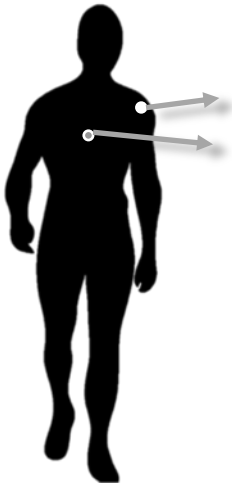
Réalisation d'un Pipeline d'analyse d'exome

UNIVERSITÉ PARIS SUD

V. 21/1/2020

GUSTAVE ROUSSY
CANCER CAMPUS GRAND PARIS

Les données



Normal tissue (blood)

Tumor tissue (non small cell lung cancer)

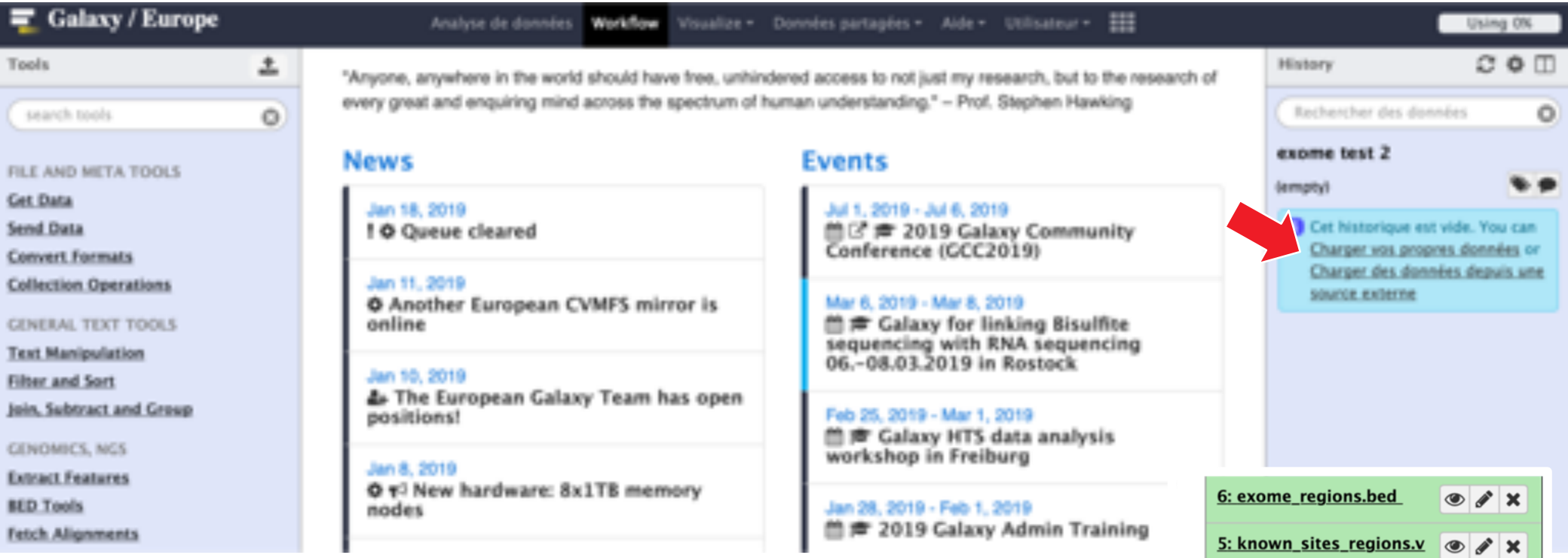
Ju et al. Genome Res. 22:436–445, 2012
100bp paired-end reads, Illumina HiSeq 2000
SRA (Sequence Read Archive): ERA148528

- 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 Time: ~20h
- **Fastq files restricted to a few regions (~112kb) to limit processing time**

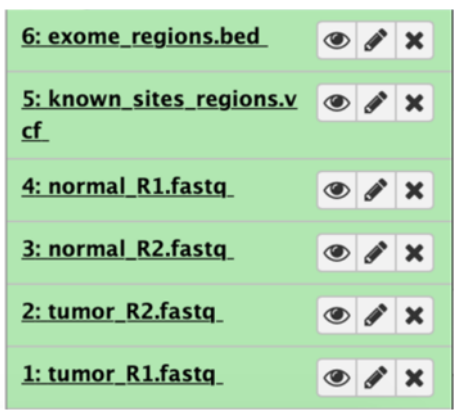
Galaxy servers

- https://usegalaxy.eu
- https://usegalaxy.fr

Chargez vos données

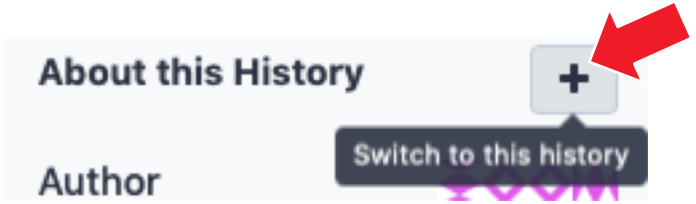


Fera apparaître:

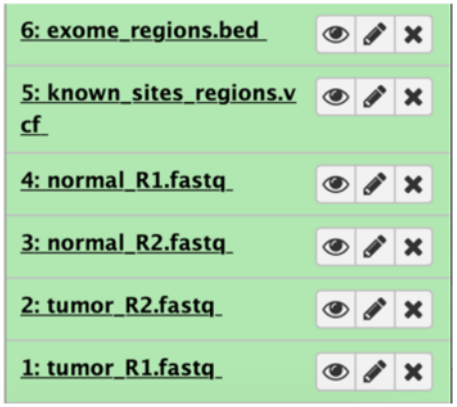


(alternativement: à partir de données partagées)

- Menu « Données partagées »
- Histories
- Choisir History « ... IFSBM ...»
- Click on history, then "+"

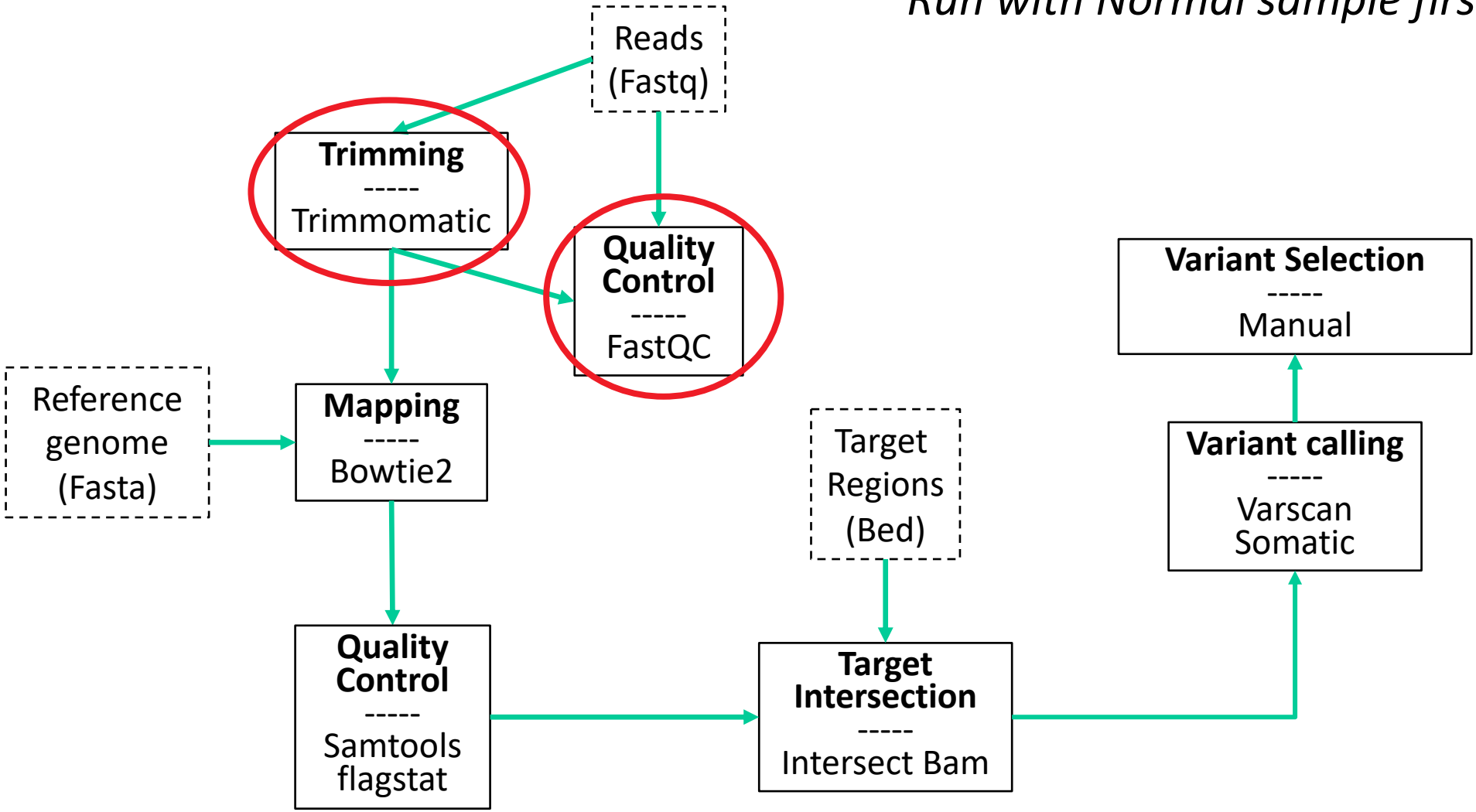


Fera apparaître:

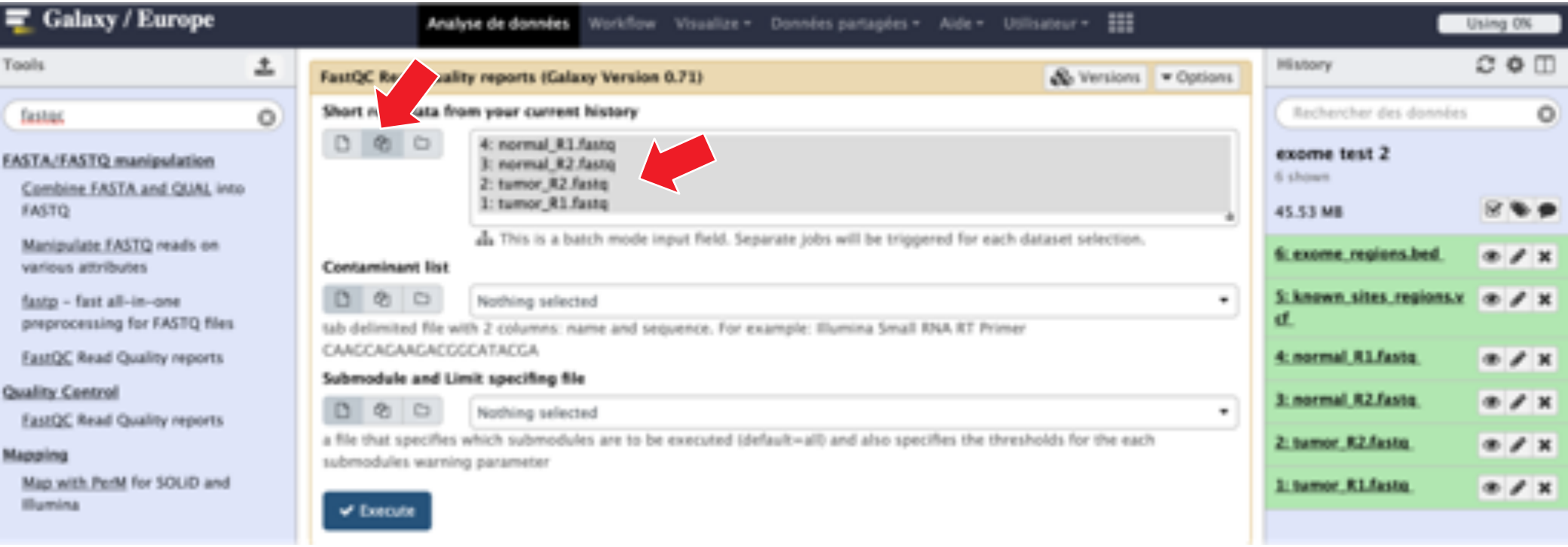


A simplified Variant Pipeline

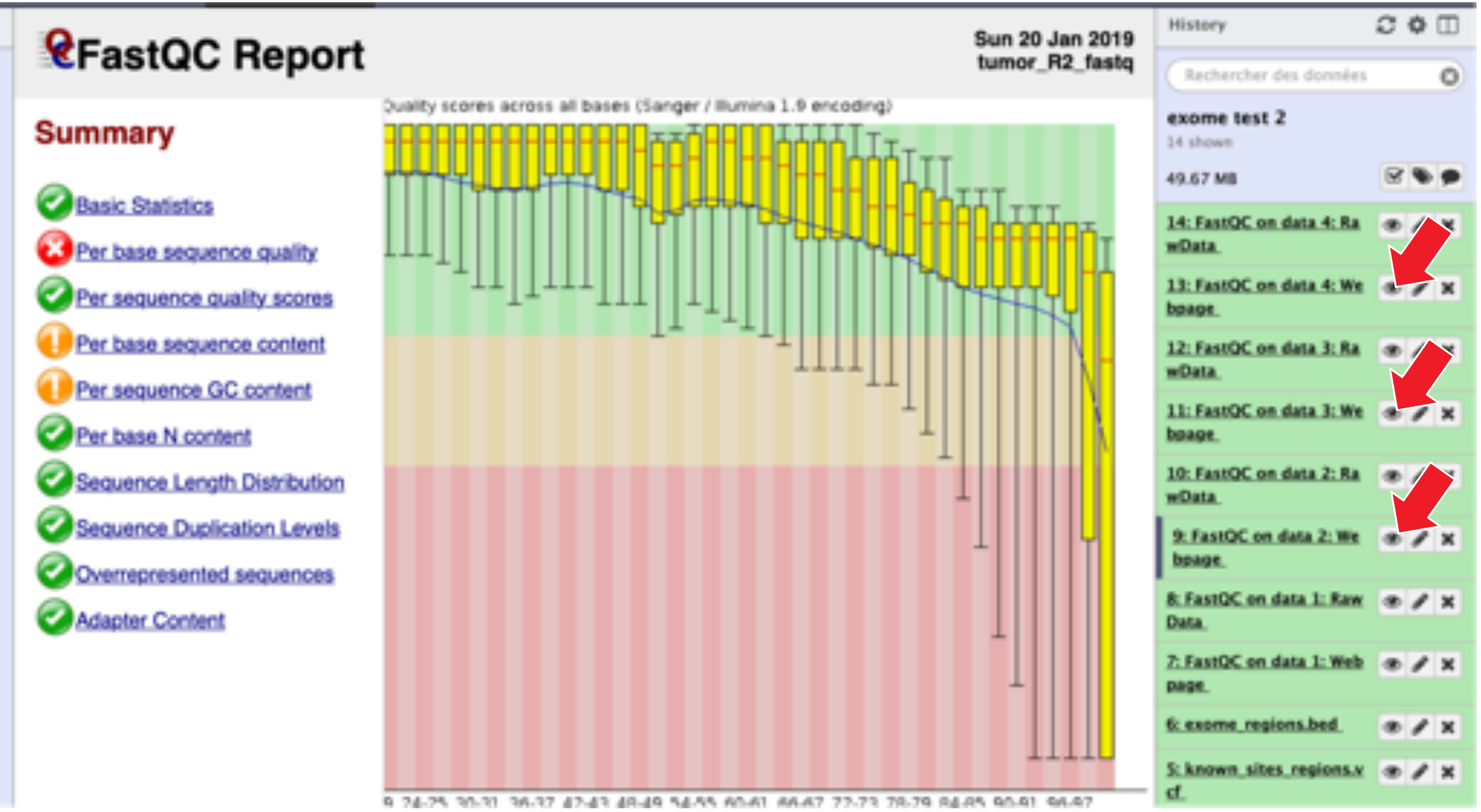
Run with Normal sample first



fastqc

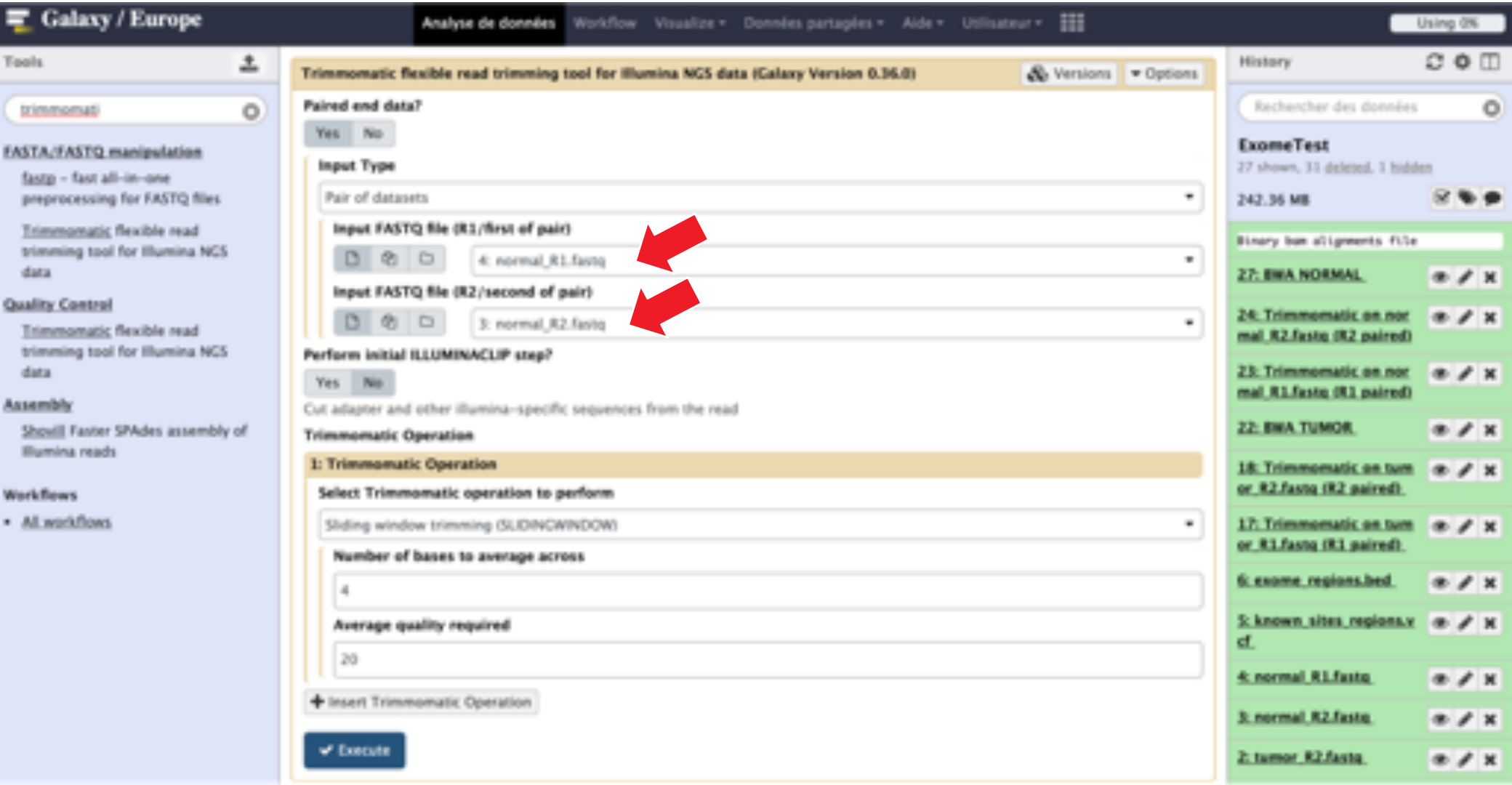

Galaxy / Europe
Analyse de données
Workflow
Visualize
Données partagées
Aide
Utilisateur
Using 0%
Tools
fastqc
FASTA/FASTQ manipulation
Combine FASTA and QUAL into FASTQ
Manipulate FASTQ reads on various attributes
fastp – fast all-in-one preprocessing for FASTQ files
FastQC Read Quality reports
Quality Control
FastQC Read Quality reports
Mapping
Map with PerM for SOLID and Illumina
Versions
Options
Short
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
CAAGCAGAAGACGGCATACGA
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

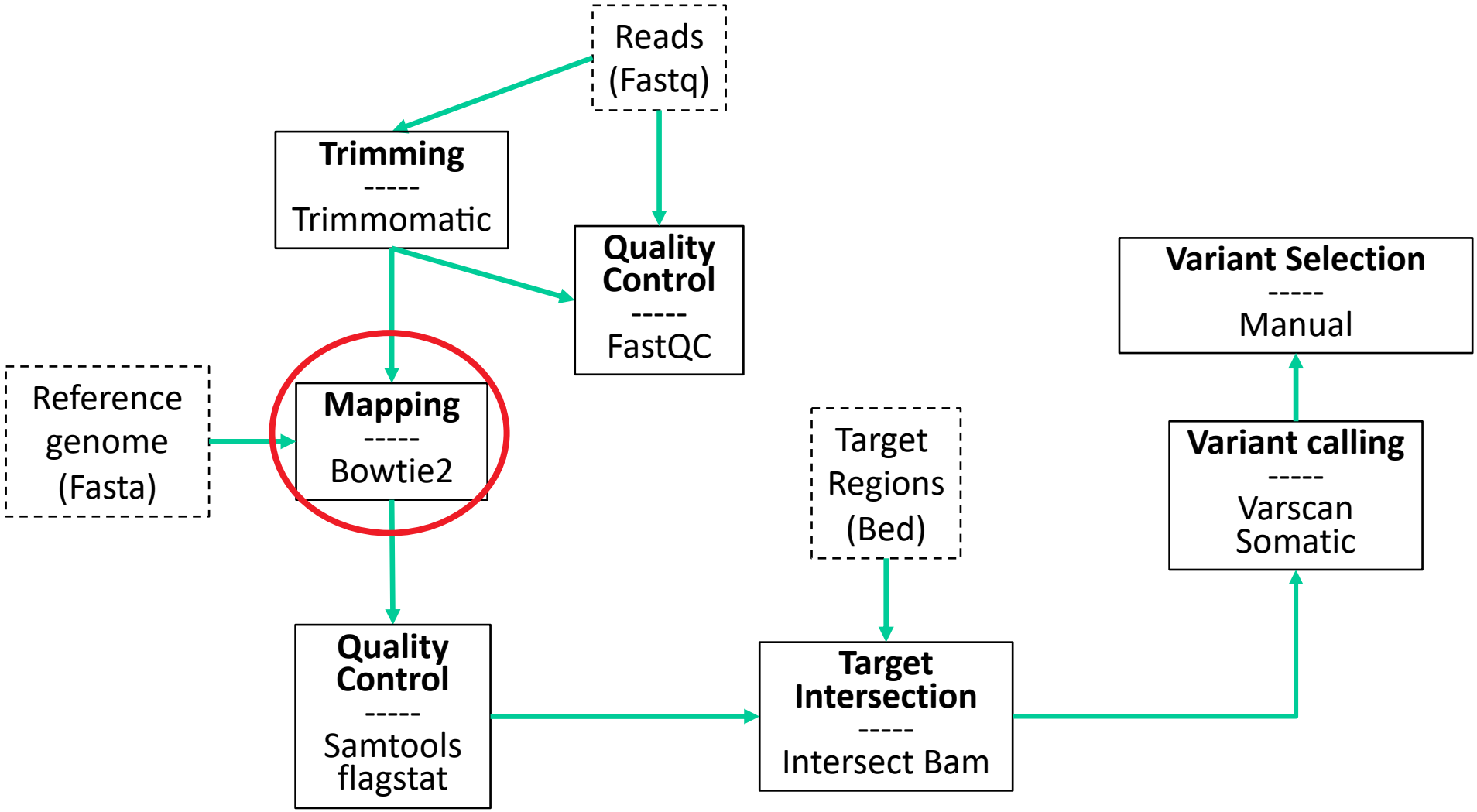
Trimmomatic



Vérifiez à nouveau les fichiers corrigés avec fastqc

Trimmomatic (fin)

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


Reads
(Fastq)
Trimming

Trimmomatic
Quality
Control

FastQC
Reference
genome
(Fasta)
Mapping

Bowtie2
Target
Regions
(Bed)
Variant Selection

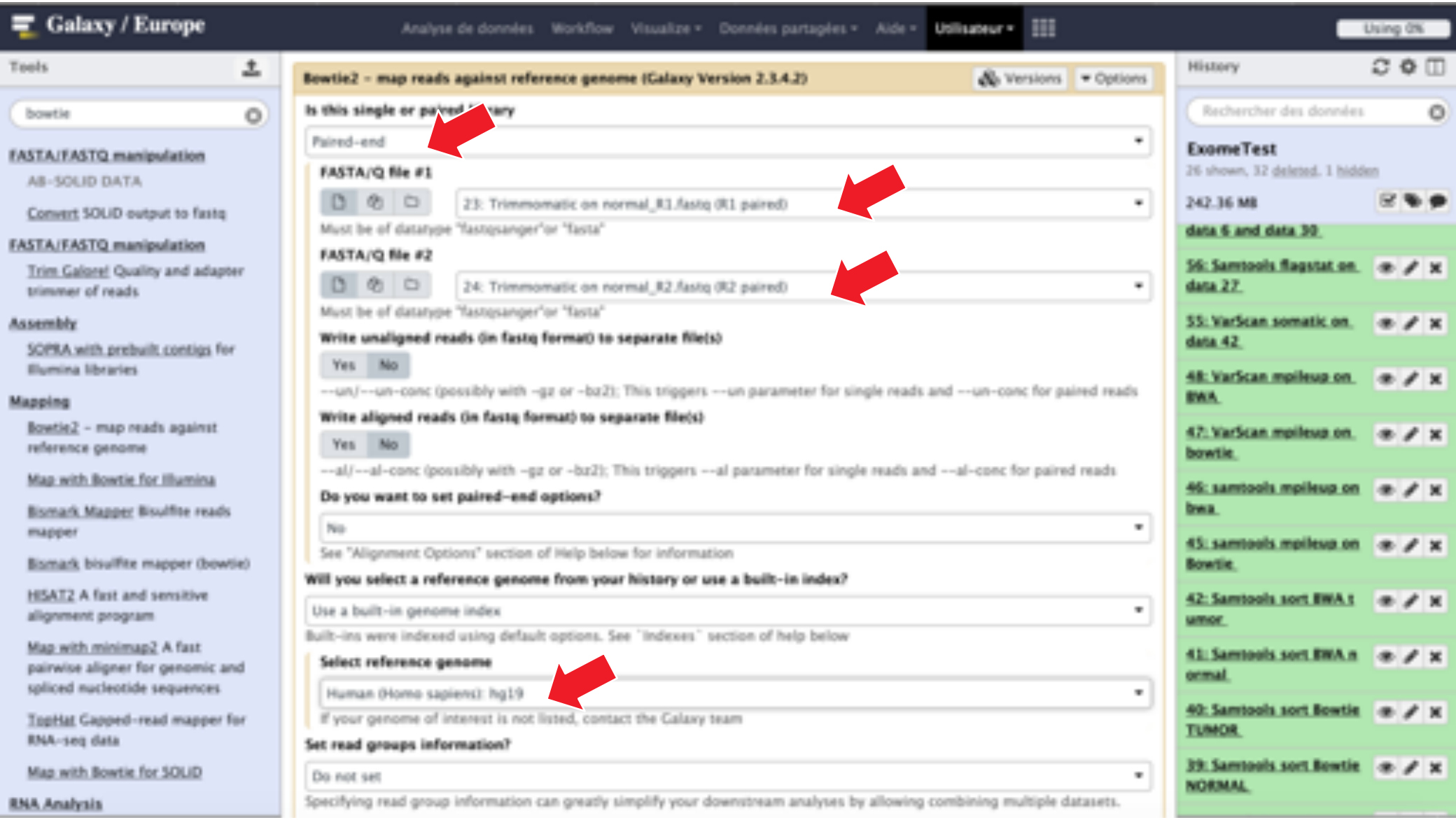
Manual
Variant calling

Varscan
Somatic
Quality
Control

Samtools
flagstat
Target
Intersection

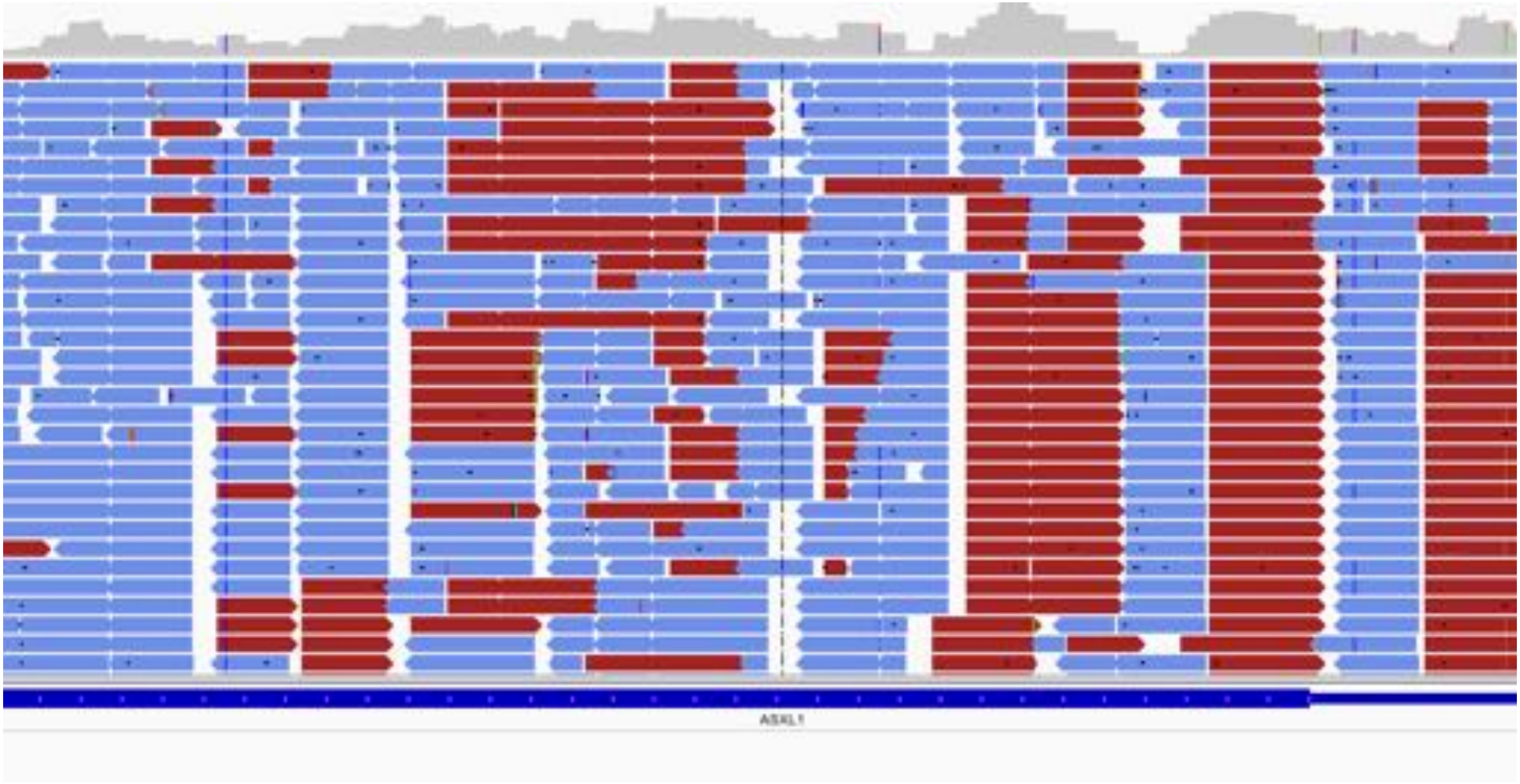
Intersect Bam

Bowtie



Check Bowtie result: what type of file is it?

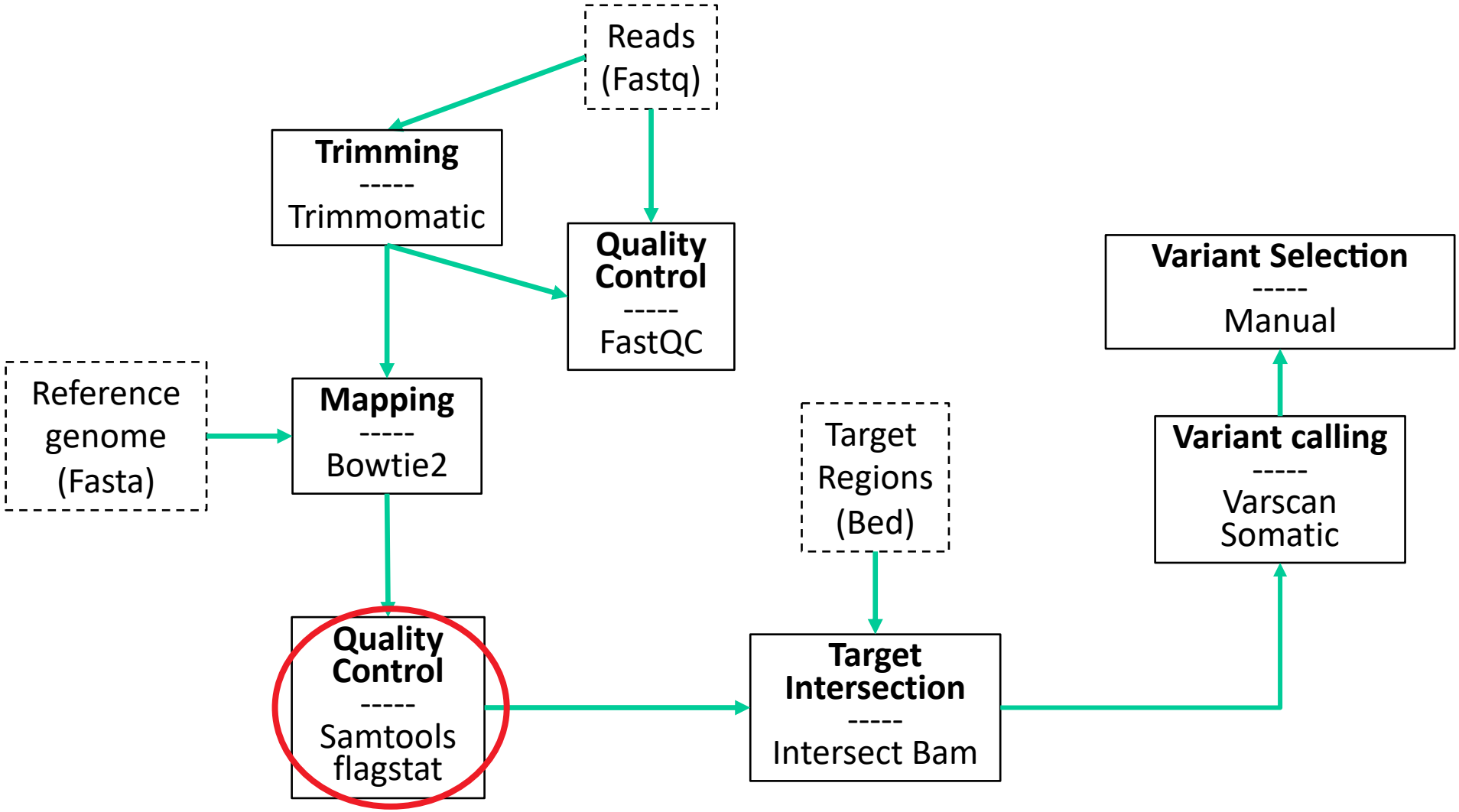
Reads alignés: le format BAM/SAM



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
Reference
genome
(Fasta)
Mapping

Bowtie2
Target
Regions
(Bed)
Variant Selection

Manual
Variant calling

Varscan
Somatic
Quality
Control

Samtools
flagstat
Target
Intersection

Intersect Bam

Samtools

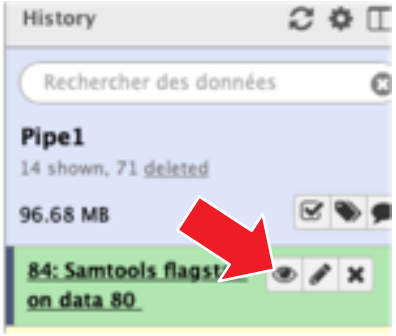
- La boîte à outils pour traiter les BAMs/SAMs
  - BAM <-> SAM
  - BAM <-> FASTQ
  - Tri de BAM
  - Indexation du BAM (création fichier .bai)
  - Obtenir un rapport sur le BAM (flagstat)

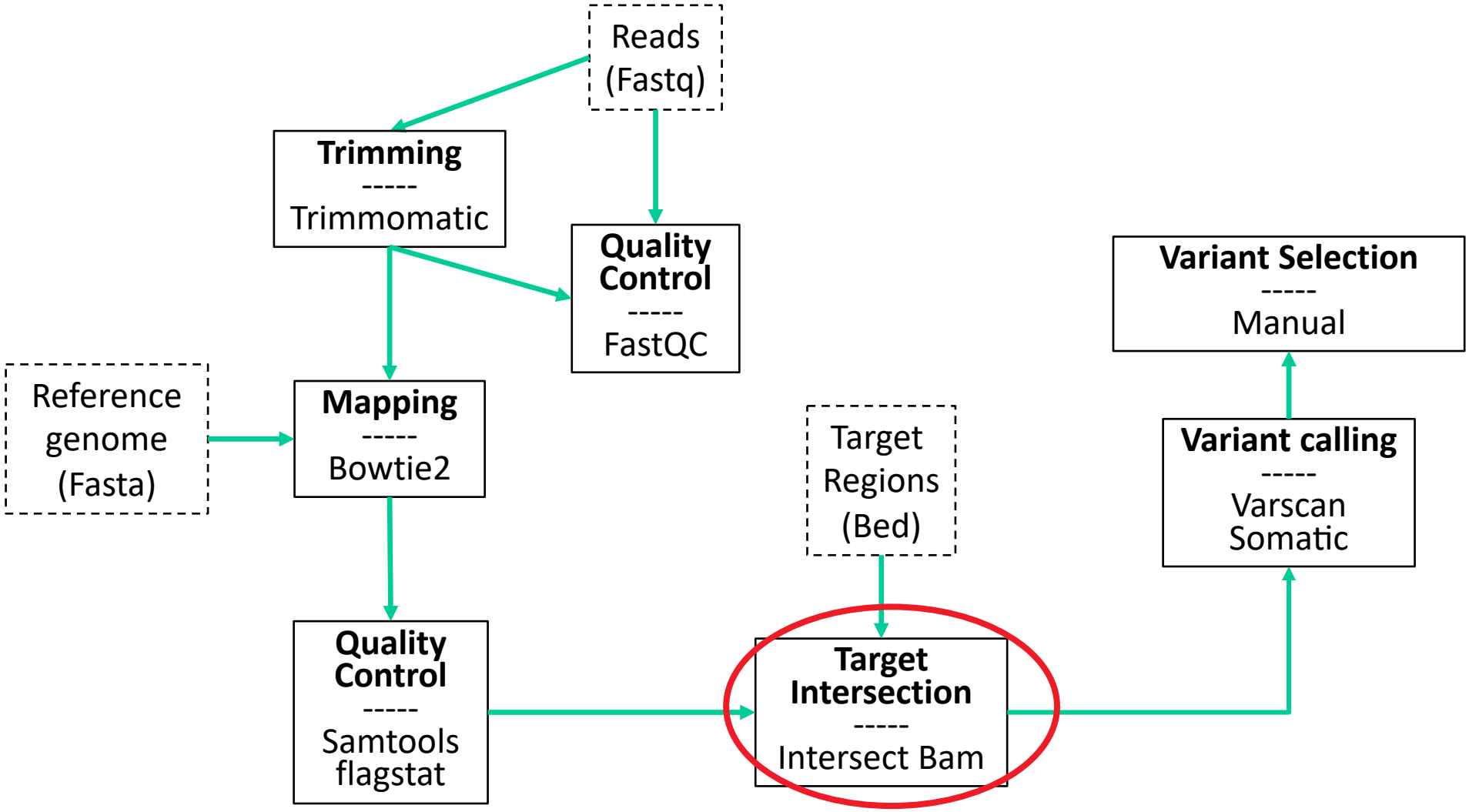
Samtools flatgstats

After renaming "Bowtie Normal"

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
Reference
genome
(Fasta)
Mapping

Bowtie2
Target
Regions
(Bed)
Variant Selection

Manual
Variant calling

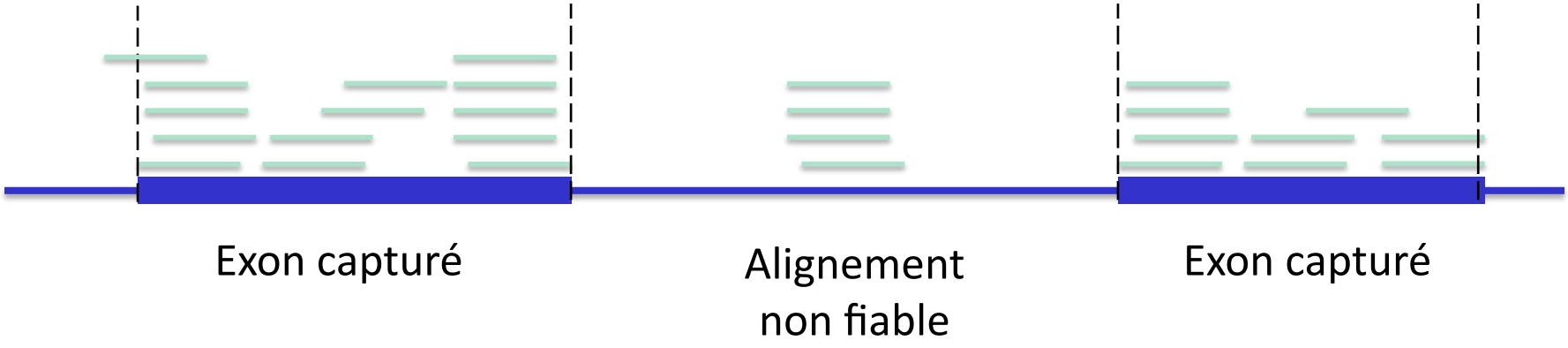
Varscan
Somatic
Quality
Control

Samtools
flagstat
Target
Intersection

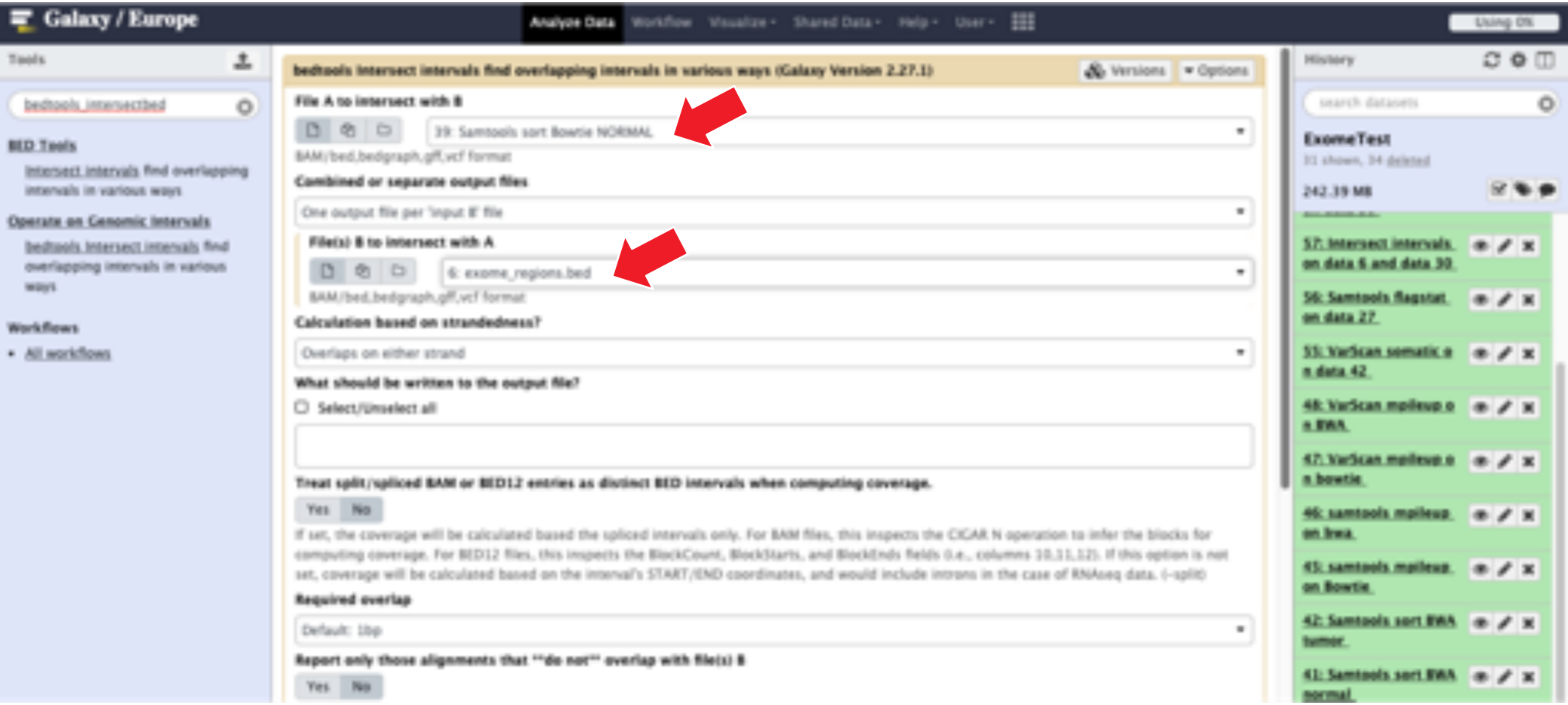
Intersect Bam

Target intersection

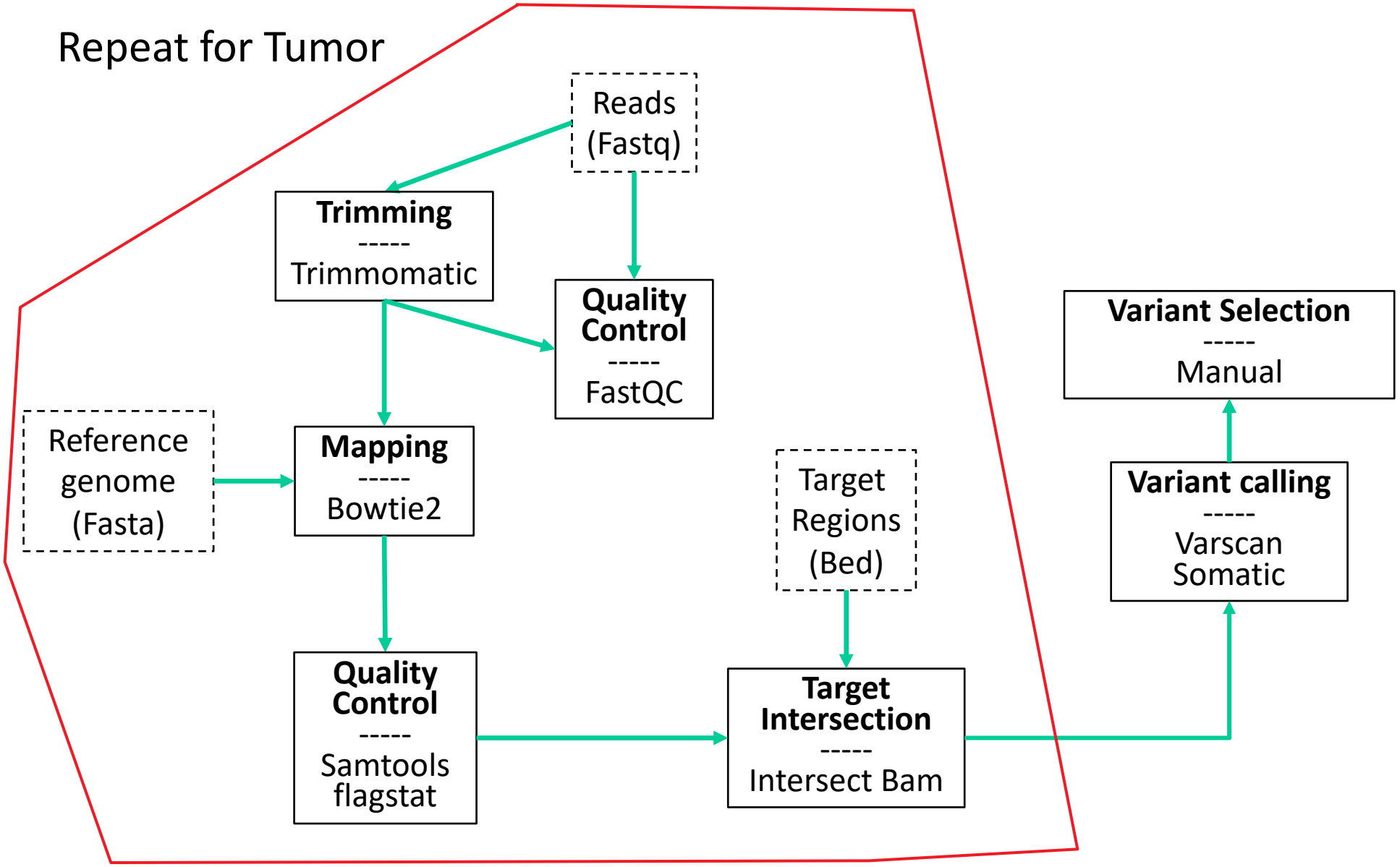
- Comparer l'alignement obtenu à la liste des positions visées par le protocole de capture



Bedtools intersect intervals



Vérifiez la réduction de taille du fichier BAM!


Repeat for Tumor
Reads
(Fastq)
Trimming

Trimmomatic
Quality
Control

FastQC
Reference
genome
(Fasta)
Mapping

Bowtie2
Target
Regions
(Bed)
Variant Selection

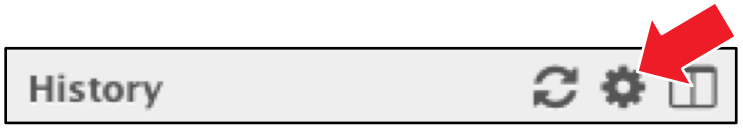
Manual
Variant calling

Varscan
Somatic
Quality
Control

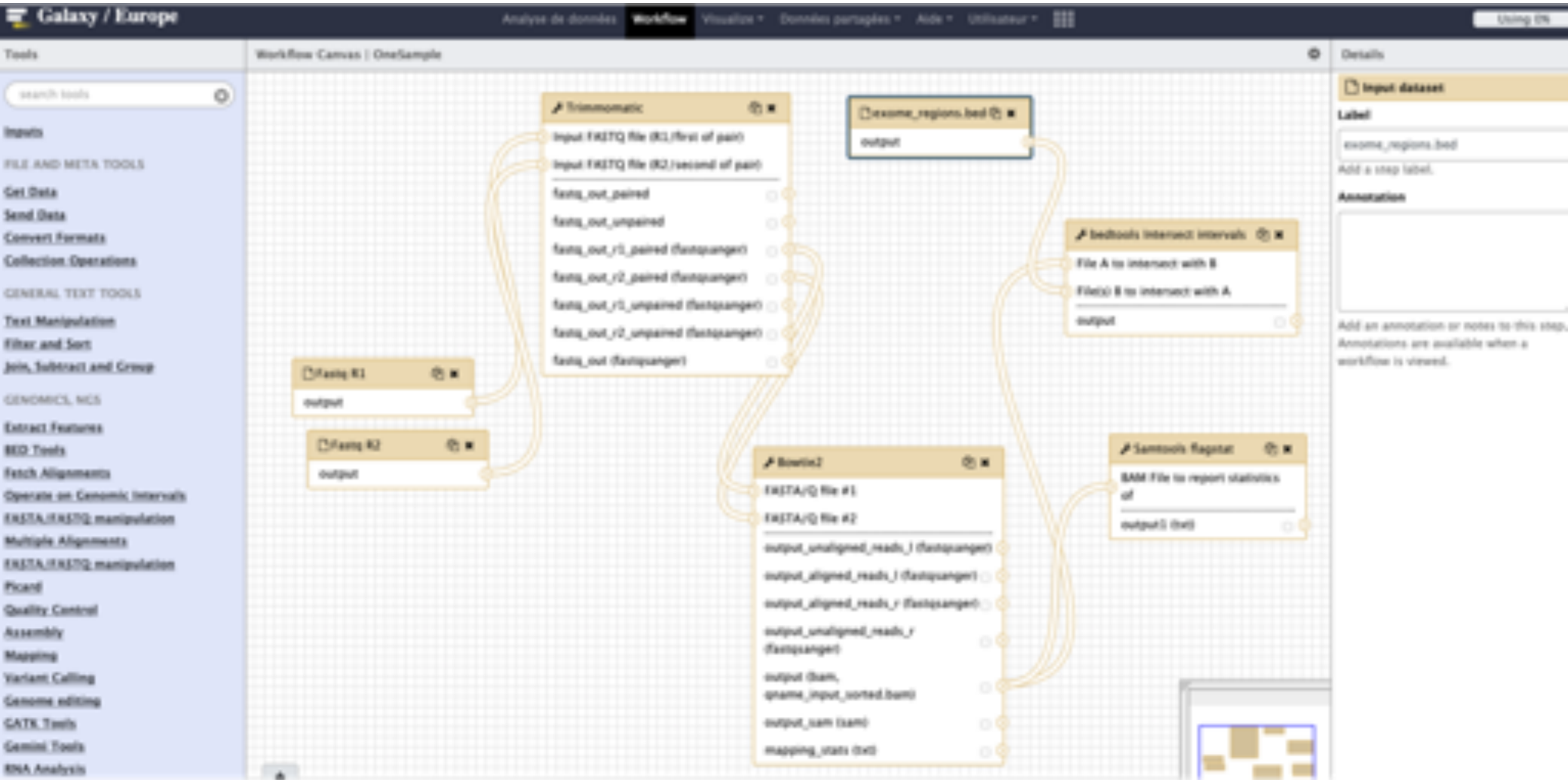
Samtools
flagstat
Target
Intersection

Intersect Bam

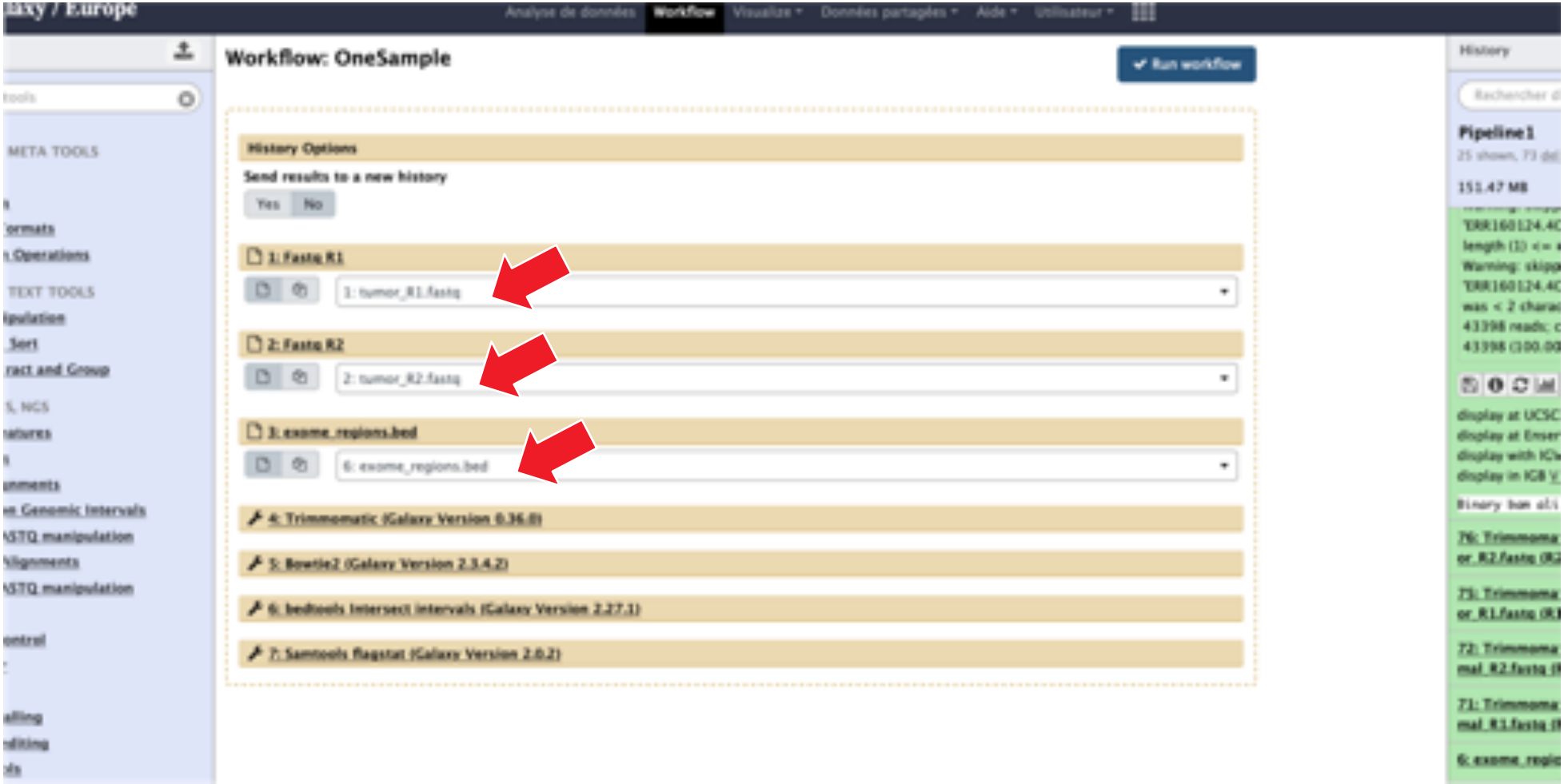
Extraire un workflow

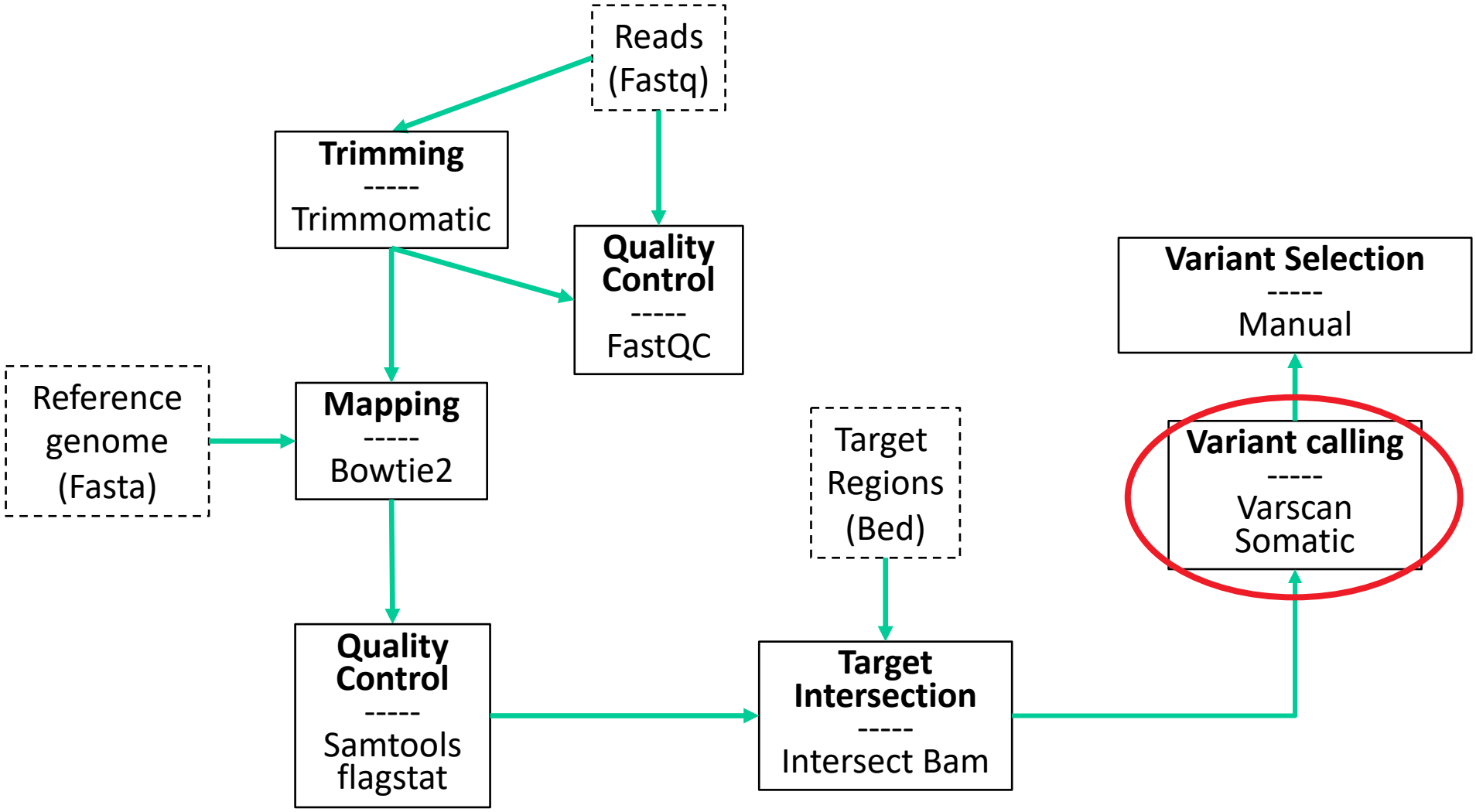


- Extraire workflow
- Le nommer
- Editer le workflow
- Choisir les données pertinentes (juste 2 fastq et regions.bed)
- Choisir les étapes de Trimmomatic à Intersect bed
- Enlever les data inutilisées (fastq tumor)
- Renommer les objets de façon générique (« sample » plutot que « normal »)
- Puis ⚙ save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


Galaxy / Europe
Analyse de données
Workflow
Visualize
Données partagées
Aide
Utilisateur
Using 0%
Tools
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
Workflow Canvas | OneSample
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)
Details
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
Reference
genome
(Fasta)
Mapping

Bowtie2
Target
Regions
(Bed)
Variant Selection

Manual
Variant calling

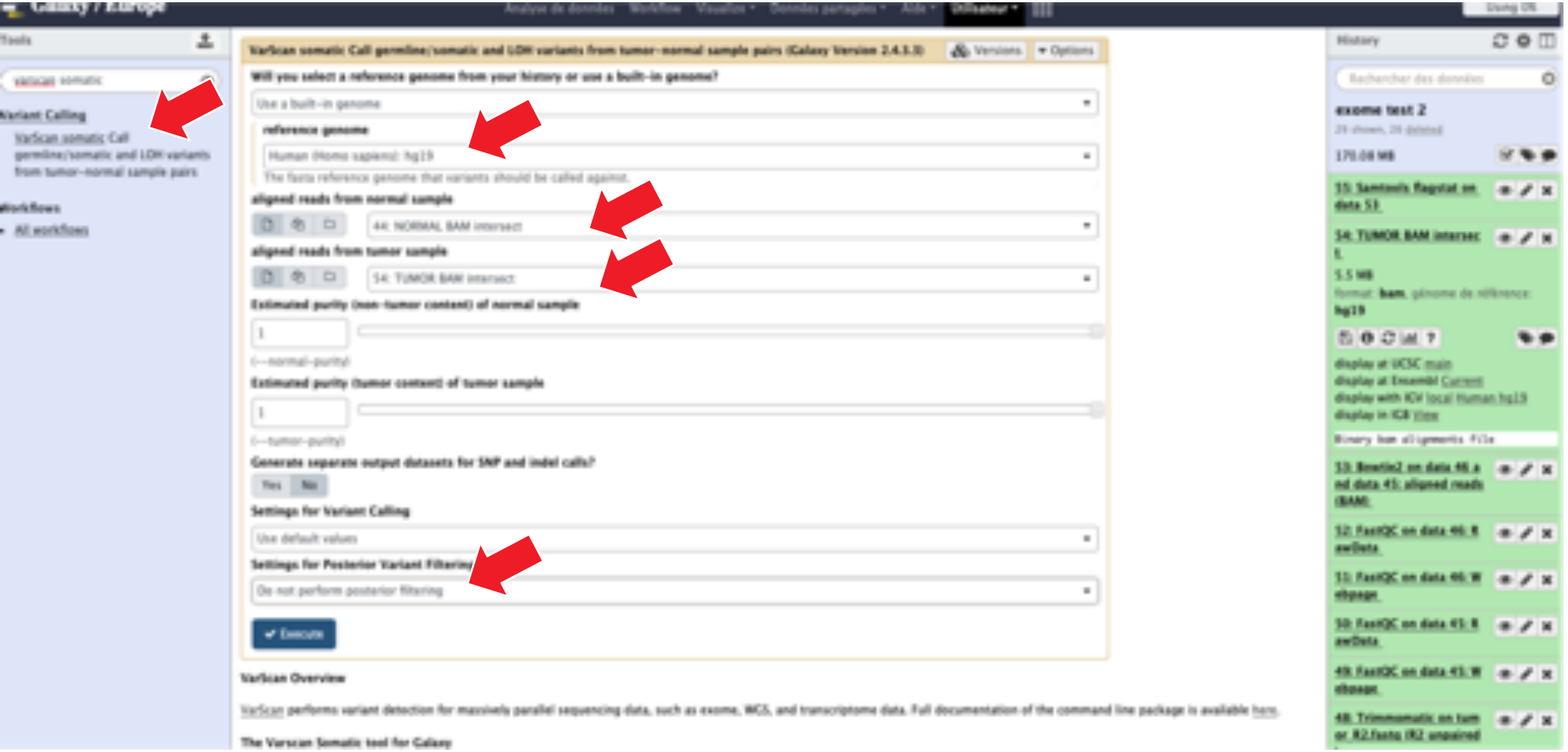
Varscan
Somatic
Quality
Control

Samtools
flagstat
Target
Intersection

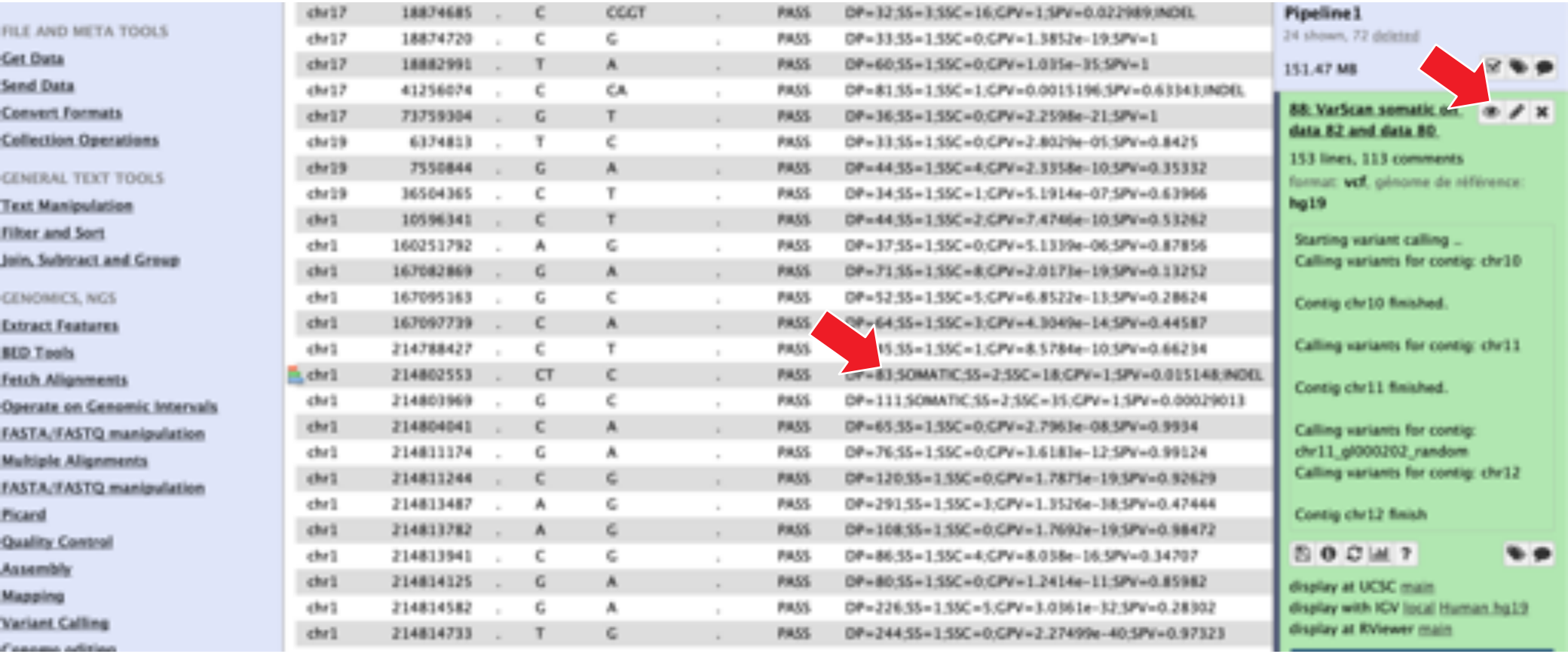
Intersect Bam

Somatic variant calling: Varscan

Attention: étape de 30min!

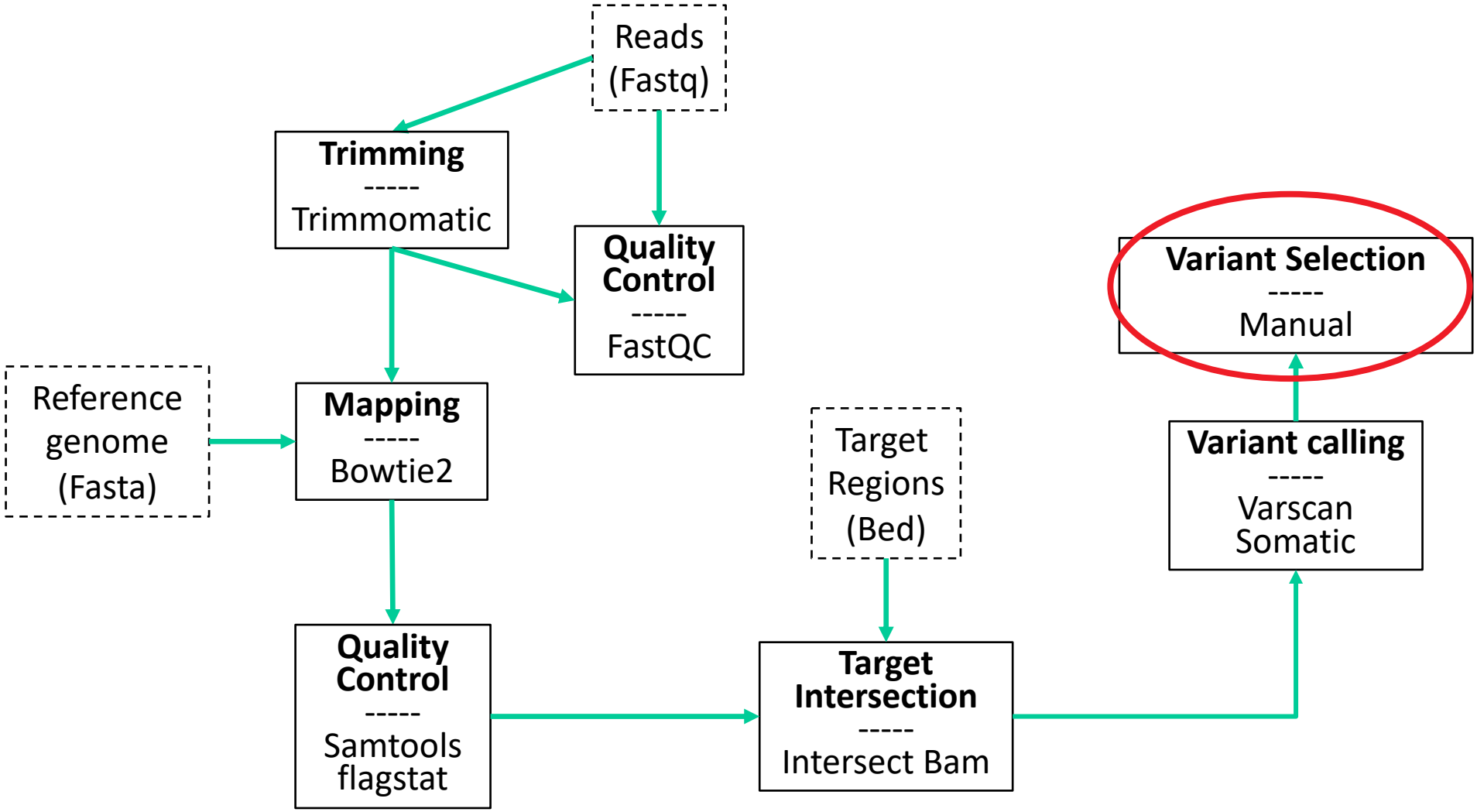


Check Varscan output



Somatic variants found?

Check "NORMAL" and "TUMOR" sample stats


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

Filter and visualize somatic variants

- Run the *grep* filter on the Varscan output with regular expression « somatic ». Check the result
- Launch IGV with hg19 reference
- Then 2 possibilities:
  - Download Normal and Tumor BAM files on your local computer (select option « download bam_index ») and load these files in IGV (« load from file »)
  - In Galaxy, click on « display with IGV local ». (will automatically connect with our local IGV session)
- Visualize somatic events.

IGV view

Bonus

Variant annotation with VEP

- Download the Varscan VCF file
- Go to https://www.ensembl.org/Tools/VEP
- Select GRCh37.p13 (=hg19)
- Launch VEP
- Display column "impact" and sort results by impact

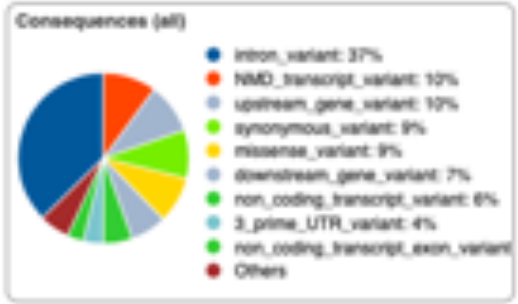
Note: the highest impact variants are not necessarily somatic!

Variant Effect Predictor results

Job details

Summary statistics

Category	Count
Variants processed	153
Variants filtered out	0
Novel / existing variants	6 (3.9) / 147 (96.1)
Overlapped genes	55
Overlapped transcripts	318
Overlapped regulatory features	23



Results preview

Navigation (per variant)

Show: 1 5 10 50 All variants

Filters

Uploaded variant | is | defined | Add

Download

All: VCF VEP TXT

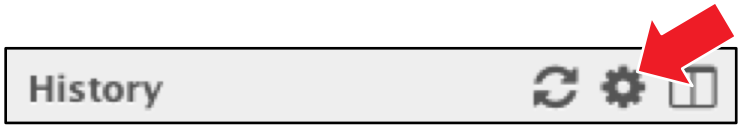
BioMart: Variants Genes

Show/hide columns (2 hidden)

Uploaded variant	Location	Allele	Consequence	Impact	Symbol	Gene	Feature type	Feature	Biotype	Exon	Intron	cDNA posit
.	1:248059779-248059779	A	frameshift_variant	HIGH	OR2W3	ENSG00000238243	Transcript	ENST00000360358	protein_coding	1/1	-	891-8
.	1:248059779-248059779	A	frameshift_variant	HIGH	OR2W3	ENSG00000238243	Transcript	ENST00000537741	protein_coding	3/3	-	1148-
.	3:121416308-121416308	T	stop_gained	HIGH	GOLGB1	ENSG00000173230	Transcript	ENST00000340645	protein_coding	13/22	-	3173
.	3:121416308-121416308	T	stop_gained	HIGH	GOLGB1	ENSG00000173230	Transcript	ENST00000393667	protein_coding	13/22	-	3173
.	3:121416308-121416308	T	stop_gained	HIGH	GOLGB1	ENSG00000173230	Transcript	ENST00000489400	protein_coding	9/9	-	2659

Annexes

Galaxy: partager ses données



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