Réalisation d'un Pipeline d'analyse d'exome

université PARIS-SACLAY | FACULTÉ DES SCIENCES D'ORSAY

GUSTAVE ROUSSY CANCER CAMPUS GRAND PARIS

V. 17/1/2023

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

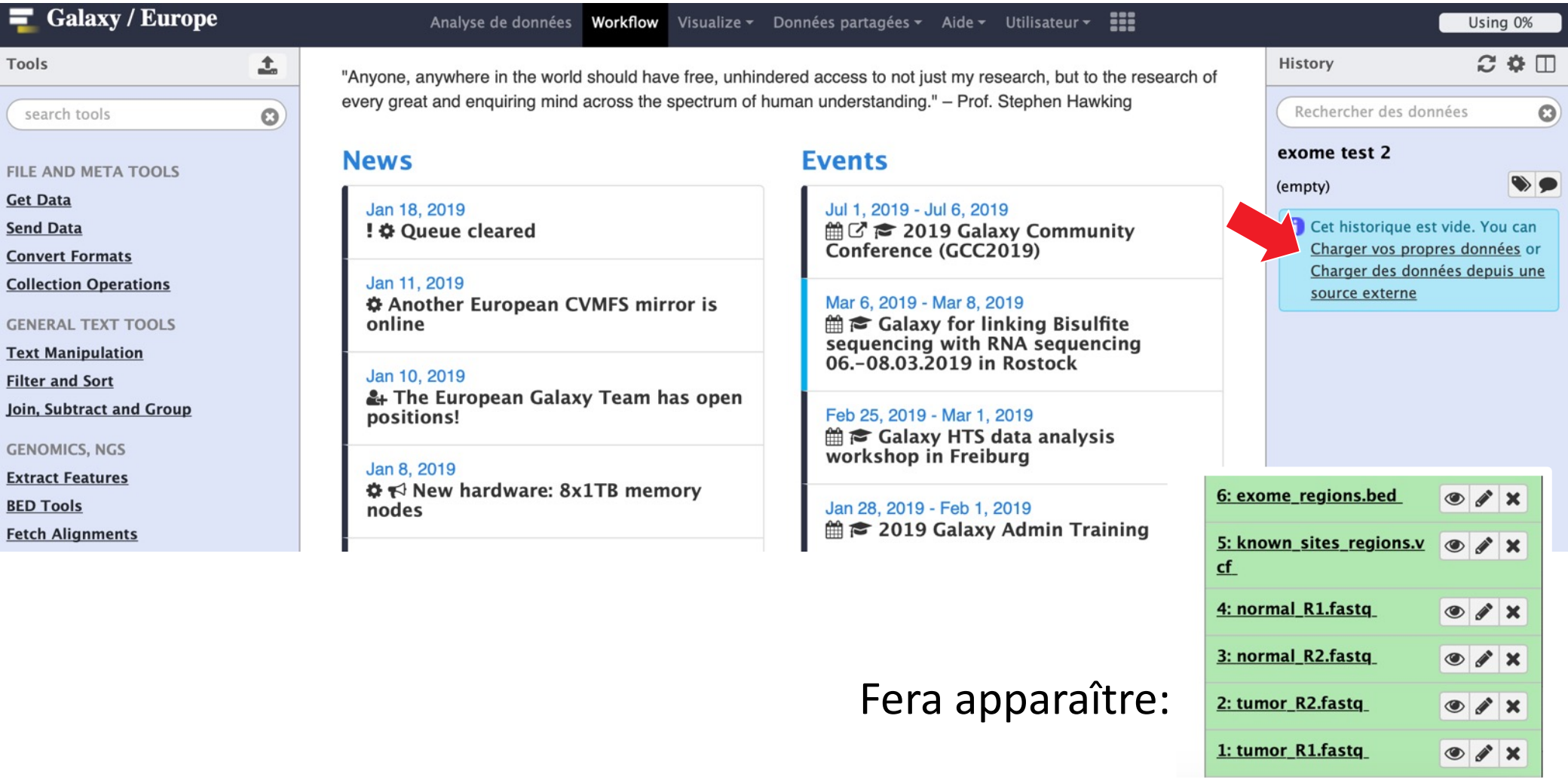
- Mean depth higher for the tumor sample (~100X) than for the normal sample (~30X) to detect somatic variant with a low allelic frequency
- Aligned Exome size: ~15 Go tumor ; ~7 Go blood
  Complete analysis processing Time: ~20h
- **Fastq files restricted to a few regions (~112kb) to limit processing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

Open access Galaxy servers

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

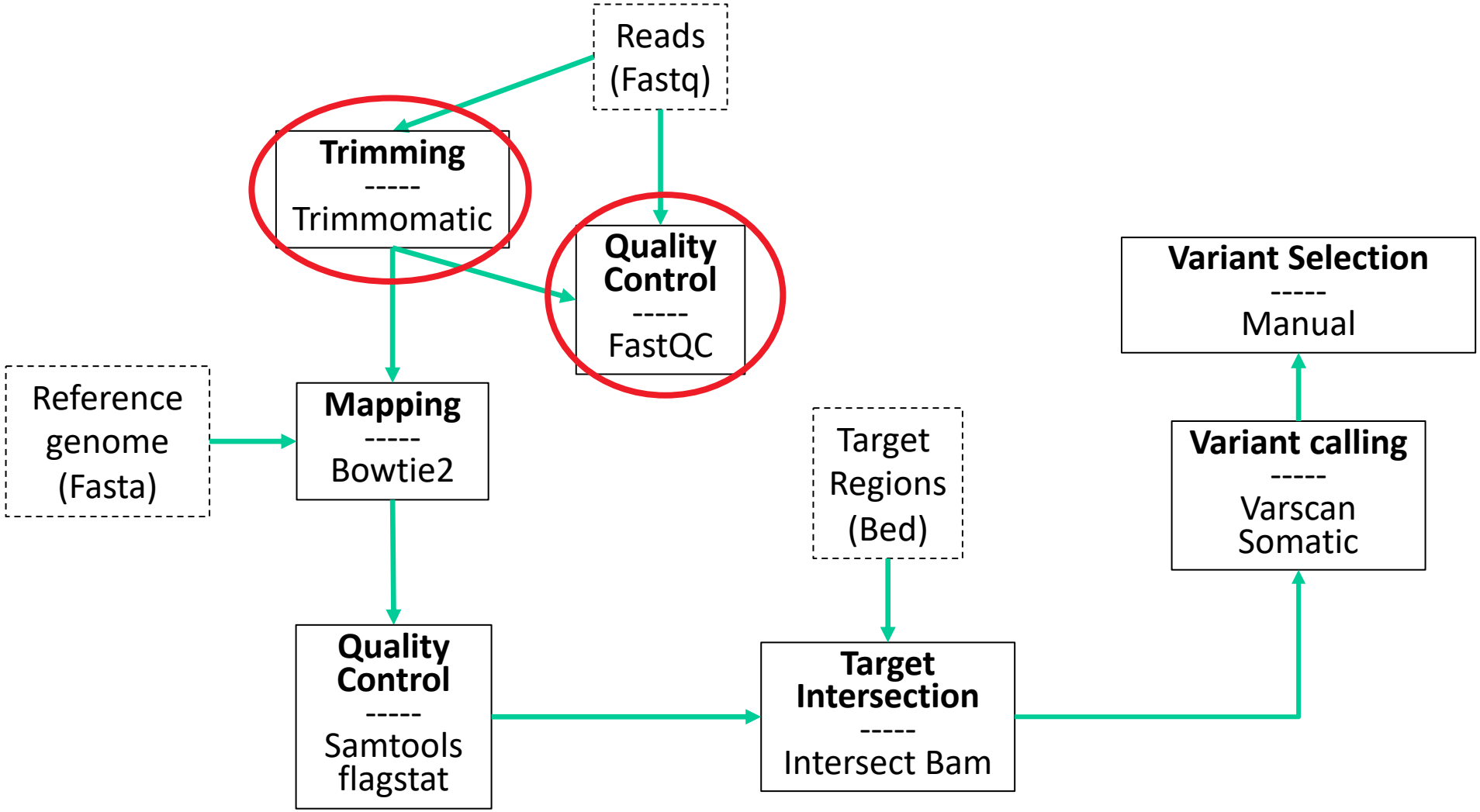
Chargez vos données

(on peut aussi déposer des fichiers .gz)



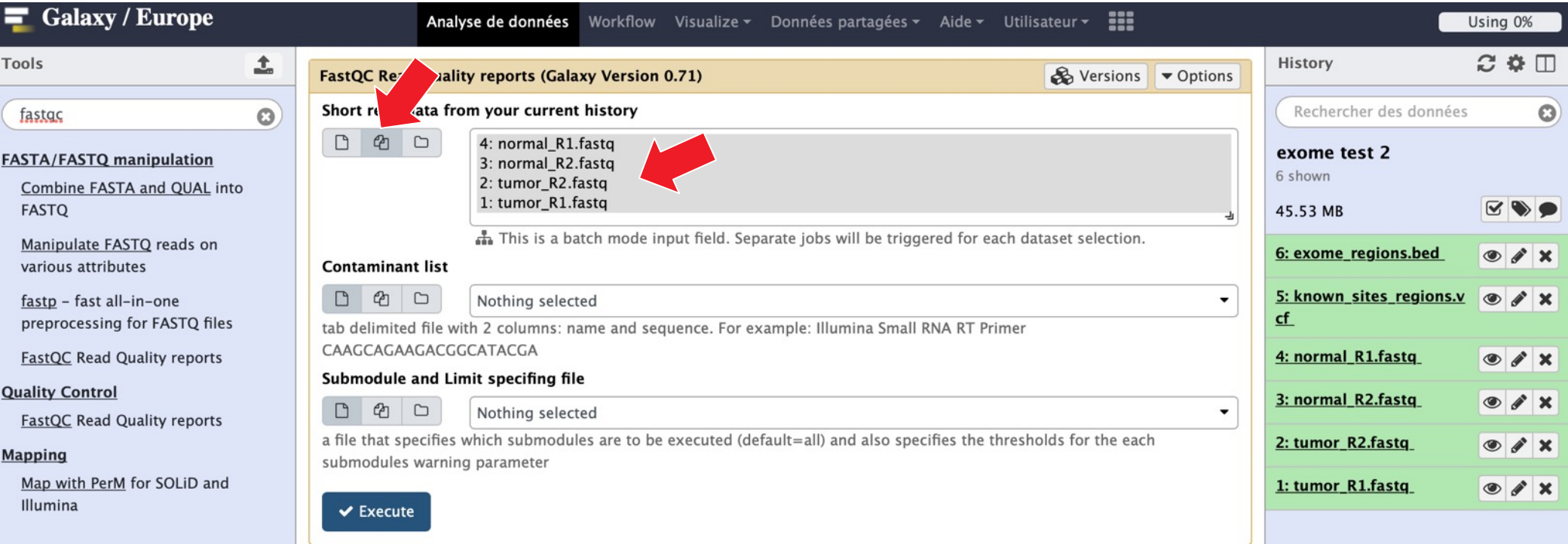
Fera apparaître:

A simplified Variant Pipeline

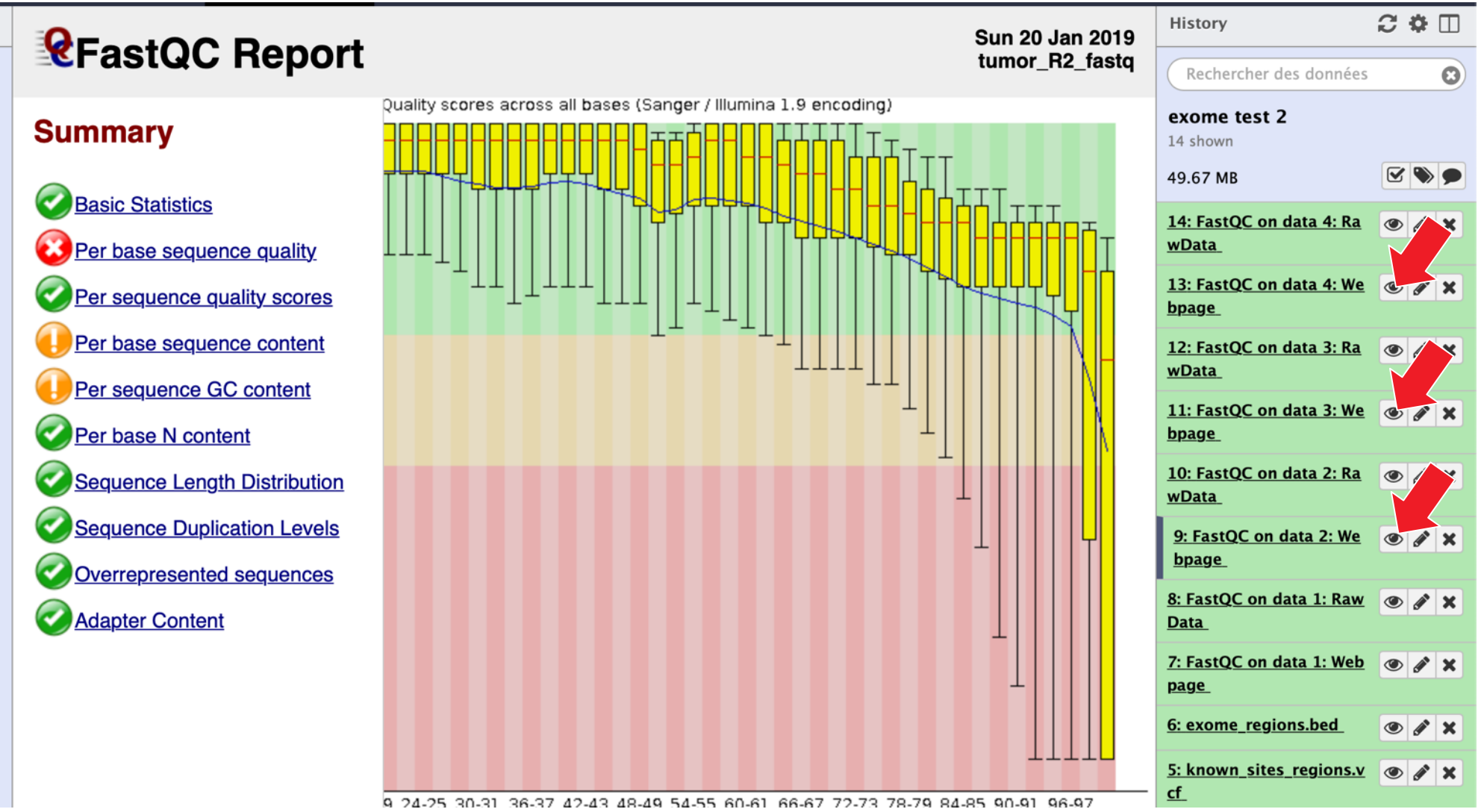


fastqc

Vérifiez les 4 fichiers fastq en mode multi-files



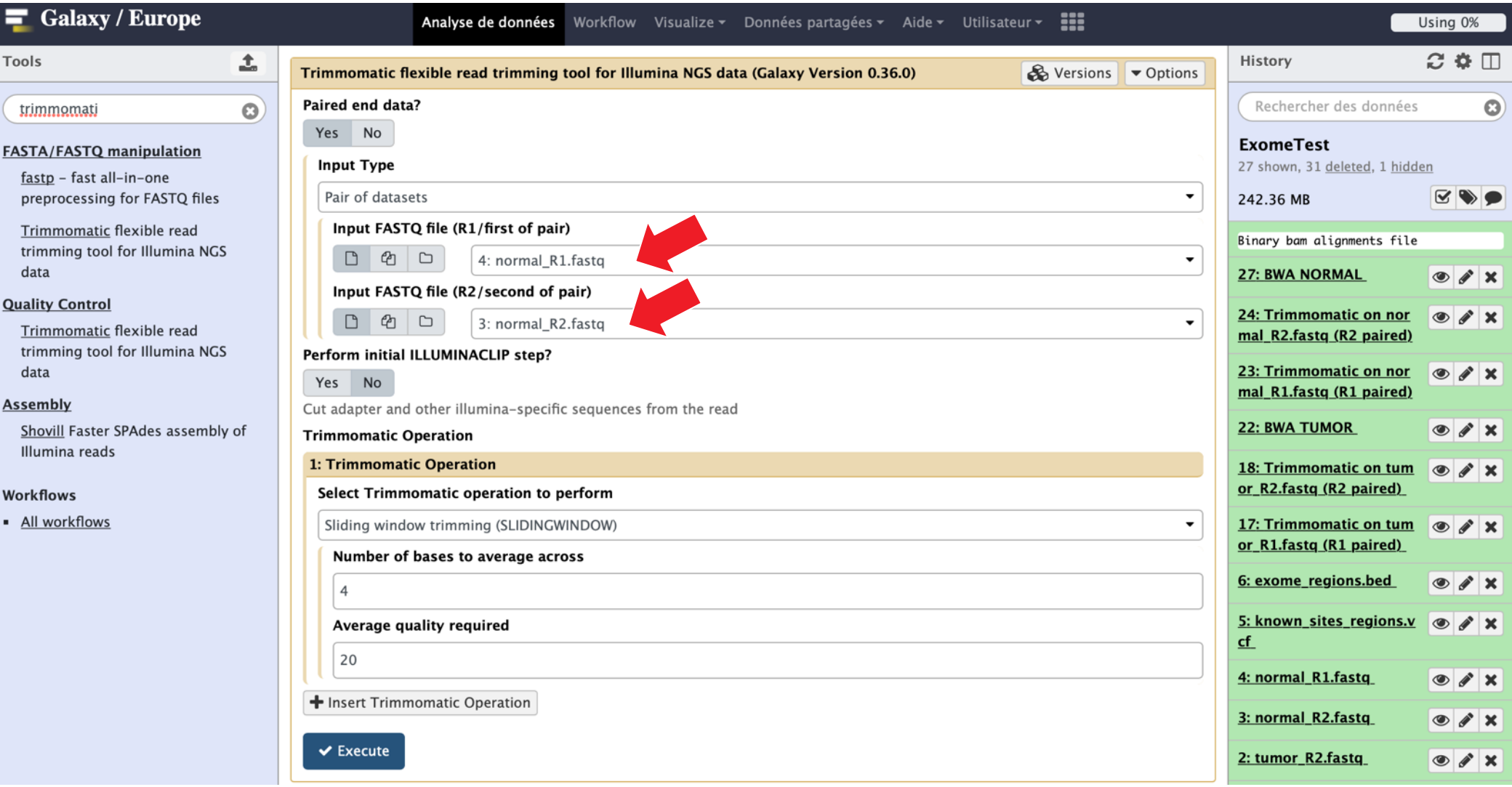
Fastqc results



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

A partir de cette étape on travaille avec une condition (normal ou tumeur), puis on sauvegardera l'ensemble du pipeline pour le rejouer sur l'autre échantillon

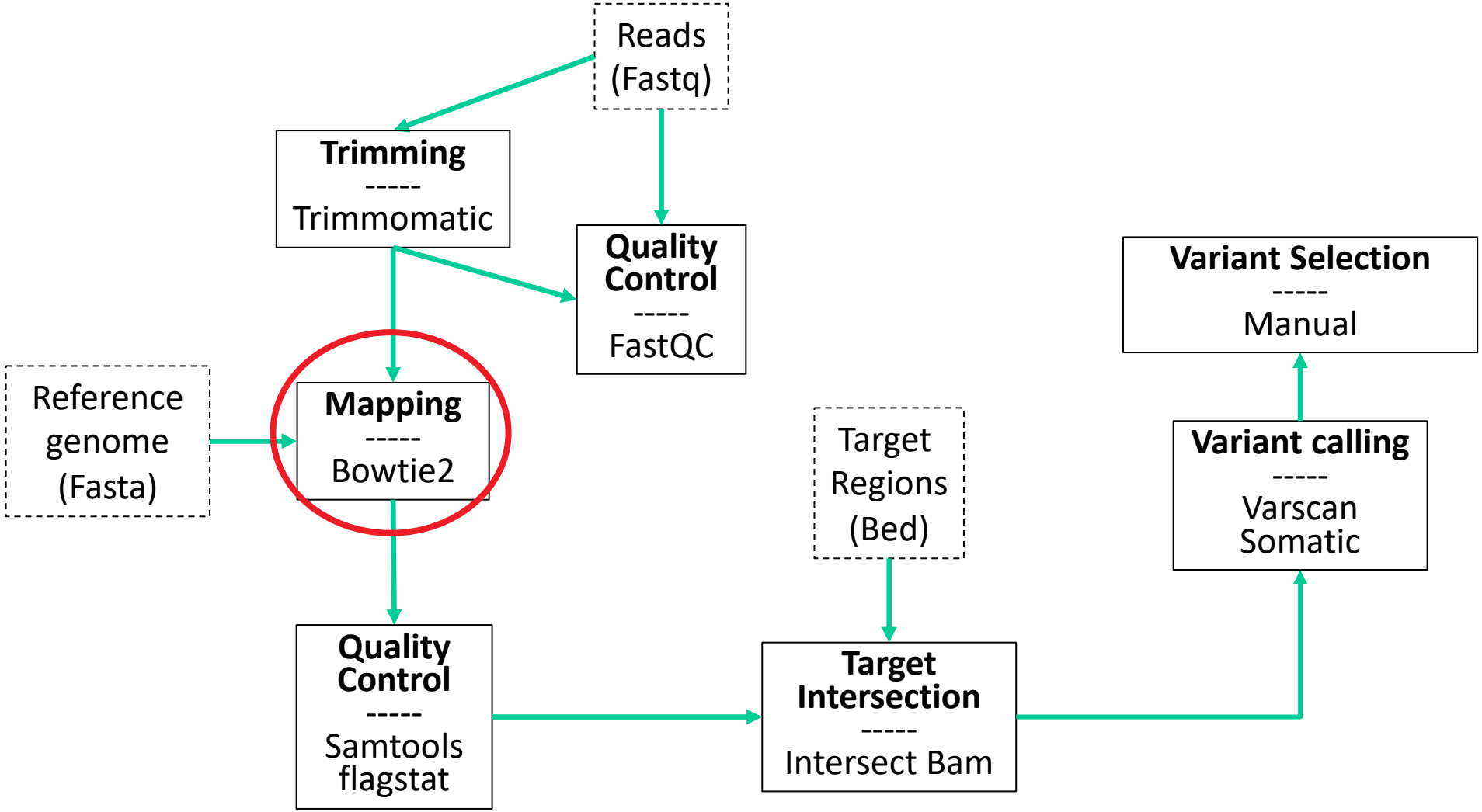
Trimmomatic



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

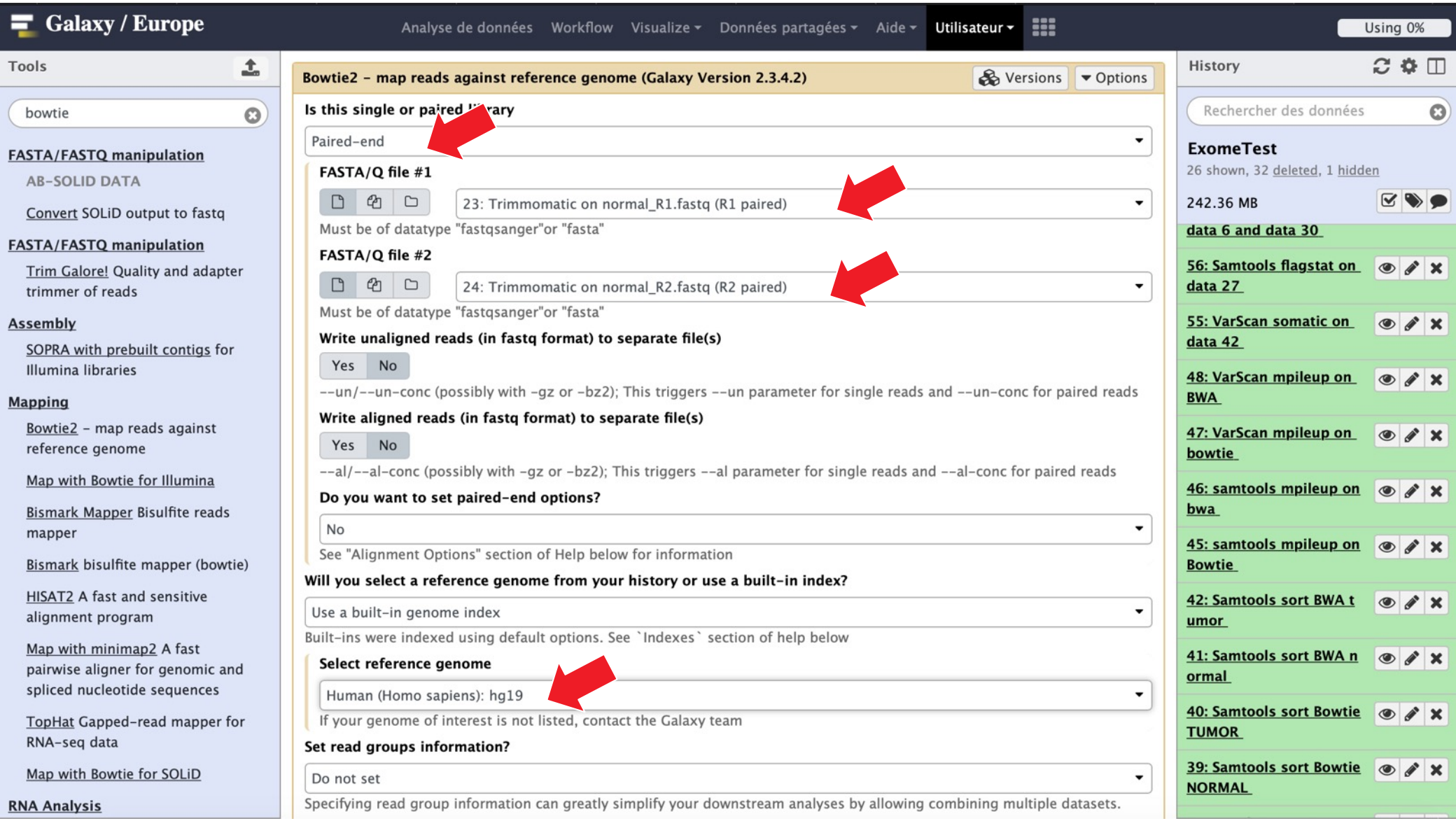
Manual
Variant calling

Varscan
Somatic
Quality
Control

Samtools
flagstat
Target
Intersection

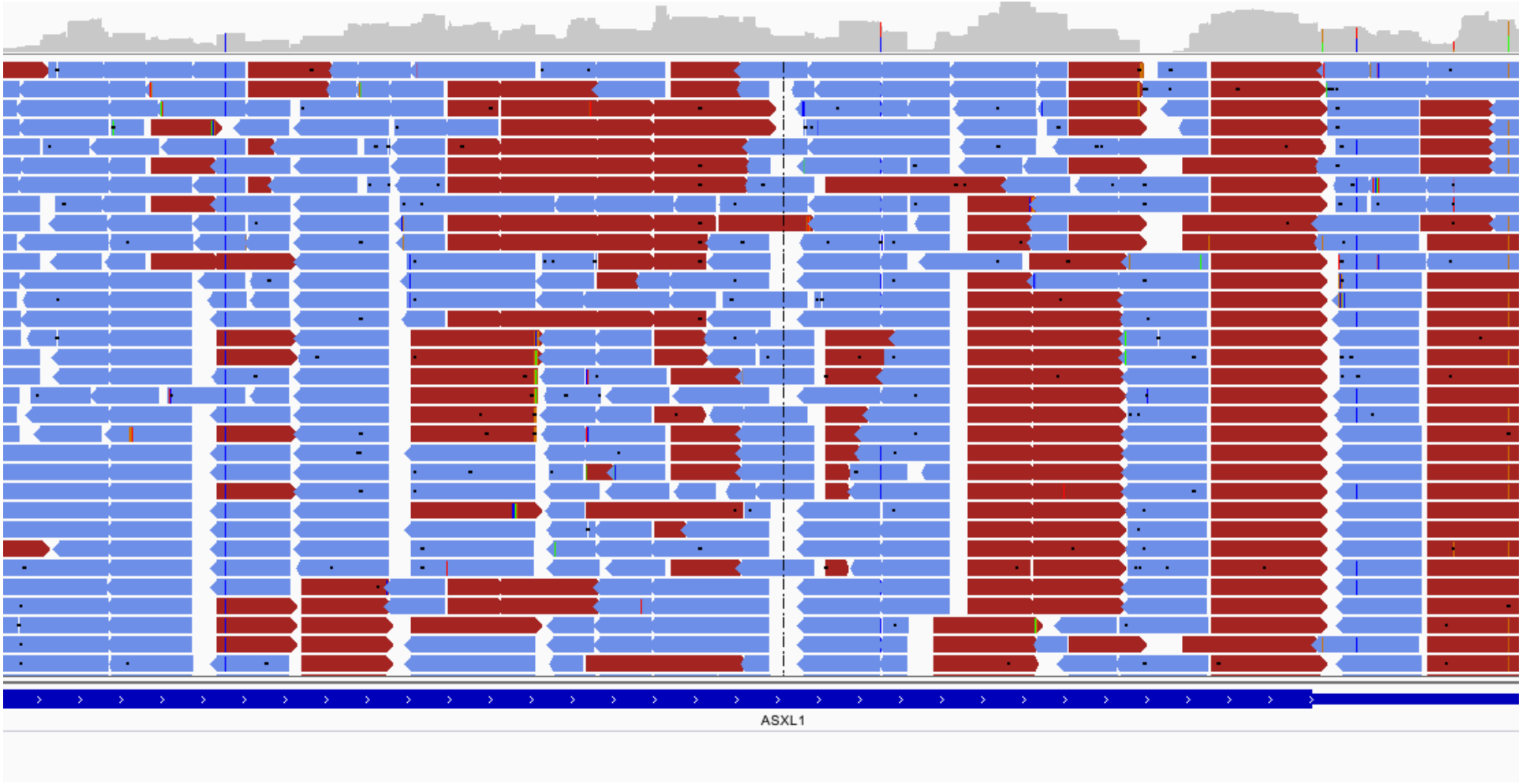
Intersect Bam

Bowtie



Check Bowtie result: what type of file is it?

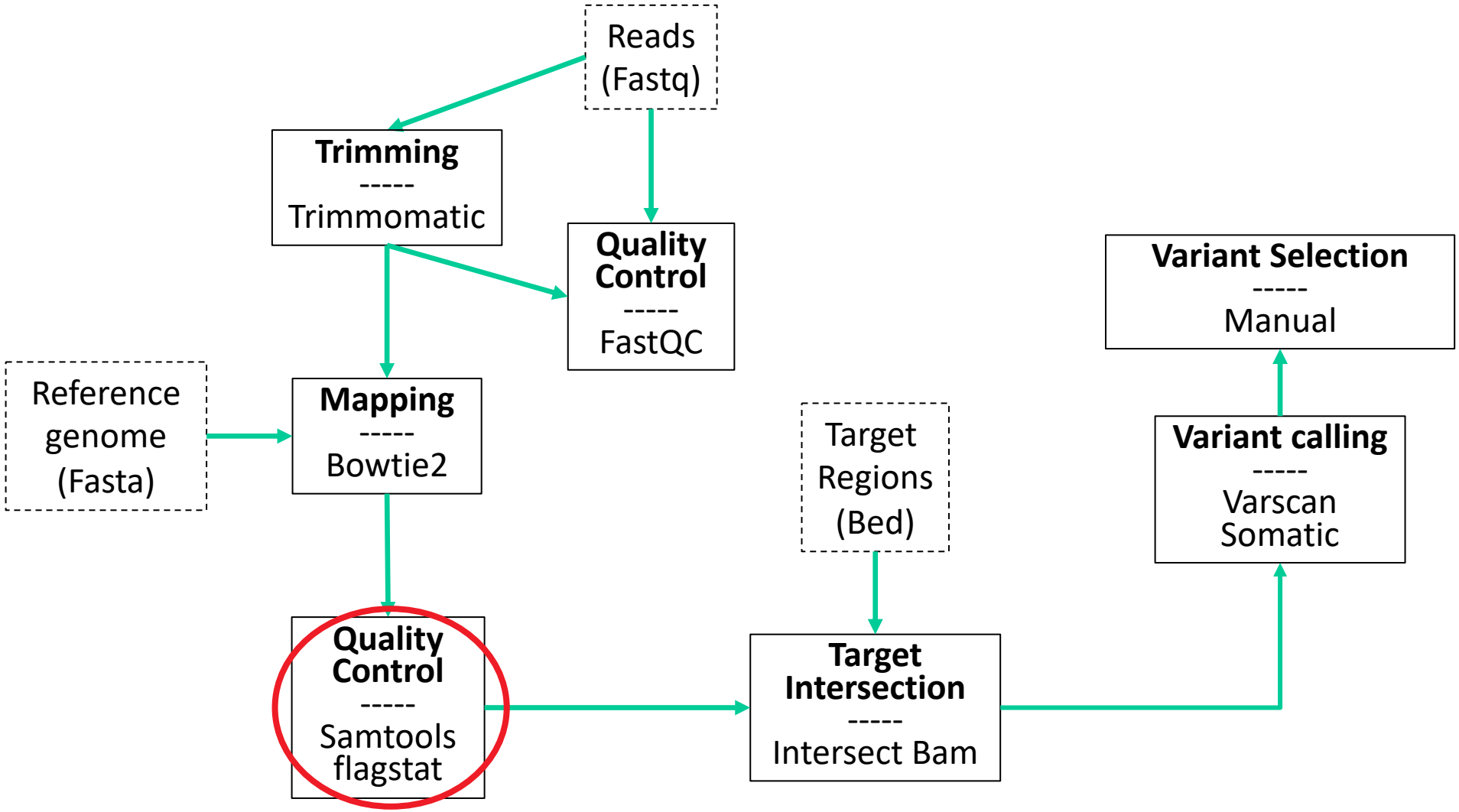
Reads alignés: le format BAM/SAM



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

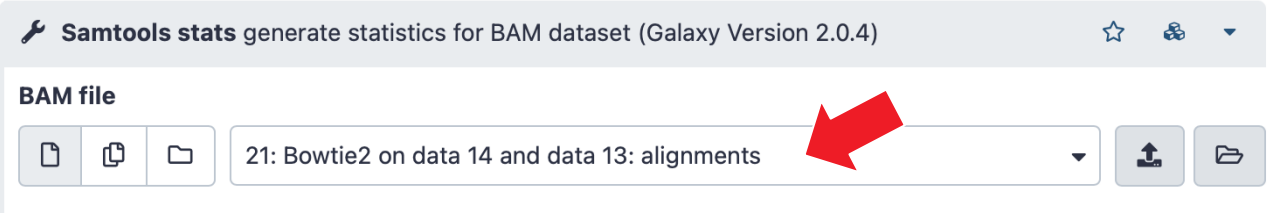
Samtools
flagstat
Target
Intersection

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)

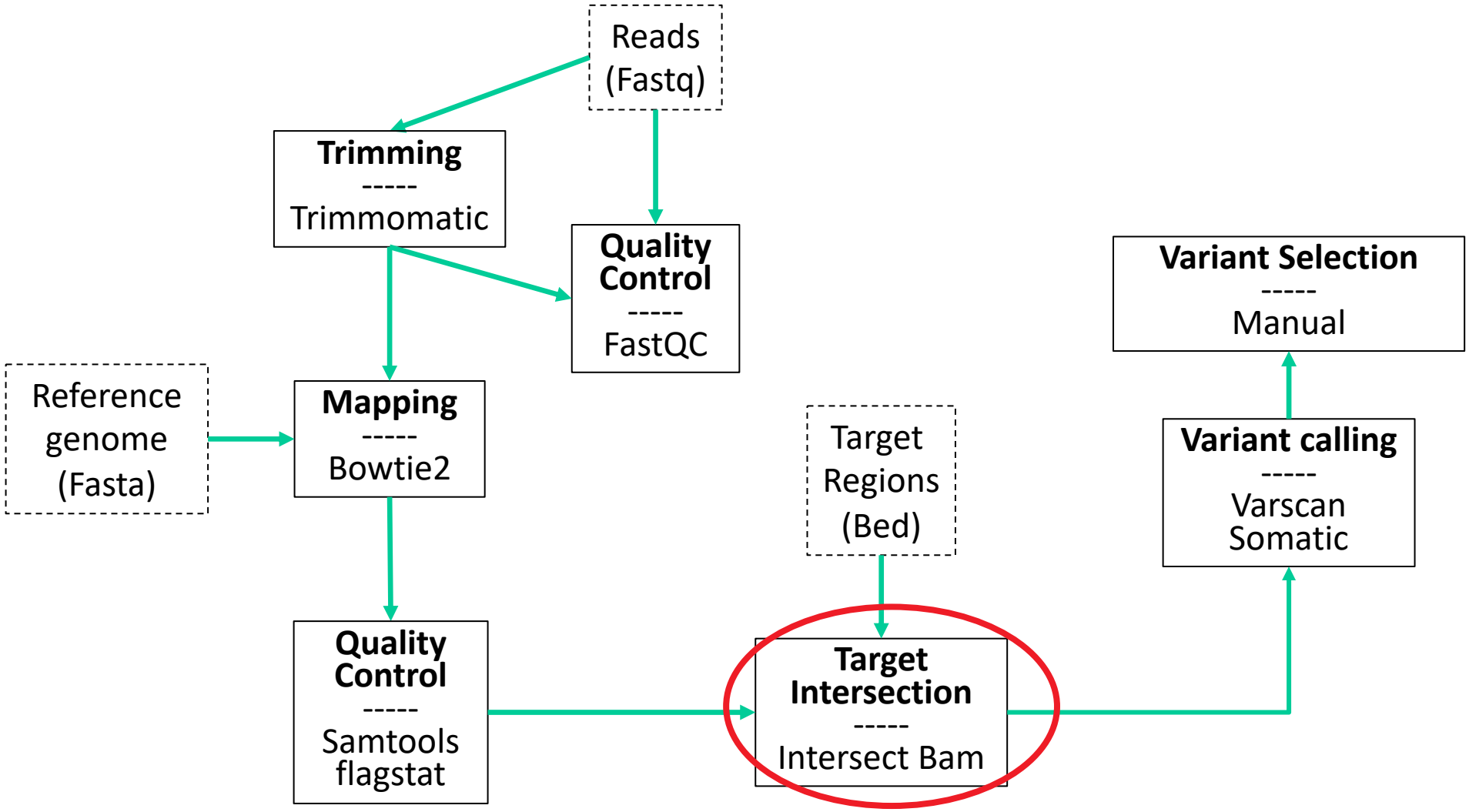
Samtools stats



résultat

# This file was produced by samtools stats (1.13+htslib-1.13) and can be plotted using plot-bamstats			
# This file contains statistics for all reads.			
# The command line was: stats -@ 0 infile			
# CHK, Checksum	[2]Read Names	[3]Sequences	[4]Qualities
# CHK, CRC32 of reads which passed filtering followed by addition (32bit overflow)			
CHK	9f0f8d14	22de9793	dcc71d6b
# Summary Numbers. Use `grep ^SN | cut -f 2-` to extract this part.			
SN	raw total sequences:	86796	# excluding supplementary and secor
SN	filtered sequences:	0	
SN	sequences:	86796	
SN	is sorted:	1	
SN	1st fragments:	43398	
SN	last fragments:	43398	
SN	reads mapped:	86738	
SN	reads mapped and paired:	86706	# paired-end technology bit set + bot
SN	reads unmapped:	58	
SN	reads properly paired:	85918	# proper-pair bit set
SN	reads paired:	86796	# paired-end technology bit set
SN	reads duplicated:	0	# PCR or optical duplicate bit set
SN	reads MQ0:	79	# mapped and MQ=0
SN	reads QC failed:	0	
SN	non-primary alignments:	0	
SN	supplementary alignments:	0	
SN	total length:	7290089	# ignores clipping
SN	total first fragment length:	3907724	# ignores clipping
SN	total last fragment length:	3382365	# ignores clipping
SN	bases mapped:	7286384	# ignores clipping
SN	bases mapped (cigar):	7286384	# more accurate
SN	bases trimmed:	0	
SN	bases duplicated:	0	
SN	mismatches:	16593	# from NM fields

Suggestion: renommer vos fichiers BAM etc avec des noms plus simples


Reads
(Fastq)
Trimming

Trimmomatic
Quality
Control

FastQC
Reference
genome
(Fasta)
Mapping

Bowtie2
Target
Regions
(Bed)
Variant Selection

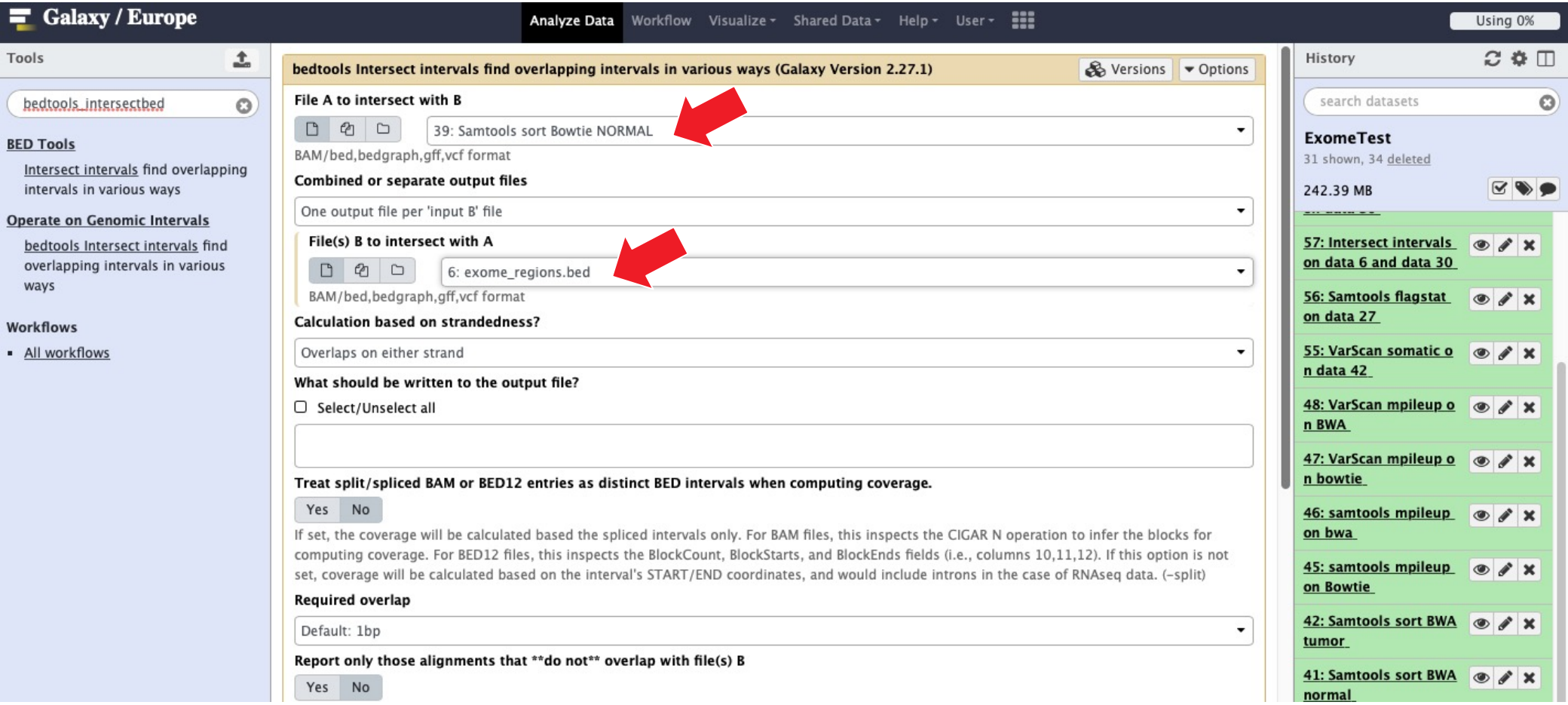
Manual
Variant calling

Varscan
Somatic
Quality
Control

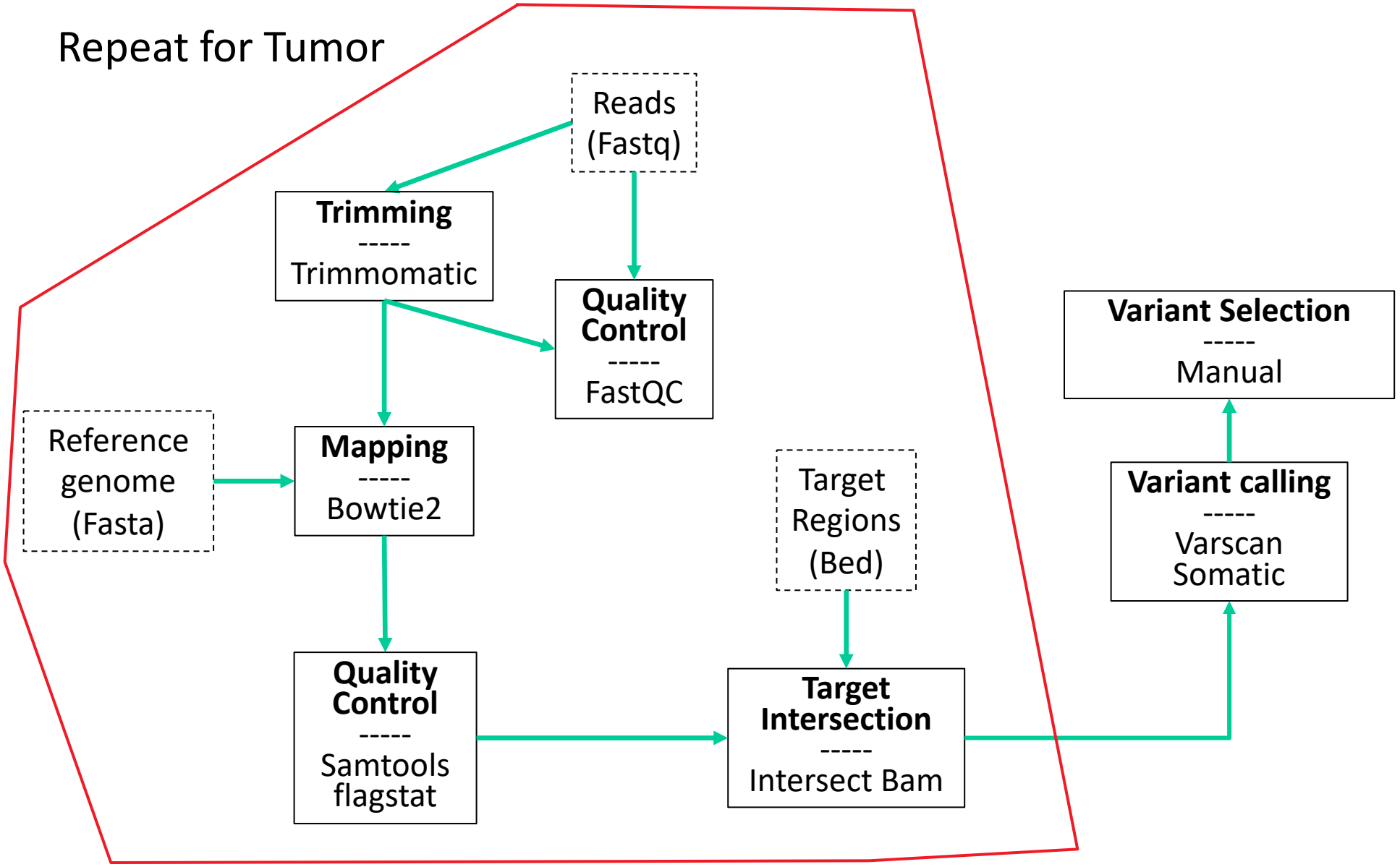
Samtools
flagstat
Target
Intersection

Intersect Bam

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

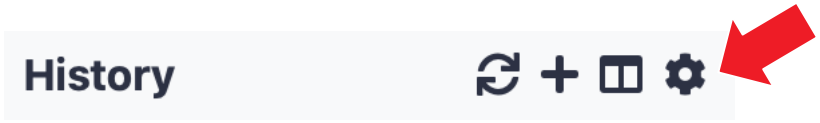
Manual
Variant calling

Varscan
Somatic
Quality
Control

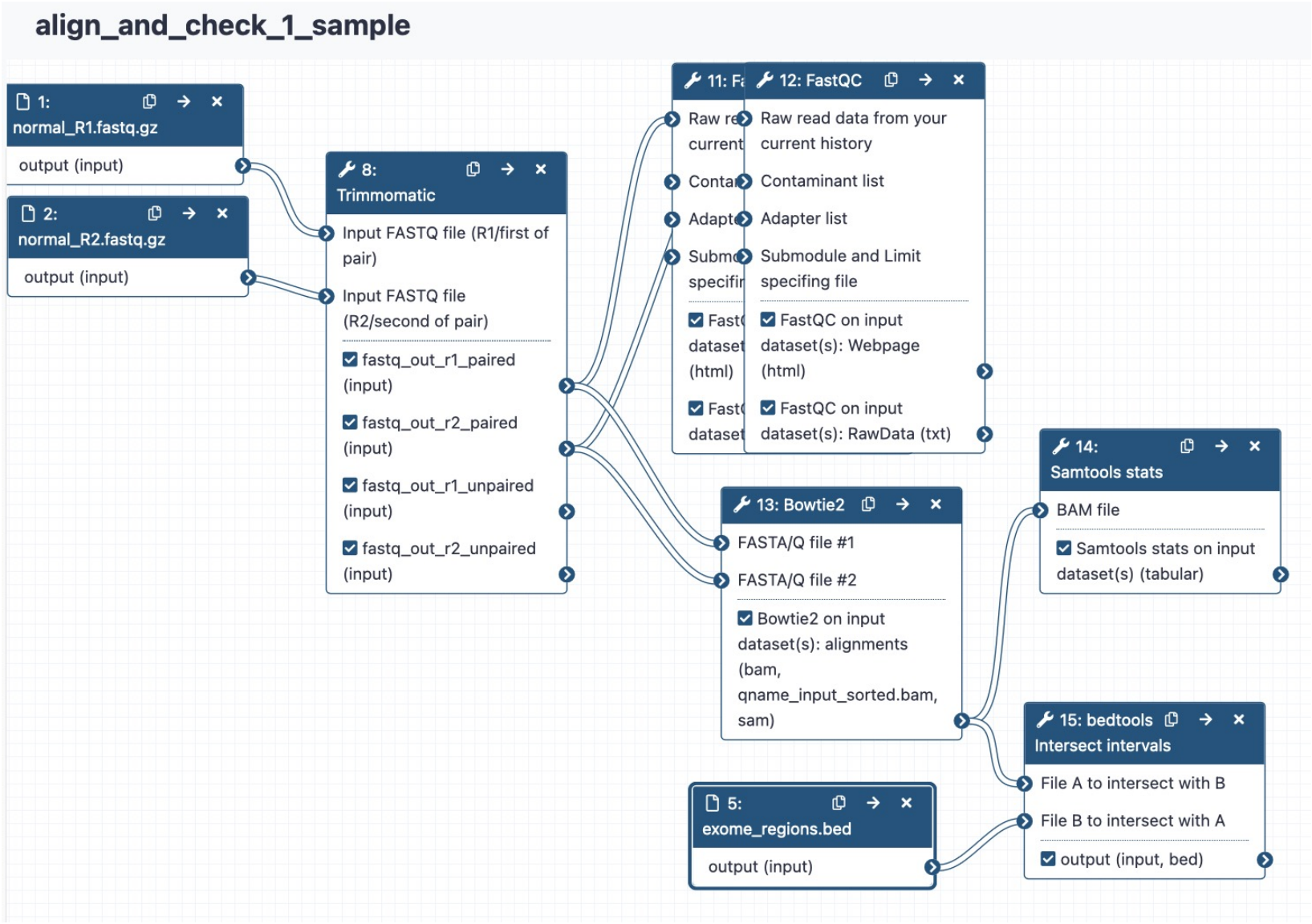
Samtools
flagstat
Target
Intersection

Intersect Bam

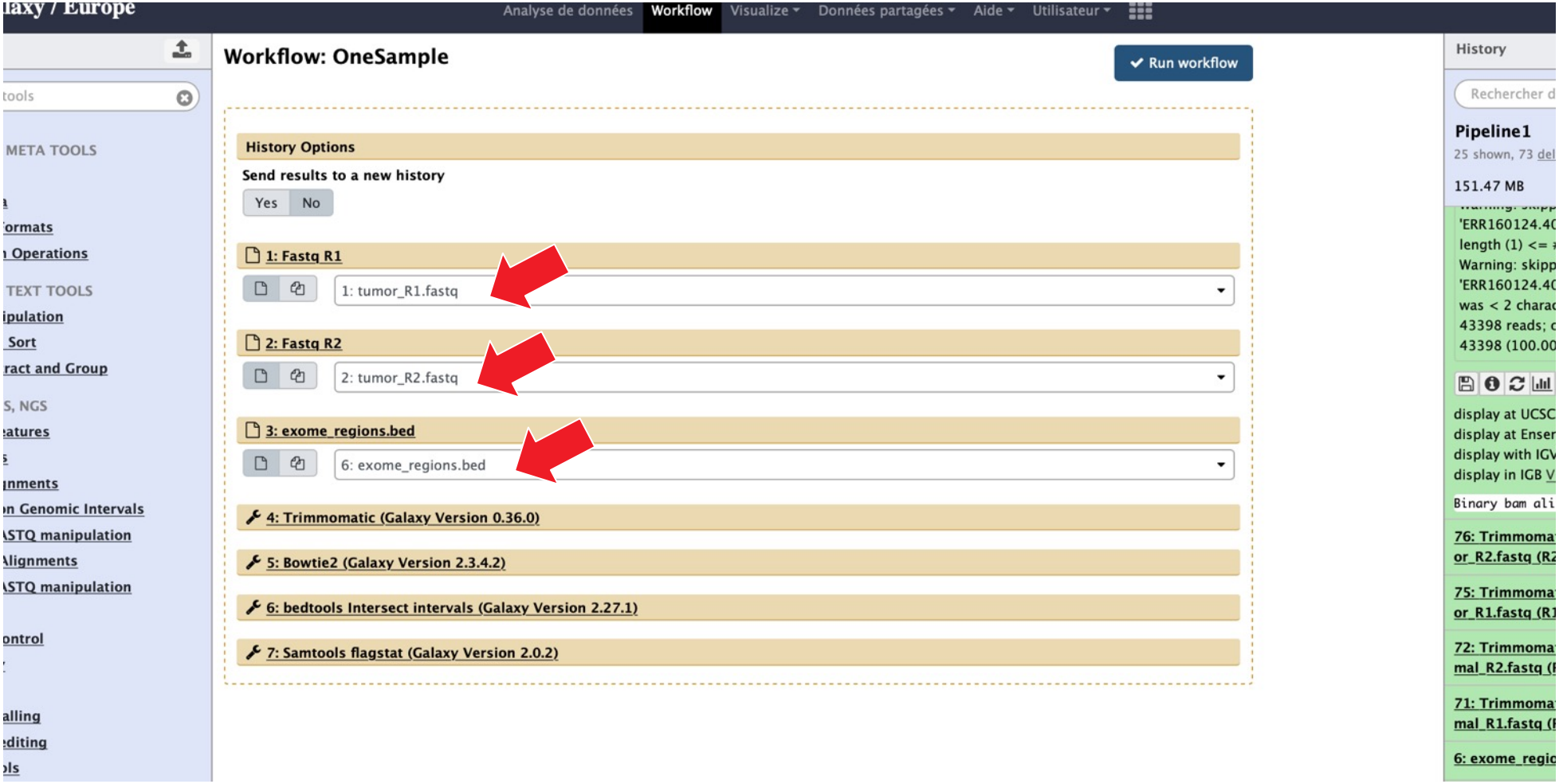
Extraire un workflow

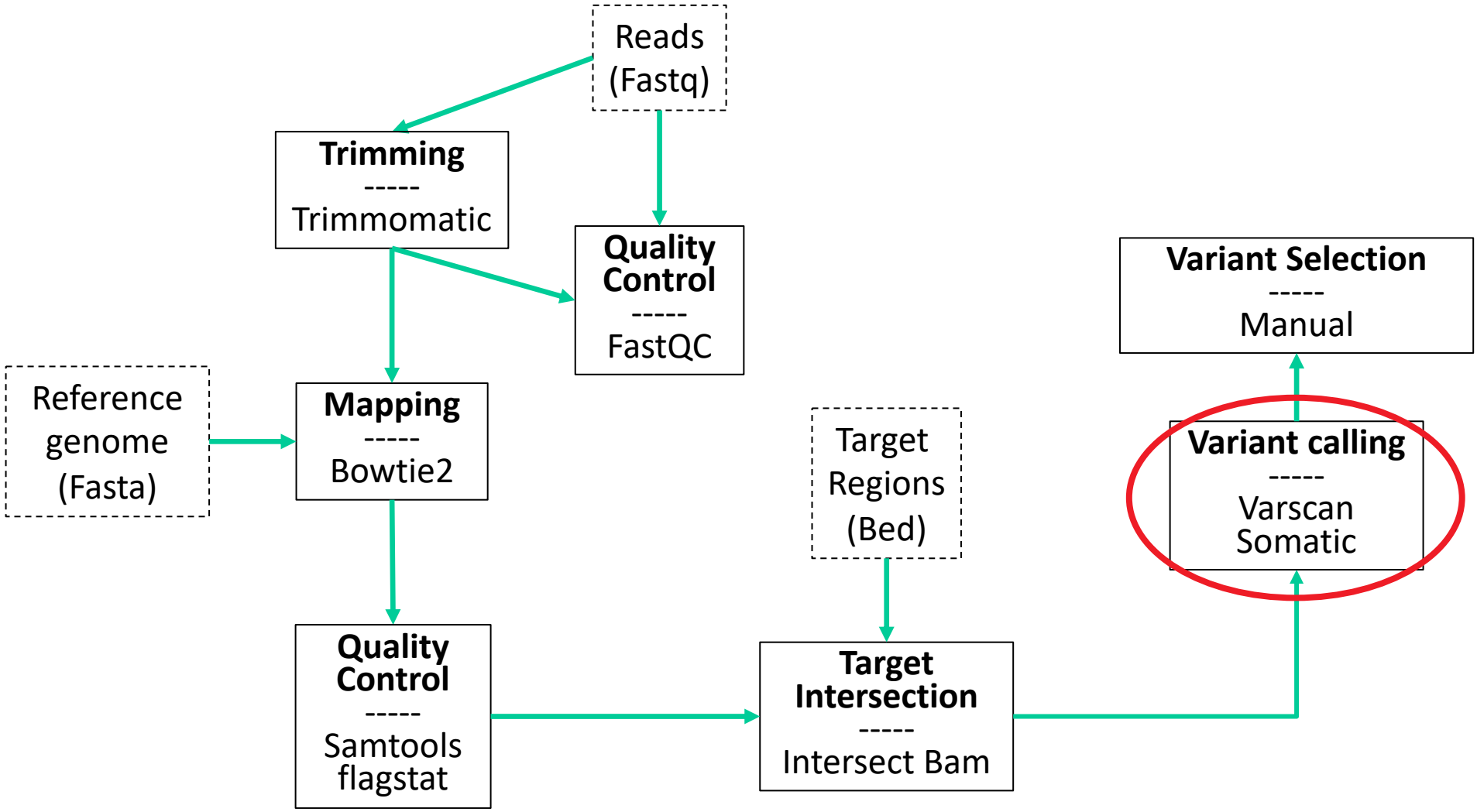


- Extraire workflow
- Le nommer + créer
- Editer le workflow
- Choisir les données pertinentes (juste 2 fastq et regions.bed)
- Choisir les étapes de Trimmomatic à Intersect bed
- Enlever les data inutilisées (fastq tumor)
- Renommer les objets de façon générique (« sample » plutot que « normal »)
-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


align_and_check_1_sample
1: normal_R1.fastq.gz
output (input)
2: normal_R2.fastq.gz
output (input)
8: Trimmomatic
Input FASTQ file (R1/first of pair)
Input FASTQ file (R2/second of pair)
fastq_out_r1_paired (input)
fastq_out_r2_paired (input)
fastq_out_r1_unpaired (input)
fastq_out_r2_unpaired (input)
11: Fa
Raw re
current
Conta
Adapt
Subm
specifi
Fast
dataset
(html)
Fast
dataset
12: FastQC
Raw read data from your current history
Contaminant list
Adapter list
Submodule and Limit specifing file
FastQC on input dataset(s): Webpage (html)
FastQC on input dataset(s): RawData (txt)
13: Bowtie2
FASTA/Q file #1
FASTA/Q file #2
Bowtie2 on input dataset(s): alignments (bam, qname_input_sorted.bam, sam)
14: Samtools stats
BAM file
Samtools stats on input dataset(s) (tabular)
15: bedtools Intersect intervals
File A to intersect with B
File B to intersect with A
output (input, bed)
5: exome_regions.bed
output (input)

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

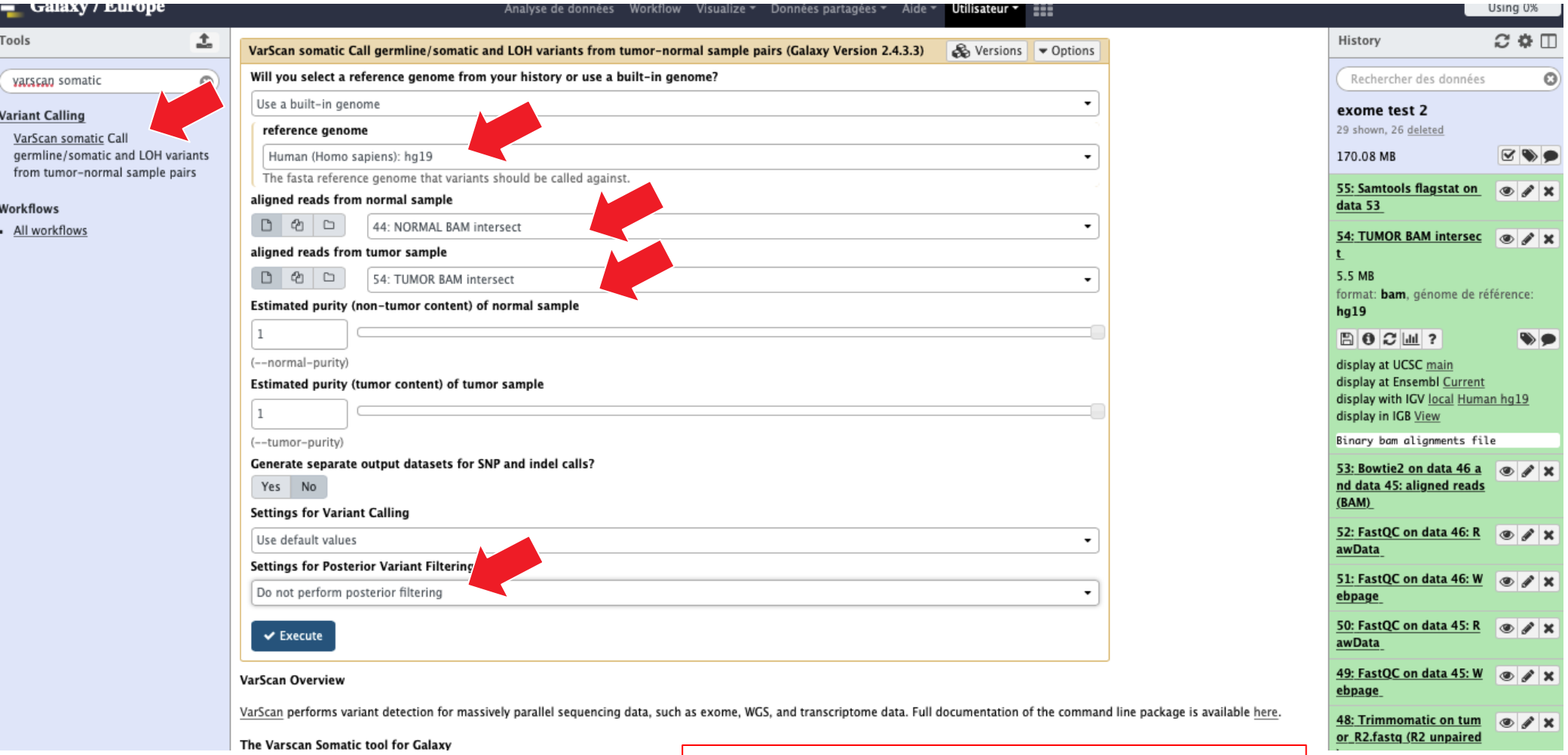
Varscan
Somatic
Quality
Control

Samtools
flagstat
Target
Intersection

Intersect Bam

Somatic variant calling: Varscan

Attention: étape de 10-30min!



Sur Galaxy.fr: version de Varscan différente exige un fichier mpileup en entrée

Vérifiez la sortie de Varscan

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

chr17	18874685	.	C	CGGT	.	PASS	DP=32;SS=3;SSC=16;GPV=1;SPV=0.022989;INDEL
chr17	18874720	.	C	G	.	PASS	DP=33;SS=1;SSC=0;GPV=1.3852e-19;SPV=1
chr17	18882991	.	T	A	.	PASS	DP=60;SS=1;SSC=0;GPV=1.035e-35;SPV=1
chr17	41256074	.	C	CA	.	PASS	DP=81;SS=1;SSC=1;GPV=0.0015196;SPV=0.63343;INDEL
chr17	73759304	.	G	T	.	PASS	DP=36;SS=1;SSC=0;GPV=2.2598e-21;SPV=1
chr19	6374813	.	T	C	.	PASS	DP=33;SS=1;SSC=0;GPV=2.8029e-05;SPV=0.8425
chr19	7550844	.	G	A	.	PASS	DP=44;SS=1;SSC=4;GPV=2.3358e-10;SPV=0.35332
chr19	36504365	.	C	T	.	PASS	DP=34;SS=1;SSC=1;GPV=5.1914e-07;SPV=0.63966
chr1	10596341	.	C	T	.	PASS	DP=44;SS=1;SSC=2;GPV=7.4746e-10;SPV=0.53262
chr1	160251792	.	A	G	.	PASS	DP=37;SS=1;SSC=0;GPV=5.1339e-06;SPV=0.87856
chr1	167082869	.	G	A	.	PASS	DP=71;SS=1;SSC=8;GPV=2.0173e-19;SPV=0.13252
chr1	167095163	.	G	C	.	PASS	DP=52;SS=1;SSC=5;GPV=6.8522e-13;SPV=0.28624
chr1	167097739	.	C	A	.	PASS	DP=64;SS=1;SSC=3;GPV=4.3049e-14;SPV=0.44587
chr1	214788427	.	C	T	.	PASS	45;SS=1;SSC=1;GPV=8.5784e-10;SPV=0.66234
chr1	214802553	.	CT	C	.	PASS	DP=83;SOMATIC;SS=2;SSC=18;GPV=1;SPV=0.015148;INDEL
chr1	214803969	.	G	C	.	PASS	DP=111;SOMATIC;SS=2;SSC=35;GPV=1;SPV=0.00029013
chr1	214804041	.	C	A	.	PASS	DP=65;SS=1;SSC=0;GPV=2.7963e-08;SPV=0.9934
chr1	214811174	.	G	A	.	PASS	DP=76;SS=1;SSC=0;GPV=3.6183e-12;SPV=0.99124
chr1	214811244	.	C	G	.	PASS	DP=120;SS=1;SSC=0;GPV=1.7875e-19;SPV=0.92629
chr1	214813487	.	A	G	.	PASS	DP=291;SS=1;SSC=3;GPV=1.3526e-38;SPV=0.47444
chr1	214813782	.	A	G	.	PASS	DP=108;SS=1;SSC=0;GPV=1.7692e-19;SPV=0.98472
chr1	214813941	.	C	G	.	PASS	DP=86;SS=1;SSC=4;GPV=8.038e-16;SPV=0.34707
chr1	214814125	.	G	A	.	PASS	DP=80;SS=1;SSC=0;GPV=1.2414e-11;SPV=0.85982
chr1	214814582	.	G	A	.	PASS	DP=226;SS=1;SSC=5;GPV=3.0361e-32;SPV=0.28302
chr1	214814733	.	T	G	.	PASS	DP=244;SS=1;SSC=0;GPV=2.27499e-40;SPV=0.97323

Pipeline1
24 shown, 72 deleted
151.47 MB
88: VarScan somatic on data 82 and data 80
153 lines, 113 comments
format: **vcf**, génome de référence: **hg19**

Starting variant calling ..
Calling variants for contig: chr10

Contig chr10 finished.

Calling variants for contig: chr11

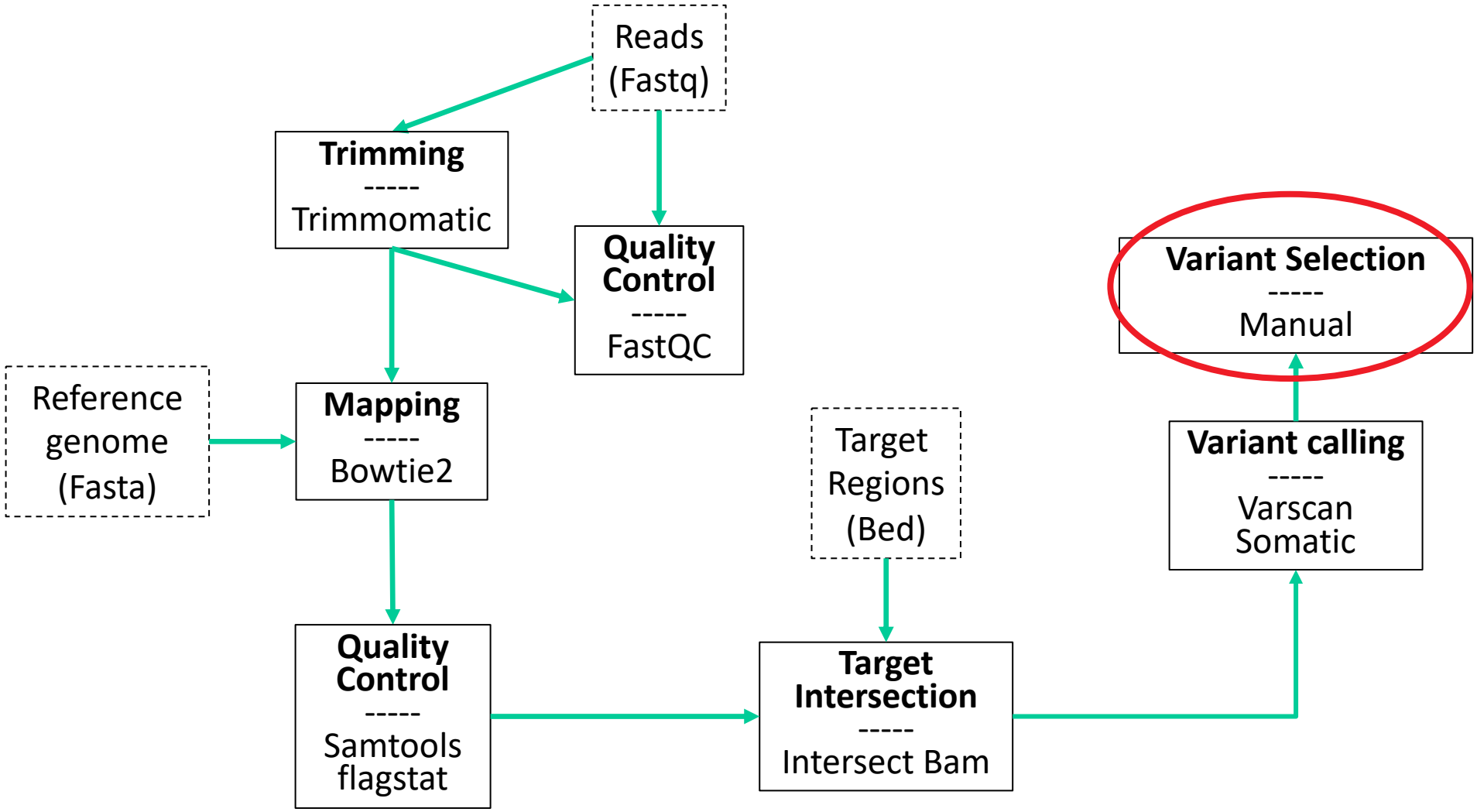
Contig chr11 finished.

Calling variants for contig: chr11_gl000202_random
Calling variants for contig: chr12

Contig chr12 finish

display at UCSC main
display with IGV local Human hg19
display at RViewer main

- Vous pouvez utiliser la fonction « grep » pour filter les lignes avec somatic ou LOH
- Vérifiez la somatic P-value (SPV), les comptages
- Regardez les sous IGV


Reads
(Fastq)
Trimming

Trimmomatic
Quality
Control

FastQC
Reference
genome
(Fasta)
Mapping

Bowtie2
Target
Regions
(Bed)
Variant Selection

Manual
Variant calling

Varscan
Somatic
Quality
Control

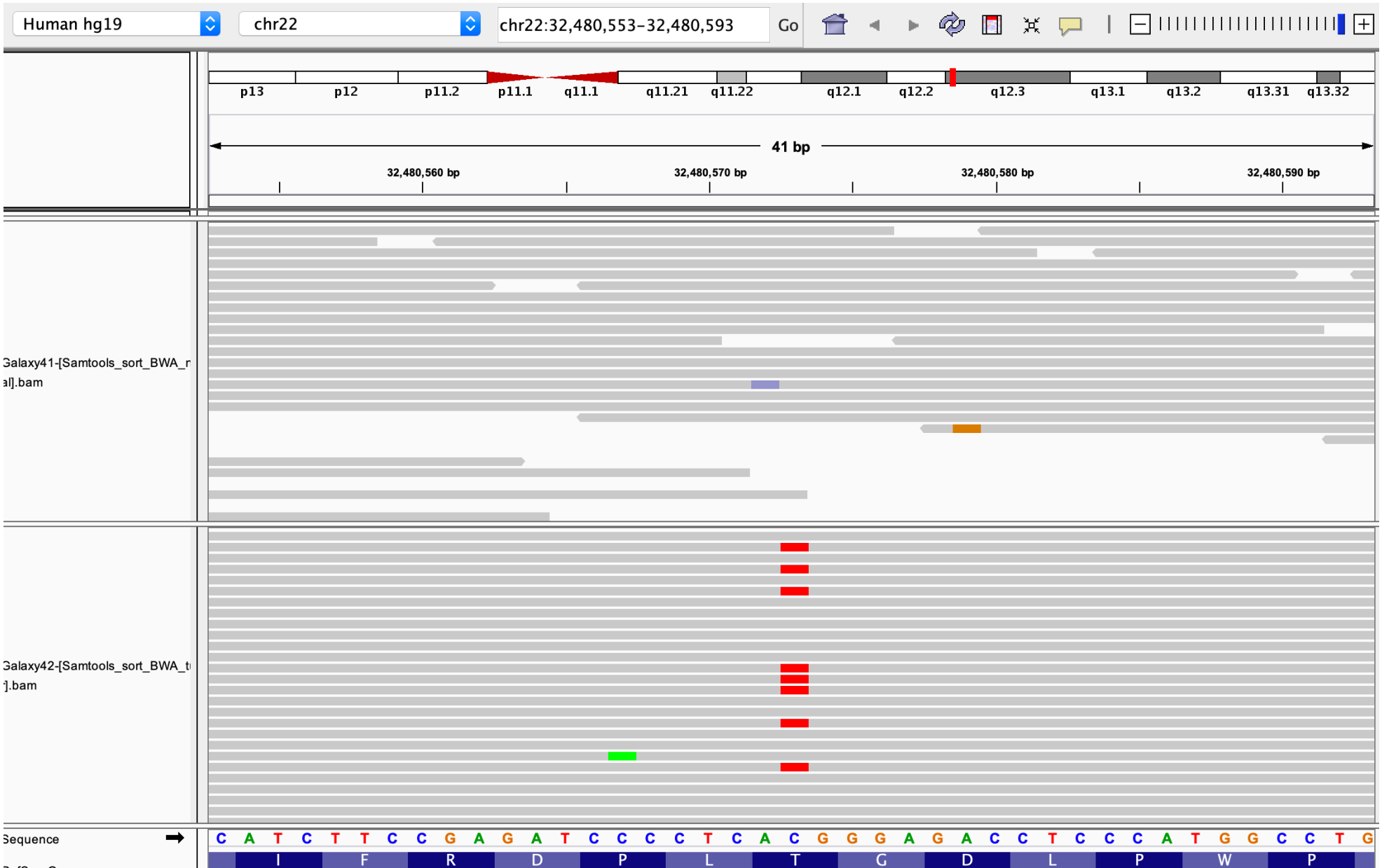
Samtools
flagstat
Target
Intersection

Intersect Bam

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 your local IGV session)
- Visualize somatic events.

IGV view


Human hg19
chr22
chr22:32,480,553-32,480,593
Go
p13
p12
p11.2
p11.1
q11.1
q11.21
q11.22
q12.1
q12.2
q12.3
q13.1
q13.2
q13.31
q13.32
41 bp
32,480,560 bp
32,480,570 bp
32,480,580 bp
32,480,590 bp
Galaxy41-[Samtools_sort_BWA_r
al].bam
Galaxy42-[Samtools_sort_BWA_t
].bam
Sequence
C A T C T T C C G A G A T C C C C T C A C G G G A G A C C T C C C A T G G C C T G
I F R D P L T G D L P W P

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

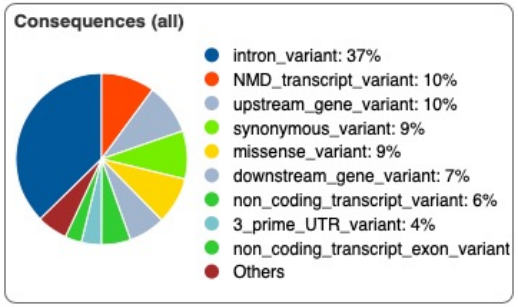
Note: the highest impact variants are not necessarily somatic!

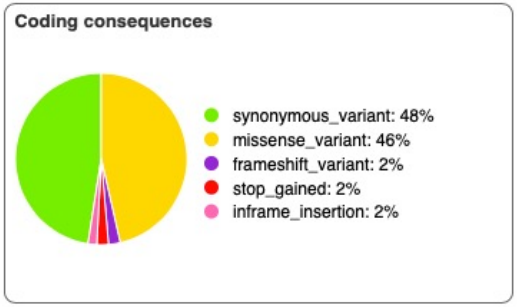
Variant Effect Predictor results

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

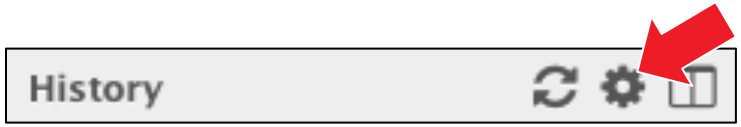
BioMart: Variants Genes

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

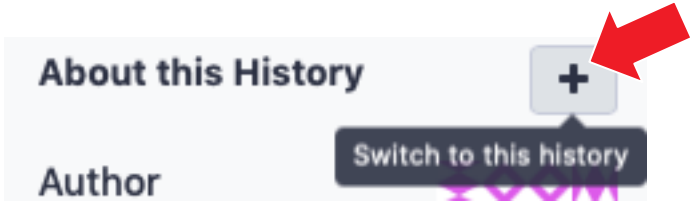
Galaxy: partager ses données



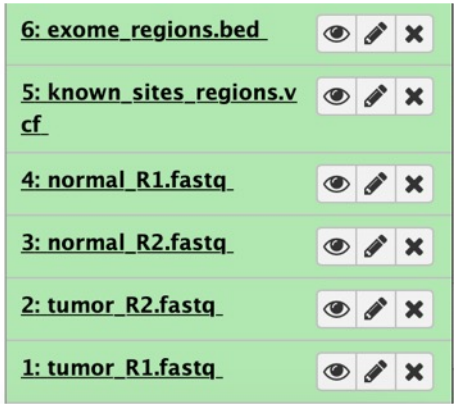
- Partager et publier
- Make History Accessible via Link
  - Cocher « also make all objects within the History accessible »

(lire des fichiers à partir de données partagées)

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



Fera apparaître:



Samtools mpileup sur fichier intersect bed

- Nécessaire sur Galaxy.fr (car version de Varscan différente, qui exige un fichier mpileup en entrée)