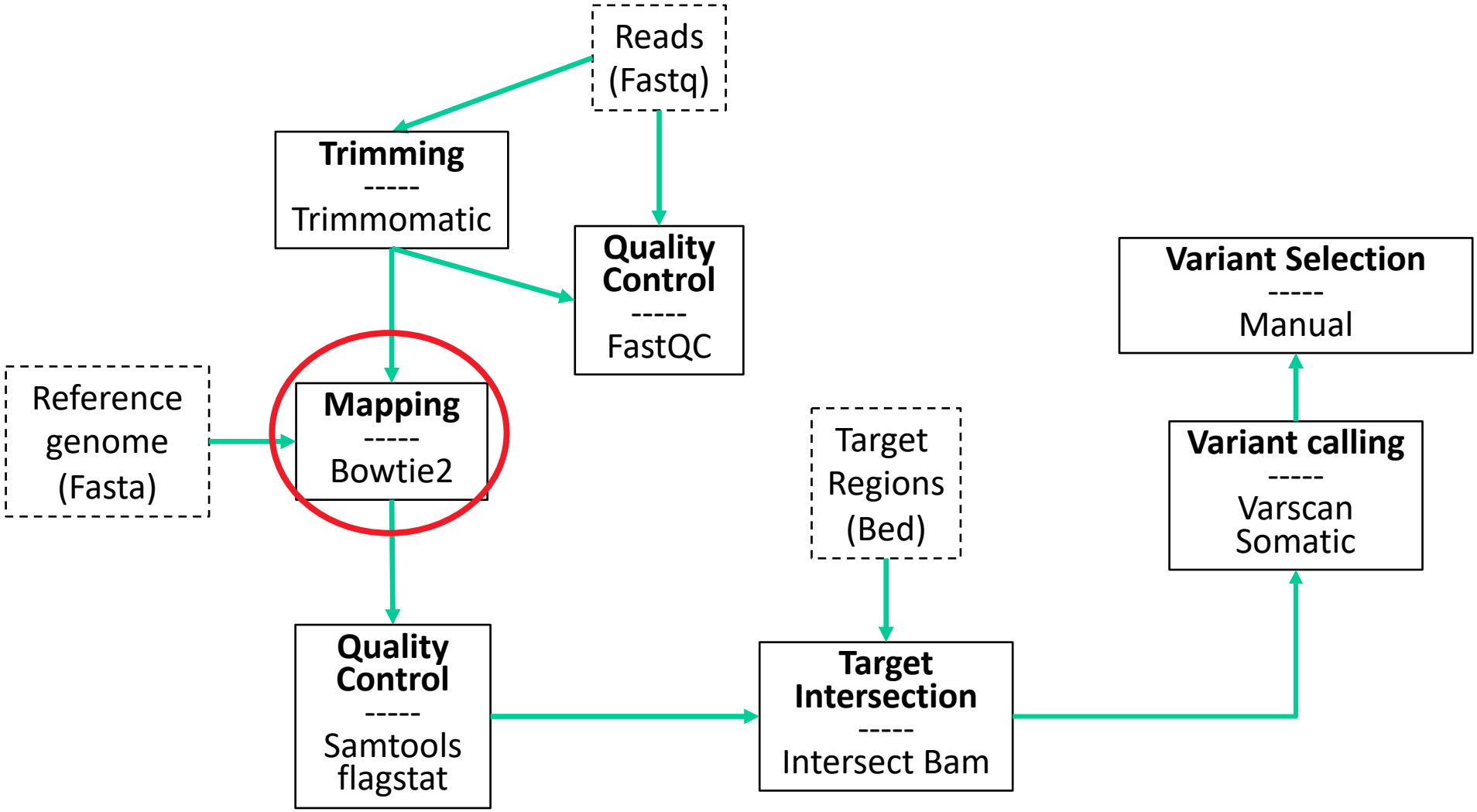
Trimmomatic



Vérifiez à nouveau les fichiers corrigés avec fastqc

Trimmomatic (fin)

- Vérifiez le gain de qualité (fastqc d’un fastq)
- Eliminez les données « unpaired »


Reads
(Fastq)
Trimming

Trimmomatic
Quality
Control

FastQC
Reference
genome
(Fasta)
Mapping

Bowtie2
Quality
Control

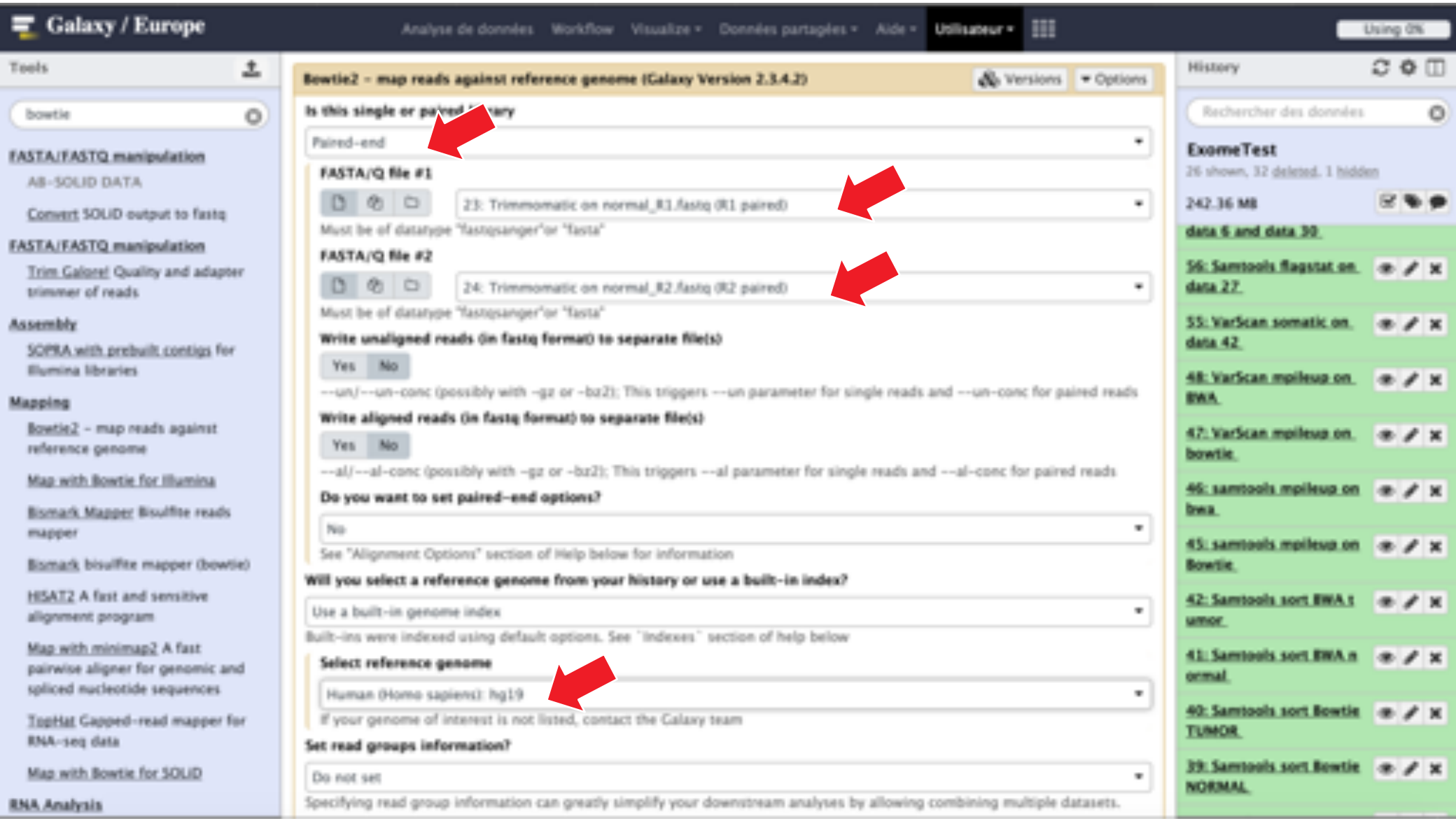
Samtools
flagstat
Target
Regions
(Bed)
Target
Intersection

Intersect Bam
Variant calling

Varscan
Somatic
Variant Selection

Manual

Bowtie

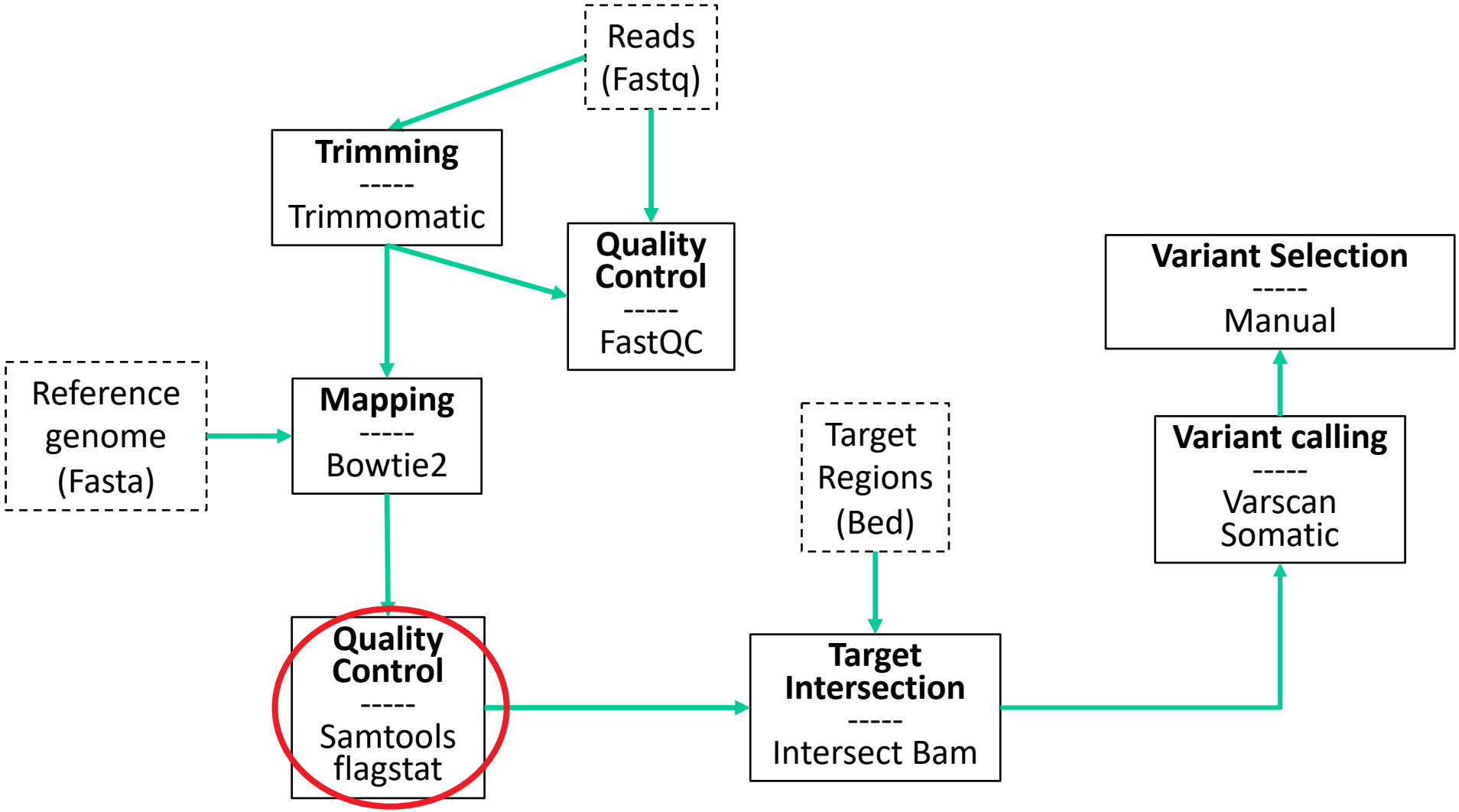

Galaxy / Europe
Analyse de données
Workflow
Visualize
Données partagées
Aide
Utilisateur
Using 0%
Tools
bowtie
FASTA/FASTQ manipulation
AB-SOLID DATA
Convert SOLiD output to fastq
FASTA/FASTQ manipulation
Trim Galore! Quality and adapter trimmer of reads
Assembly
SOPRA with prebuilt contigs for Illumina libraries
Mapping
Bowtie2 - map reads against reference genome
Map with Bowtie for Illumina
Bismark Mapper Bisulfite reads mapper
Bismark bisulfite mapper (bowtie)
HISAT2 A fast and sensitive alignment program
Map with minimap2 A fast pairwise aligner for genomic and spliced nucleotide sequences
TopHat Gapped-read mapper for RNA-seq data
Map with Bowtie for SOLiD
RNA Analysis
Bowtie2 - map reads against reference genome (Galaxy Version 2.3.4.2)
Versions
Options
Is this single or paired library
Paired-end
FASTA/Q file #1
23: Trimmomatic on normal_R1.fastq (R1 paired)
Must be of datatype "fastqsanger" or "fasta"
FASTA/Q file #2
24: Trimmomatic on normal_R2.fastq (R2 paired)
Must be of datatype "fastqsanger" or "fasta"
Write unaligned reads (in fastq format) to separate file(s)
Yes
No
--un/--un-conc (possibly with .gz or .bz2); This triggers --un parameter for single reads and --un-conc for paired reads
Write aligned reads (in fastq format) to separate file(s)
Yes
No
--al/--al-conc (possibly with .gz or .bz2); This triggers --al parameter for single reads and --al-conc for paired reads
Do you want to set paired-end options?
No
See "Alignment Options" section of Help below for information
Will you select a reference genome from your history or use a built-in index?
Use a built-in genome index
Built-ins were indexed using default options. See "Indexes" section of help below
Select reference genome
Human (Homo sapiens): hg19
If your genome of interest is not listed, contact the Galaxy team
Set read groups information?
Do not set
Specifying read group information can greatly simplify your downstream analyses by allowing combining multiple datasets.
History
Rechercher des données
ExomeTest
26 shown, 32 deleted, 1 hidden
242.36 MB
data 6 and data 30
56: Samtools flagstat on data 27
55: VarScan somatic on data 42
48: VarScan mpileup on BWA
47: VarScan mpileup on bowtie
46: samtools mpileup on bwa
45: samtools mpileup on Bowtie
42: Samtools sort BWA tumor
41: Samtools sort BWA normal
40: Samtools sort Bowtie TUMOR
39: Samtools sort Bowtie NORMAL

Reads alignés: le format BAM/SAM

BAM format

Rappel BAM:

```
@RG     ID:group1       SM:1425_CD34    PL:ILLUMINA     LB:lib1 PU:unit1
@PG     ID:bwa  PN:bwa  VN:0.7.12-r1039 CL:bwa mem -M -t 2 -A 2 -E 1 -R @RG\tID:group1\tSM:1425_CD34\tPL:ILLUMINA\tLB:lib1\tPU:unit1 /root/myd
ERR166338.13782800      83      chr13   32890449        60      101M    =       32890343        -207    GGGACTGAATTAGAATTCAAACAAATTTTCCAGCGCTT
ERR166338.13782800      163     chr13   32890343        60      75M     =       32890449        207     CACTAGCCACGTTTCGAGTGCTTAATGTGGCTAGTGGC
ERR166338.26716588      99      chr13   32890406        60      101M    =       32890553        222     AATGTTCCCATCCTCACAGTAAGCTGTTACCGTTCCAG
ERR166338.26716588      147     chr13   32890553        60      75M     =       32890406        -222    TTGCAGACTTATTTACCAAGCATTGGAGGAATATCGTA
ERR166338.27259961      99      chr13   32890496        60      101M    =       32890558        137     ACCTCAGTCACATAATAAGGAATGCATCCCTGTGTAAG
ERR166338.27259961      147     chr13   32890558        60      75M     =       32890496        -137    GACTTATTTACCAAGCATTGGAGGAATATCGTAGGTAA
ERR166338.63037998      99      chr13   32890496        60      101M    =       32890558        137     ACCTCAGTCACATAATAAGGAATGCATCCCTGTGTAAG
ERR166338.63037998      147     chr13   32890558        60      75M     =       32890496        -137    GACTTATTTACCAAGCATTGGAGGAATATCGTAGGTAA
```


Reads
(Fastq)
Trimming

Trimmomatic
Quality
Control

FastQC
Variant Selection

Manual
Reference
genome
(Fasta)
Mapping

Bowtie2
Target
Regions
(Bed)
Variant calling

Varscan
Somatic
Quality
Control

Samtools
flagstat
Target
Intersection

Intersect Bam

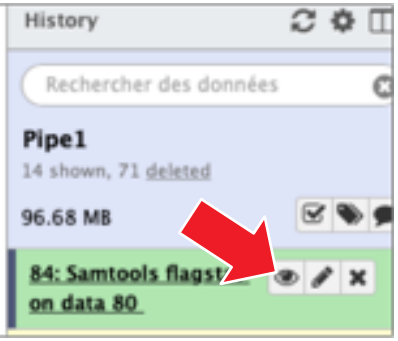
Samtools

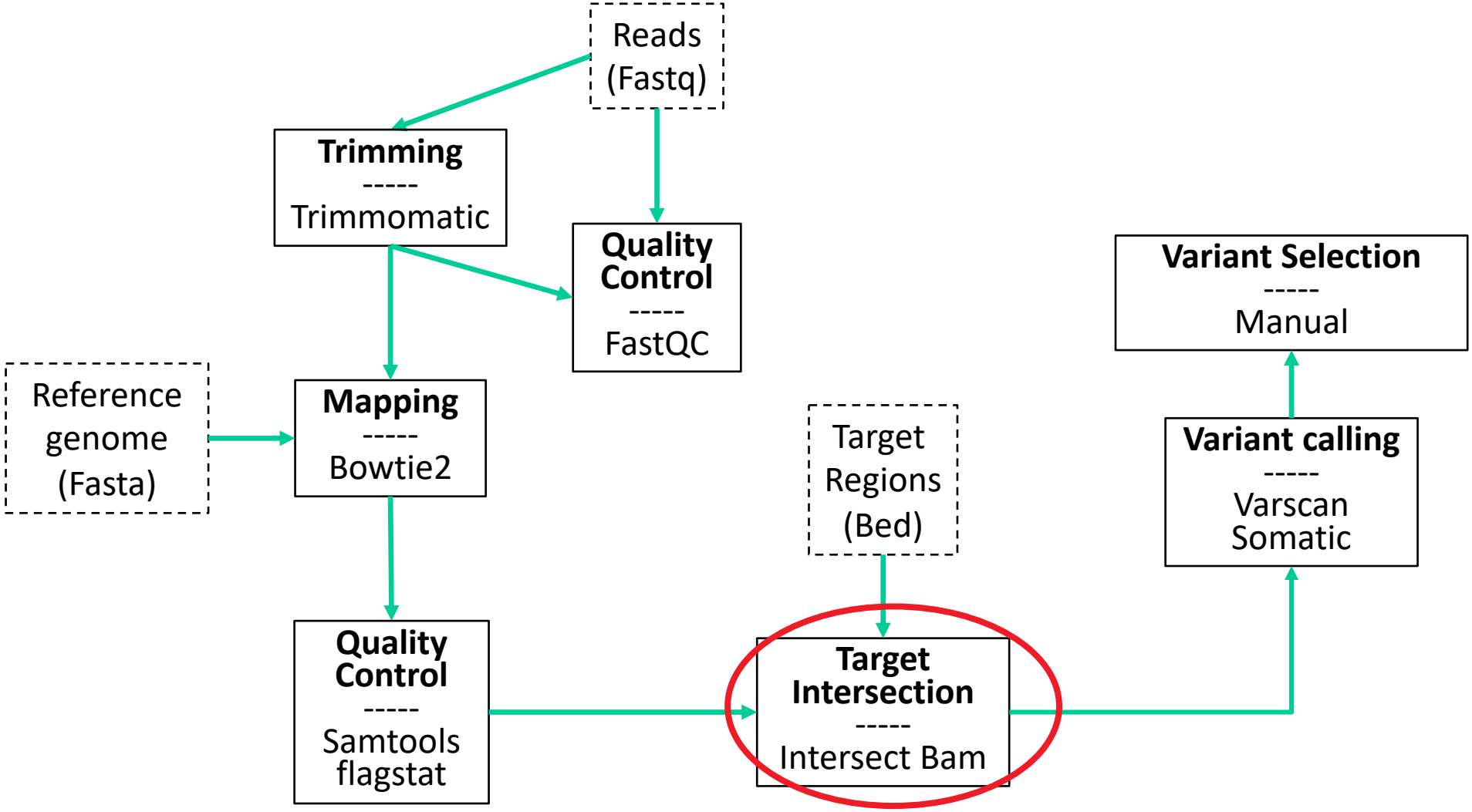
- La boîte à outils pour traiter les BAMs/SAMs
  - BAM <-> SAM
  - BAM <-> FASTQ
  - Tri de BAM
  - Indexation du BAM (création fichier .bai)

Samtools flatgstats

résultat

```
86796 + 0 in total (QC-passed reads + QC-failed reads)
0 + 0 secondary
0 + 0 supplementary
0 + 0 duplicates
86738 + 0 mapped (99.93% : N/A)
86796 + 0 paired in sequencing
43398 + 0 read1
43398 + 0 read2
86152 + 0 properly paired (99.26% : N/A)
86706 + 0 with itself and mate mapped
32 + 0 singletons (0.04% : N/A)
76 + 0 with mate mapped to a different chr
19 + 0 with mate mapped to a different chr (mapQ>=5)
```




Reads
(Fastq)
Trimming

Trimmomatic
Quality
Control

FastQC
Variant Selection

Manual
Reference
genome
(Fasta)
Mapping

Bowtie2
Target
Regions
(Bed)
Variant calling

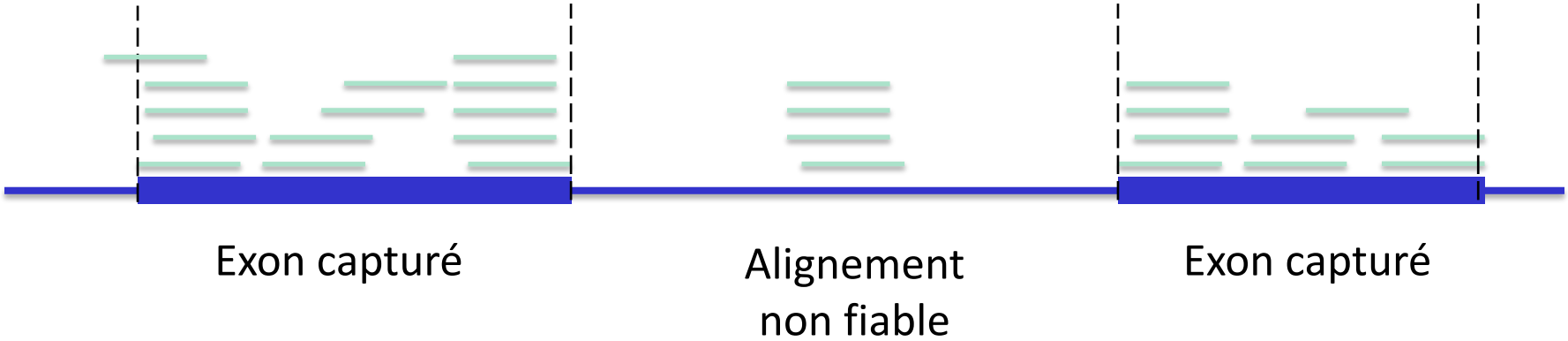
Varscan
Somatic
Quality
Control

Samtools
flagstat
Target
Intersection

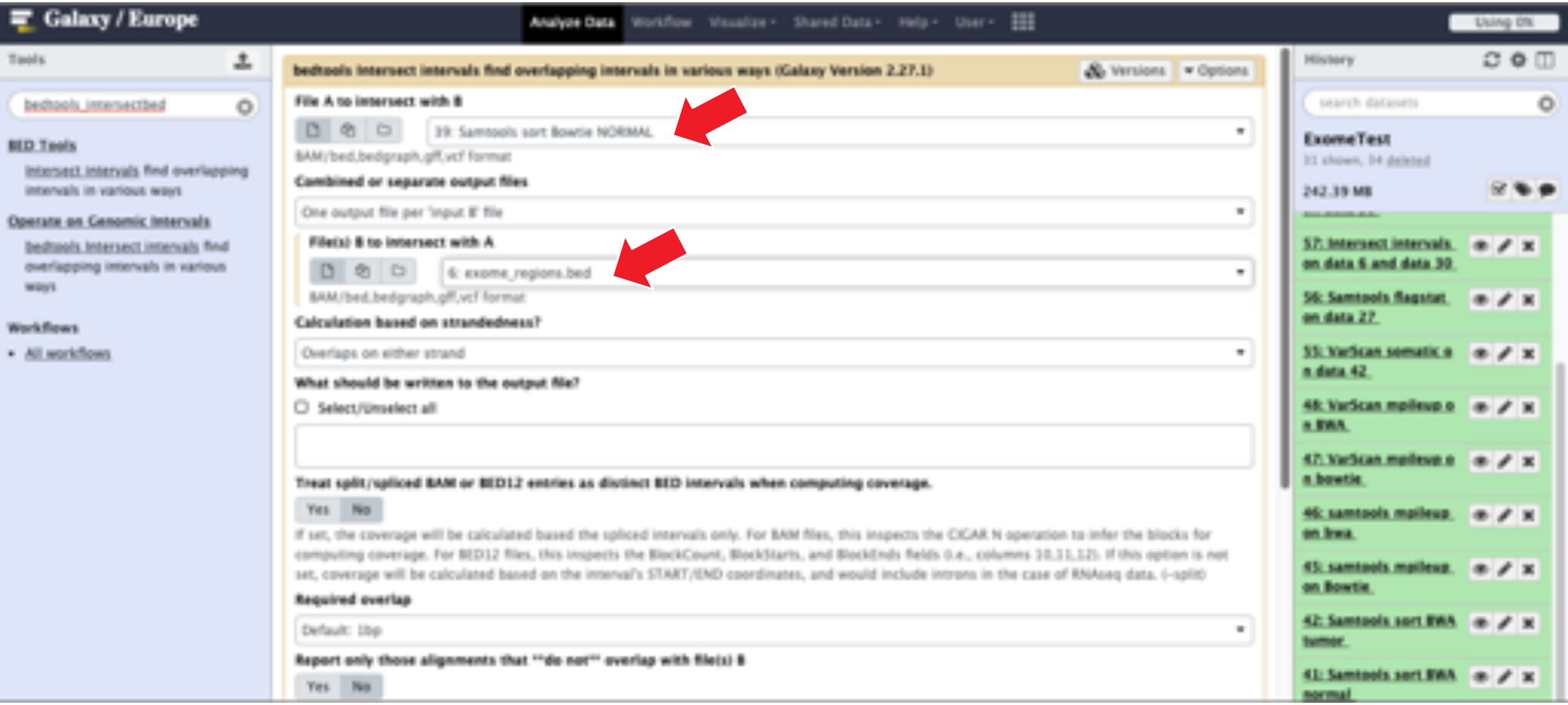
Intersect Bam

Target intersection

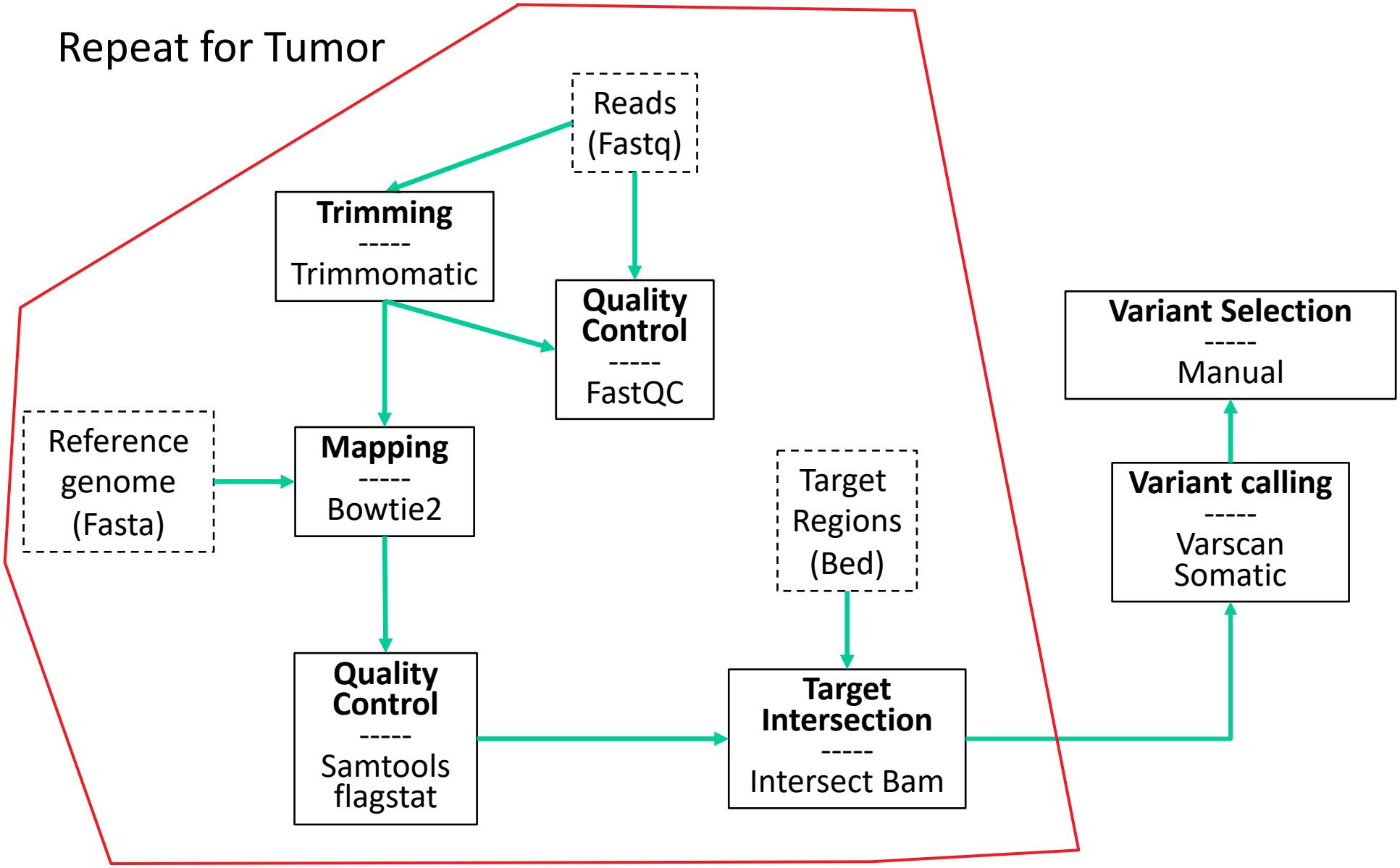
- Comparer l'alignement obtenu à la liste des positions visées par le protocole de capture



Bedtools intersect



Vérifiez la réduction de taille du fichier BAM!


Repeat for Tumor
Reads
(Fastq)
Trimming

Trimmomatic
Quality
Control

FastQC
Reference
genome
(Fasta)
Mapping

Bowtie2
Quality
Control

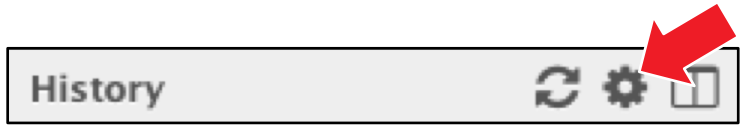
Samtools
flagstat
Target
Regions
(Bed)
Target
Intersection

Intersect Bam
Variant calling

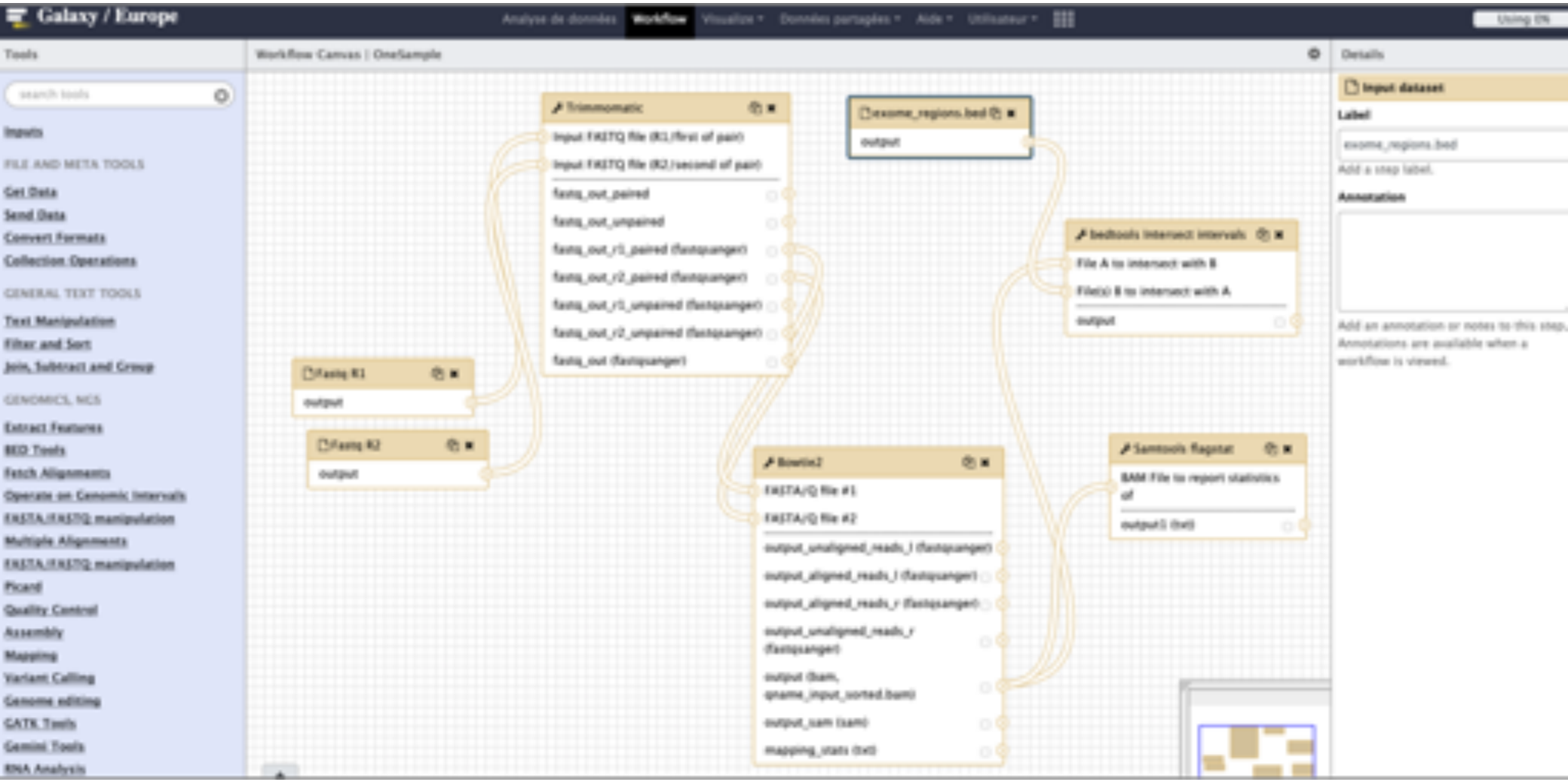
Varscan
Somatic
Variant Selection

Manual

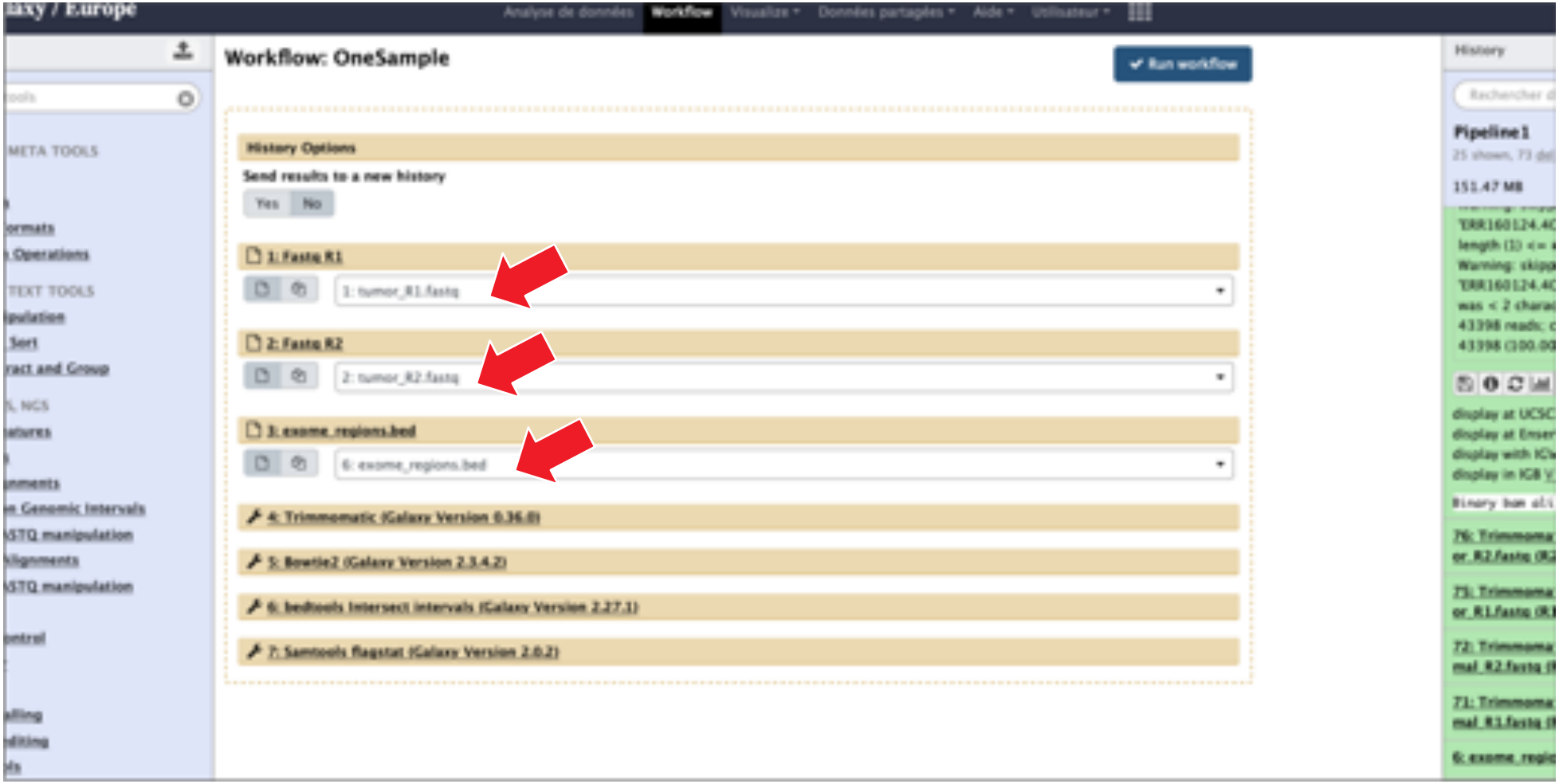
Extraire un workflow

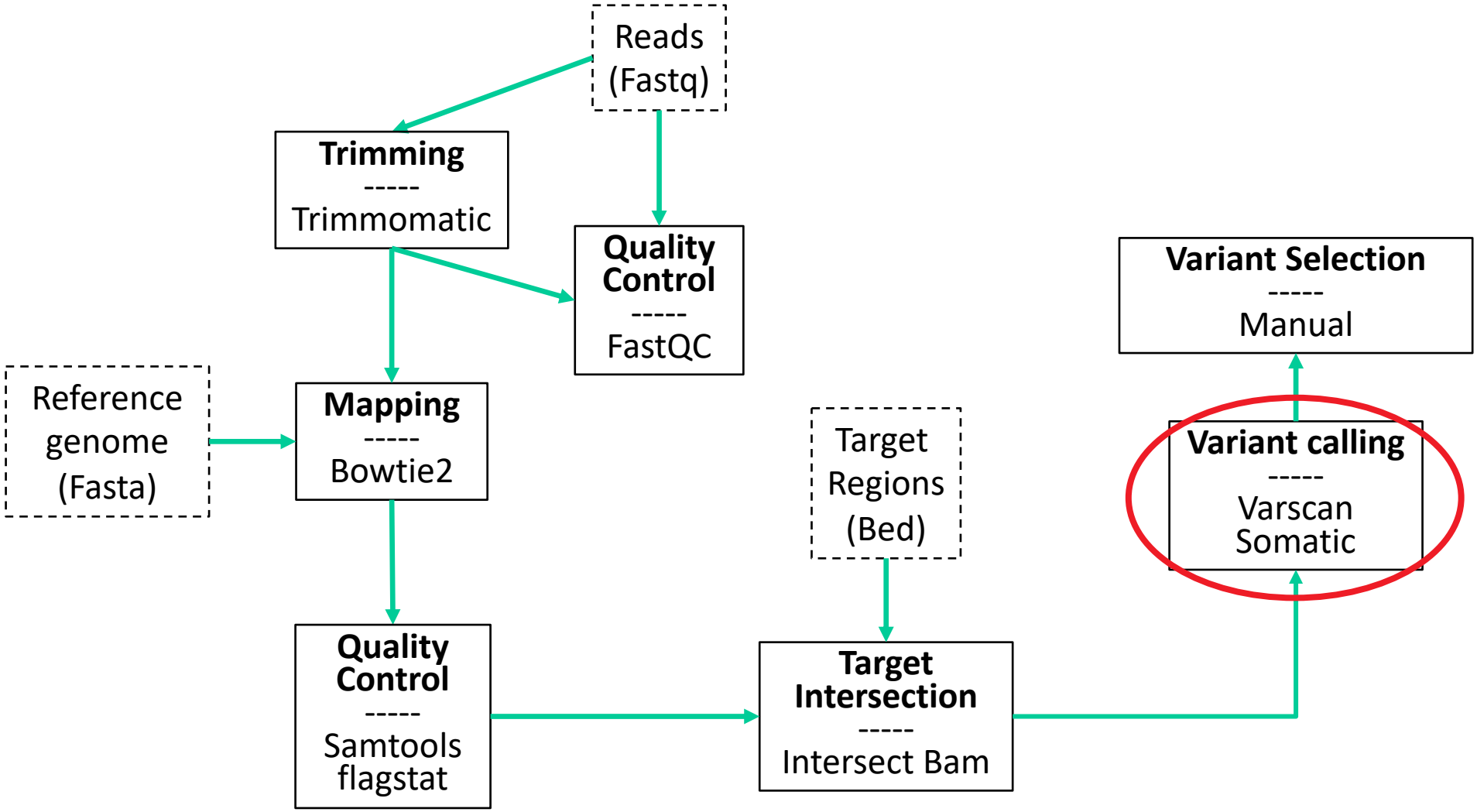


- Extraire workflow
- Le nommer
- Choisir les données pertinentes (juste 2 fastq et regions.bed)
- Choisir les étapes de Trimmomatic à Intersect bed
- Enlever les data inutilisées (fastq tumor)
- Renommer les objets de façon générique (« sample » plutot que « normal »)
- Puis ⚙ save workflow


Galaxy / Europe
Analyse de données
Workflow
Visualize
Données partagées
Aide
Utilisateur
Using 0%
Tools
Workflow Canvas | OneSample
Details
search tools
Inputs
FILE AND META TOOLS
Get Data
Send Data
Convert Formats
Collection Operations
GENERAL TEXT TOOLS
Text Manipulation
Filter and Sort
Join, Subtract and Group
GENOMICS, NGS
Extract Features
BED Tools
Fetch Alignments
Operate on Genomic Intervals
FASTA/FASTQ manipulation
Multiple Alignments
FASTA/FASTQ manipulation
Picard
Quality Control
Assembly
Mapping
Variant Calling
Genome editing
GATK Tools
Gemini Tools
RNA Analysis
Trimmomatic
Input FASTQ file (R1/first of pair)
Input FASTQ file (R2/second of pair)
fastq_out_paired
fastq_out_unpaired
fastq_out_r1_paired (fastqsanger)
fastq_out_r2_paired (fastqsanger)
fastq_out_r1_unpaired (fastqsanger)
fastq_out_r2_unpaired (fastqsanger)
fastq_out (fastqsanger)
exome_regions.bed
output
bedtools Intersect intervals
File A to intersect with B
File(s) B to intersect with A
output
Fastq R1
output
Fastq R2
output
Bowtie2
FASTA/Q file #1
FASTA/Q file #2
output_unaligned_reads_l (fastqsanger)
output_aligned_reads_l (fastqsanger)
output_aligned_reads_r (fastqsanger)
output_unaligned_reads_r (fastqsanger)
output (bam, qname_input_sorted.bam)
output_sam (sam)
mapping_stats (txt)
Samtools flagstat
BAM File to report statistics of
output1 (txt)
Input dataset
Label
exome_regions.bed
Add a step label.
Annotation
Add an annotation or notes to this step. Annotations are available when a workflow is viewed.

Maintenant lancez le workflow sur les données Tumor (run workflow)




Reads
(Fastq)
Trimming

Trimmomatic
Quality
Control

FastQC
Variant Selection

Manual
Reference
genome
(Fasta)
Mapping

Bowtie2
Target
Regions
(Bed)
Variant calling

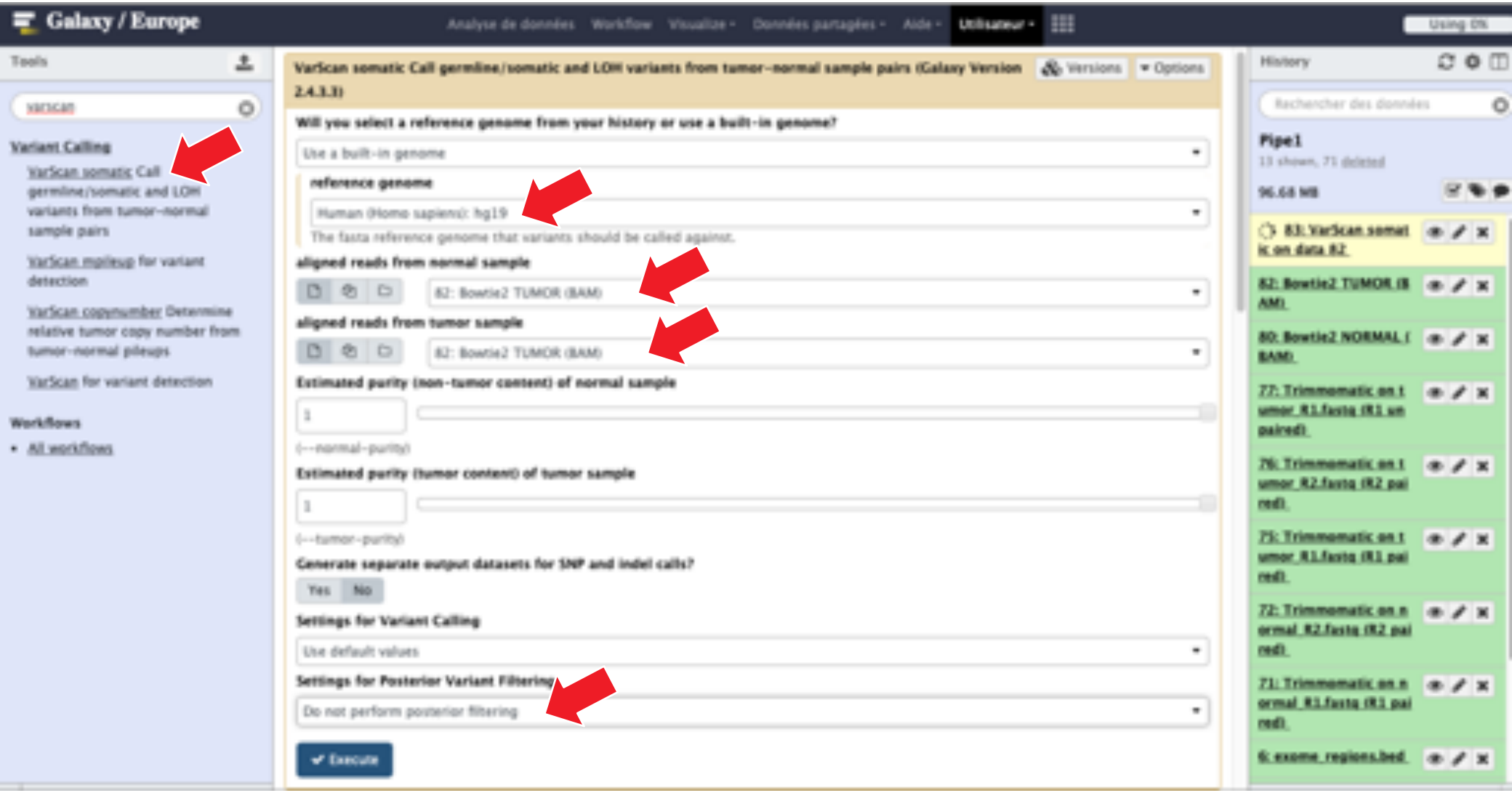
Varscan
Somatic
Quality
Control

Samtools
flagstat
Target
Intersection

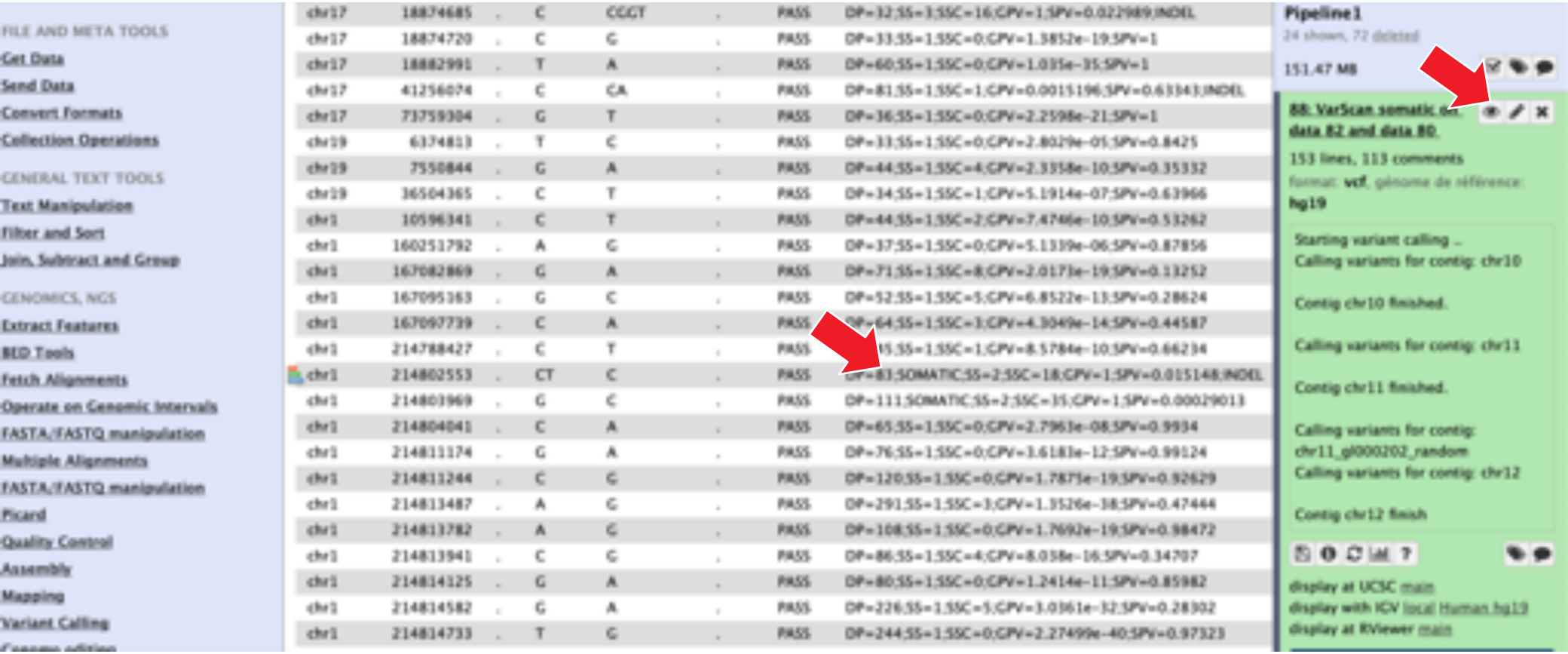
Intersect Bam

Somatic variant calling: Varscan

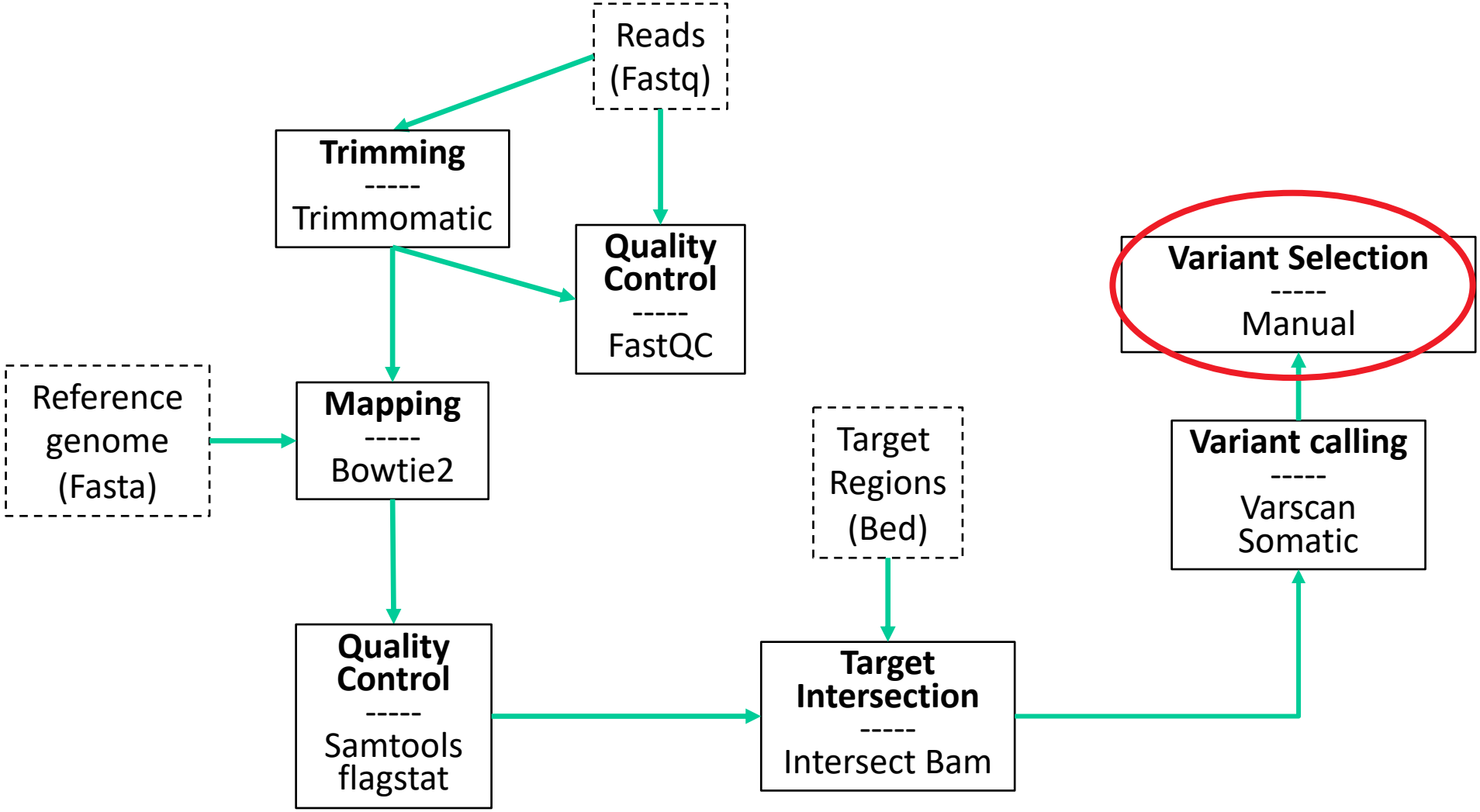
Attention: étape de 30min!



Check Varscan output



How many variants? Somatic variants found?


Reads
(Fastq)
Trimming

Trimmomatic
Quality
Control

FastQC
Variant Selection

Manual
Reference
genome
(Fasta)
Mapping

Bowtie2
Target
Regions
(Bed)
Variant calling

Varscan
Somatic
Quality
Control

Samtools
flagstat
Target
Intersection

Intersect Bam

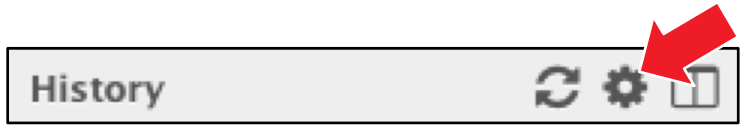
Filter and visualize somatic variants

- Run the *grep* filter on the Varscan output with regulat expression « somatic ». Check the result
- Download Normal and Tumor BAM files on your local computer
  - (select option « download bam_index »)
- Launch IGV with hg19 reference
- In IGV, load normal and tumor BAM files
- Visualize somatic events.

IGV view

Annexes

Galaxy: partager ses données



- Partager et publier
- Make History Accessible via Link
  - Cocher « also make all objects within the History accessible »

Galaxy: récupérer des données partagées

- Menu « Données partagées »
- Histories
- Choisir History « ... IFSBM ...»
- Import History