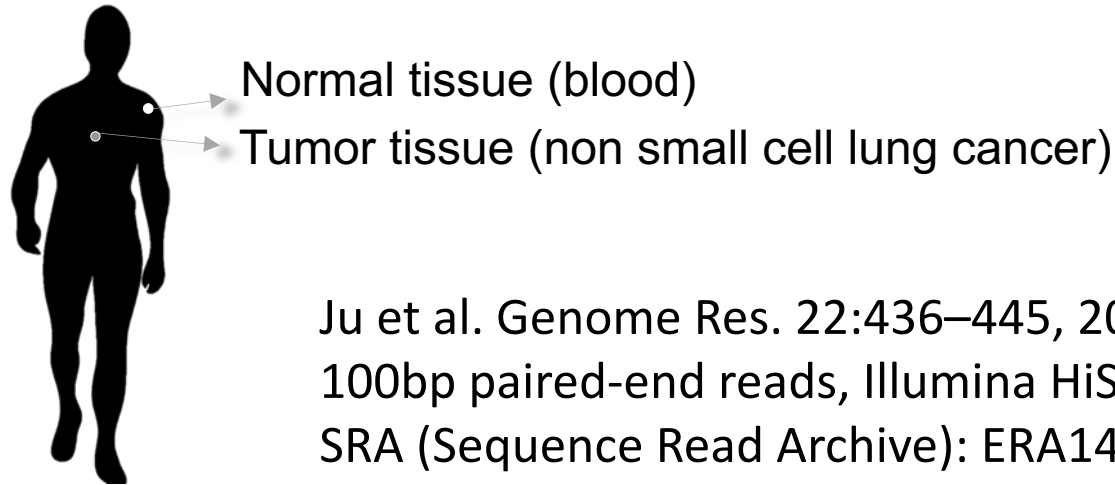


# Réalisation d'un Pipeline d'analyse d'exome

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

# Open access Galaxy servers

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

# Chargez vos données

(on peut aussi déposer des fichiers .gz)

The screenshot shows the Galaxy / Europe web interface. The top navigation bar includes links for 'Analyse de données', 'Workflow', 'Visualize', 'Données partagées', 'Aide', 'Utilisateur', and a 'Using 0%' indicator. The left sidebar contains a 'Tools' section with a search bar and categories like 'FILE AND META TOOLS', 'GENERAL TEXT TOOLS', and 'GENOMICS, NGS'. The main content area is divided into 'News' and 'Events' sections. The 'History' panel on the right shows a message: 'Cet historique est vide. You can Charger vos propres données or Charger des données depuis une source externe'. A red arrow points to this message. Below the message, a list of datasets is shown, including '6: exome\_regions.bed', '5: known\_sites\_regions.vcf', '4: normal\_R1.fastq', '3: normal\_R2.fastq', '2: tumor\_R2.fastq', and '1: tumor\_R1.fastq'.

Galaxy / Europe

Analyse de données Workflow Visualize Données partagées Aide Utilisateur Using 0%

Tools

search tools

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](#)

News

Jan 18, 2019  
! Queue cleared

Jan 11, 2019  
Another European CVMFS mirror is online

Jan 10, 2019  
The European Galaxy Team has open positions!

Jan 8, 2019  
New hardware: 8x1TB memory nodes

Events

Jul 1, 2019 - Jul 6, 2019  
2019 Galaxy Community Conference (GCC2019)

Mar 6, 2019 - Mar 8, 2019  
Galaxy for linking Bisulfite sequencing with RNA sequencing 06.-08.03.2019 in Rostock

Feb 25, 2019 - Mar 1, 2019  
Galaxy HTS data analysis workshop in Freiburg

Jan 28, 2019 - Feb 1, 2019  
2019 Galaxy Admin Training

History

Rechercher des données

exome test 2  
(empty)

Cet historique est vide. You can [Charger vos propres données](#) or [Charger des données depuis une source externe](#)

6: [exome\\_regions.bed](#)

5: [known\\_sites\\_regions.vcf](#)

4: [normal\\_R1.fastq](#)

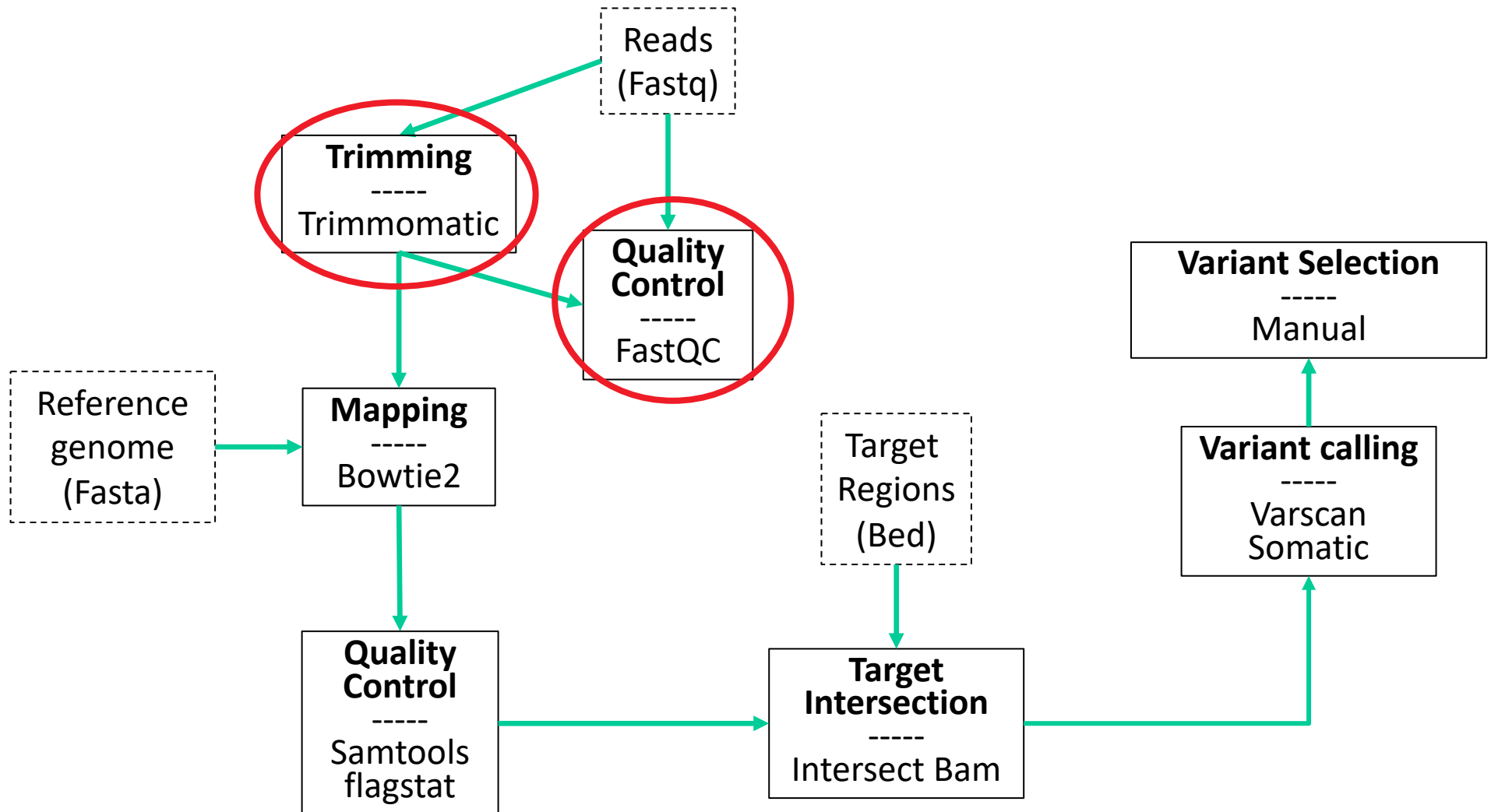
3: [normal\\_R2.fastq](#)

2: [tumor\\_R2.fastq](#)

1: [tumor\\_R1.fastq](#)

Fera apparaître:

# A simplified Variant Pipeline



# fastqc

Vérifiez les 4 fichiers fastq en mode multi-files

The screenshot displays the Galaxy web interface for the 'FastQC Read Quality reports' tool (Galaxy Version 0.71). The interface is divided into several sections:

- Top Navigation Bar:** Contains links for 'Analyse de données', 'Workflow', 'Visualize', 'Données partagées', 'Aide', 'Utilisateur', and a 'Using 0%' status indicator.
- Left Sidebar:** Lists tool categories: 'FASTA/FASTQ manipulation', 'Quality Control', and 'Mapping'. Under 'FASTA/FASTQ manipulation', 'fastqc' is highlighted.
- Central Tool Panel:**
  - Short read data from your current history:** A text input field containing a list of files: '4: normal\_R1.fastq', '3: normal\_R2.fastq', '2: tumor\_R2.fastq', and '1: tumor\_R1.fastq'. A red arrow points to this field.
  - Contaminant list:** A dropdown menu currently showing 'Nothing selected'.
  - Submodule and Limit specifying file:** A dropdown menu currently showing 'Nothing selected'.
  - Execute Button:** A blue button with a checkmark and the text 'Execute'.
- Right Sidebar:** Contains a 'History' section with a search bar and a list of recent jobs. The list includes 'exome test 2' (45.53 MB) and several files: '6: exome\_regions.bed', '5: known\_sites\_regions.vcf', '4: normal\_R1.fastq', '3: normal\_R2.fastq', '2: tumor\_R2.fastq', and '1: tumor\_R1.fastq'. A red arrow points to the file list in the central panel.

# Fastqc results

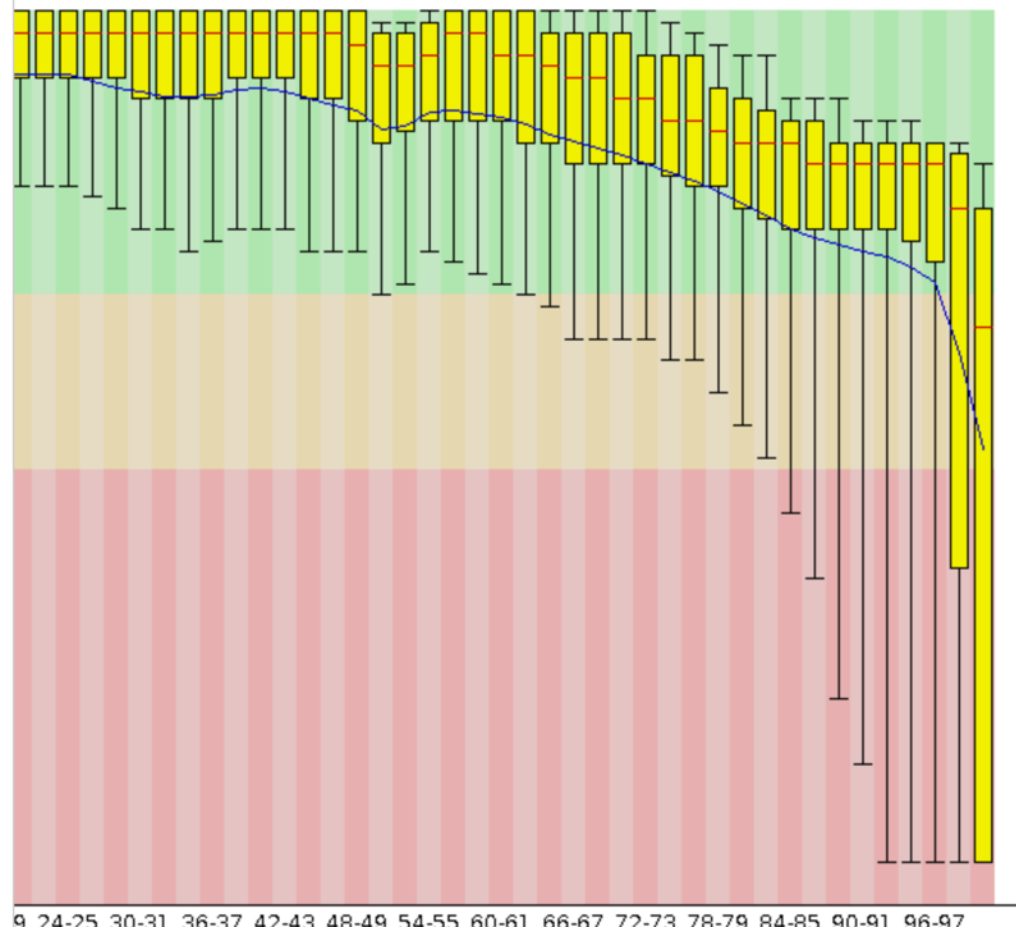
## FastQC Report

Sun 20 Jan 2019  
tumor\_R2\_fastq

### Summary

- ✓ Basic Statistics
- ✗ Per base sequence quality
- ✓ Per sequence quality scores
- ! Per base sequence content
- ! Per sequence GC content
- ✓ Per base N content
- ✓ Sequence Length Distribution
- ✓ Sequence Duplication Levels
- ✓ Overrepresented sequences
- ✓ Adapter Content

Quality scores across all bases (Sanger / Illumina 1.9 encoding)



History



Rechercher des données

exome test 2

14 shown

49.67 MB



14: FastQC on data 4: RawData



13: FastQC on data 4: Webpage



12: FastQC on data 3: RawData



11: FastQC on data 3: Webpage



10: FastQC on data 2: RawData



9: FastQC on data 2: Webpage



8: FastQC on data 1: RawData



7: FastQC on data 1: Webpage



6: exome\_regions.bed



5: known\_sites\_regions.vcf



- 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

Galaxy / Europe

Analyse de données Workflow Visualize Données partagées Aide Utilisateur Using 0%

Tools

trimmomatici

FASTA/FASTQ manipulation

fastp – fast all-in-one preprocessing for FASTQ files

Trimmomatic flexible read trimming tool for Illumina NGS data

Quality Control

Trimmomatic flexible read trimming tool for Illumina NGS data

Assembly

Shovill Faster SPAdes assembly of Illumina reads

Workflows

All workflows

Trimmomatic flexible read trimming tool for Illumina NGS data (Galaxy Version 0.36.0)

Paired end data?

Yes No

Input Type

Pair of datasets

Input FASTQ file (R1/first of pair)

4: normal\_R1.fastq

Input FASTQ file (R2/second of pair)

3: normal\_R2.fastq

Perform initial ILLUMINACLIP step?

Yes No

Cut adapter and other illumina-specific sequences from the read

Trimmomatic Operation

1: Trimmomatic Operation

Select Trimmomatic operation to perform

Sliding window trimming (SLIDINGWINDOW)

Number of bases to average across

4

Average quality required

20

+ Insert Trimmomatic Operation

Execute

History

Rechercher des données

ExomeTest

27 shown, 31 deleted, 1 hidden

242.36 MB

Binary bam alignments file

27: BWA NORMAL

24: Trimmomatic on normal\_R2.fastq (R2 paired)

23: Trimmomatic on normal\_R1.fastq (R1 paired)

22: BWA TUMOR

18: Trimmomatic on tumor\_R2.fastq (R2 paired)

17: Trimmomatic on tumor\_R1.fastq (R1 paired)

6: exome\_regions.bed

5: known\_sites\_regions.vcf

4: normal\_R1.fastq

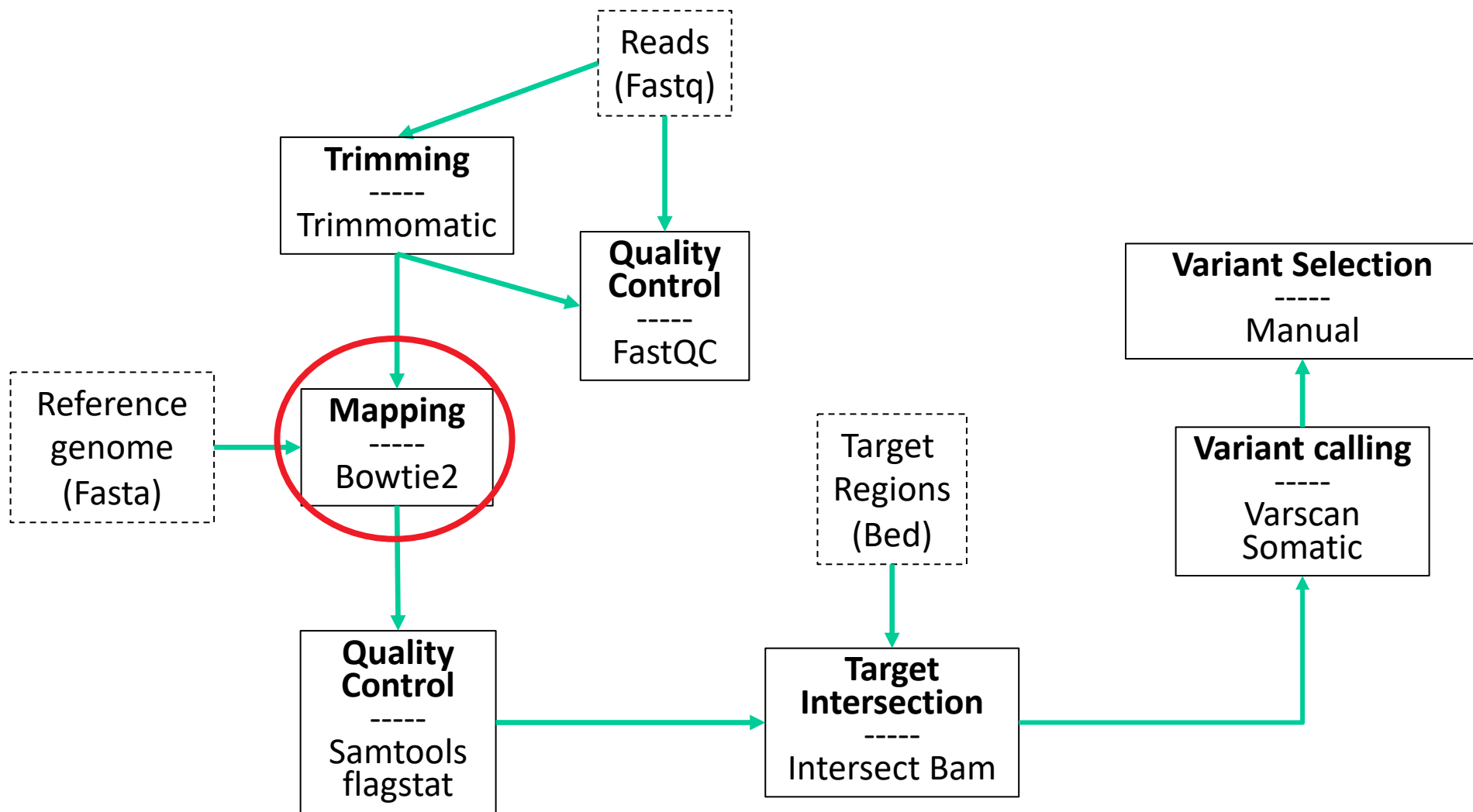
3: normal\_R2.fastq

2: tumor\_R2.fastq

Vérifiez à nouveau les fichiers corrigés avec fastqc

# Trimmomatic (fin)

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



# 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**

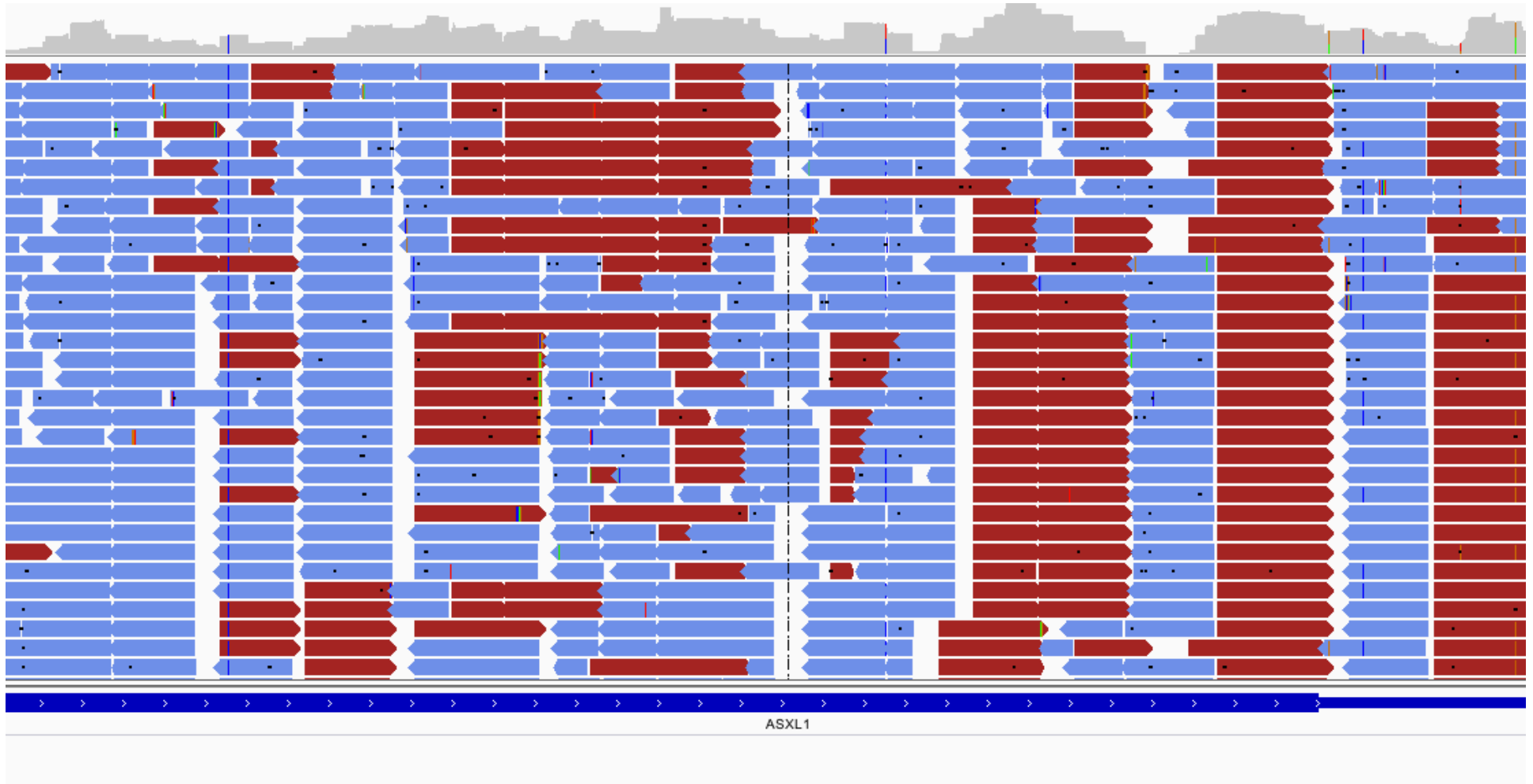
**41: Samtools sort BWA normal**

**40: Samtools sort Bowtie TUMOR**

**39: Samtools sort Bowtie NORMAL**

Check Bowtie result: what type of file is it?

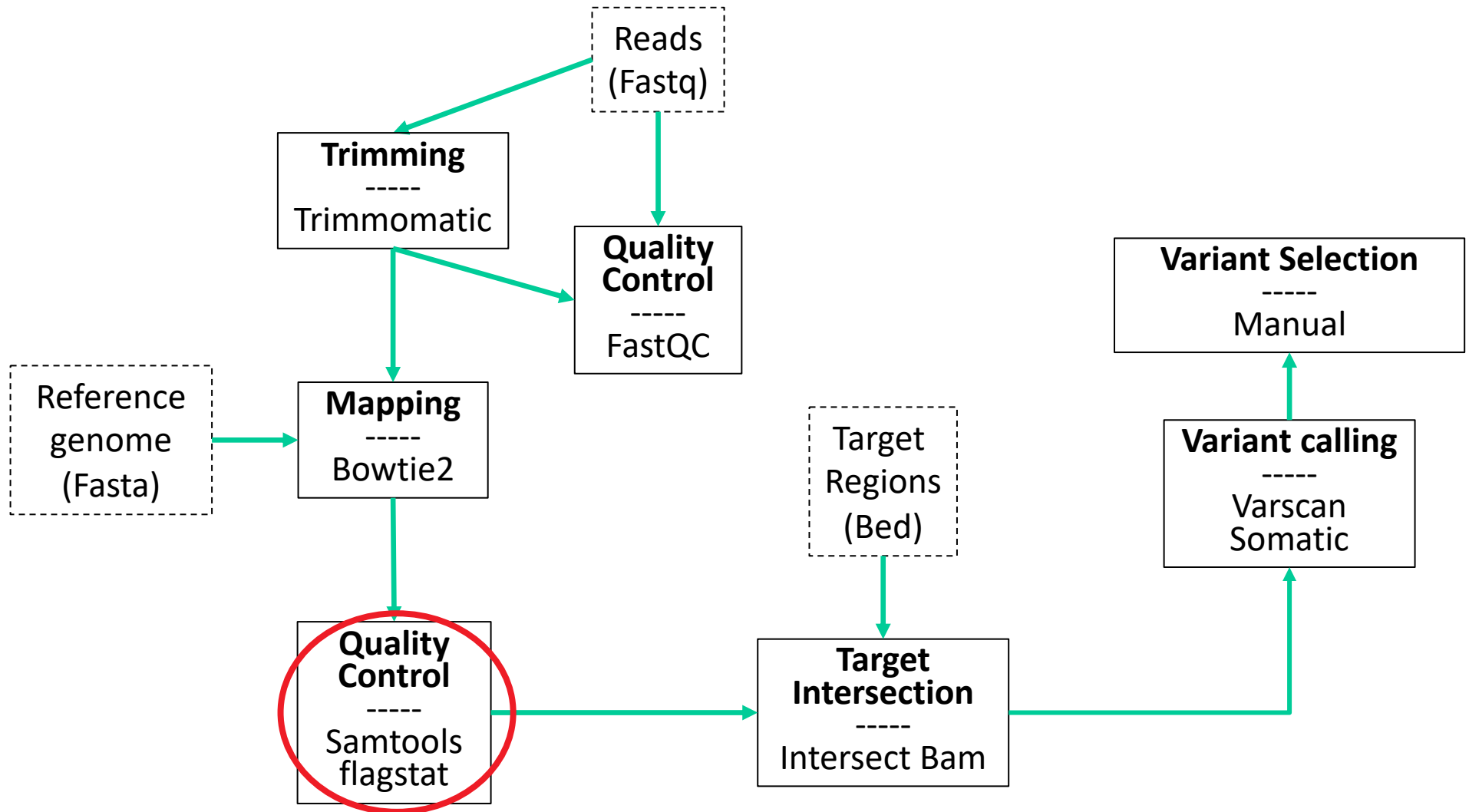
# Reads alignés: le format BAM/SAM



# 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
```







# 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

 **Samtools stats** generate statistics for BAM dataset (Galaxy Version 2.0.4)   

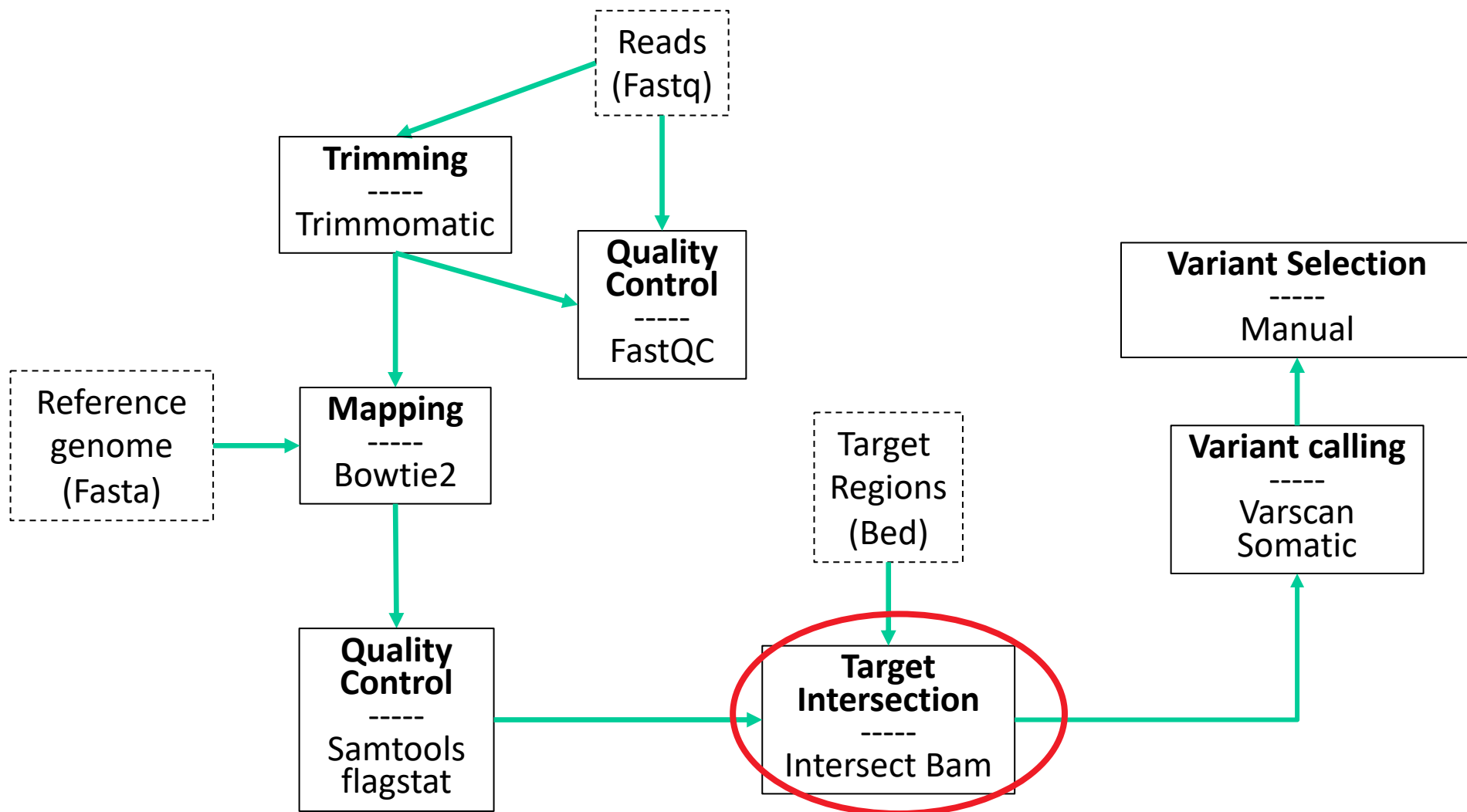
**BAM file**

   21: Bowtie2 on data 14 and data 13: alignments   

**réultat**

```
# 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 second
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



# Bedtools intersect intervals

**Galaxy / Europe** Analyze Data Workflow Visualize Shared Data Help User Using 0%

**Tools** [bedtools\\_intersectbed](#)

**BED Tools**  
[Intersect intervals](#) find overlapping intervals in various ways

**Operate on Genomic Intervals**  
[bedtools Intersect intervals](#) find overlapping intervals in various ways

**Workflows**  
All workflows

**bedtools Intersect intervals find overlapping intervals in various ways (Galaxy Version 2.27.1)** Versions Options

**File A to intersect with B**  
39: Samtools sort Bowtie NORMAL  
BAM/bed,bedgraph,gff,vcf format

**Combined or separate output files**  
One output file per 'input B' file

**File(s) B to intersect with A**  
6: exome\_regions.bed  
BAM/bed,bedgraph,gff,vcf format

**Calculation based on strandedness?**  
Overlaps on either strand

**What should be written to the output file?**  
☐ Select/Unselect all

**Treat split/spliced BAM or BED12 entries as distinct BED intervals when computing coverage.**  
Yes No  
If set, the coverage will be calculated based the spliced intervals only. For BAM files, this inspects the CIGAR N operation to infer the blocks for computing coverage. For BED12 files, this inspects the BlockCount, BlockStarts, and BlockEnds fields (i.e., columns 10,11,12). If this option is not set, coverage will be calculated based on the interval's START/END coordinates, and would include introns in the case of RNAseq data. (-split)

**Required overlap**  
Default: 1bp

**Report only those alignments that **\*\*do not\*\*** overlap with file(s) B**  
Yes No

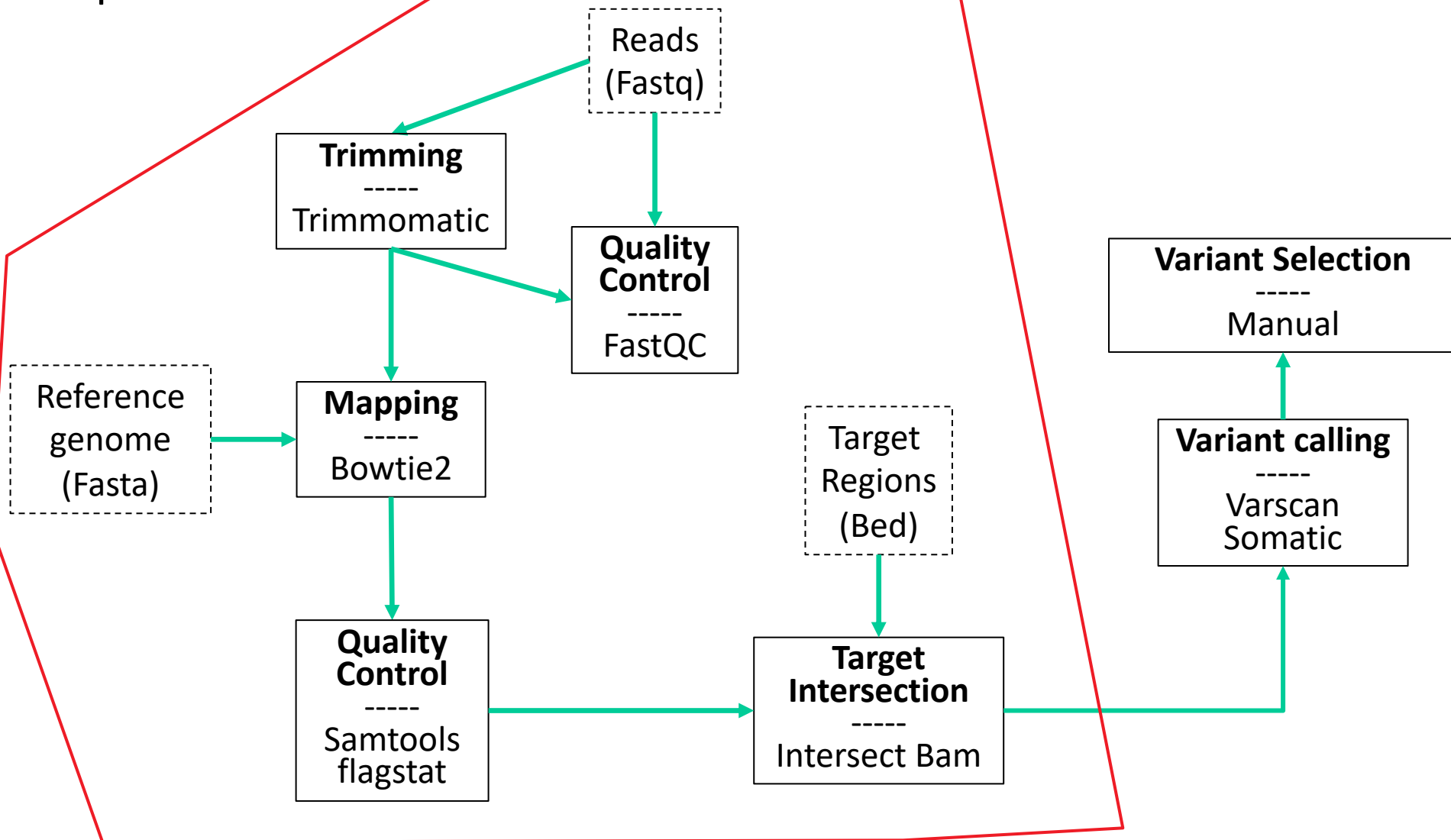
**History** search datasets

**ExomeTest**  
31 shown, 34 deleted  
242.39 MB

- 57: Intersect intervals on data 6 and data 30
- 56: Samtools flagstat on data 27
- 55: VarScan somatic o n data 42
- 48: VarScan mpileup o n BWA
- 47: VarScan mpileup o n bowtie
- 46: samtools mpileup on bwa
- 45: samtools mpileup on Bowtie
- 42: Samtools sort BWA tumor
- 41: Samtools sort BWA normal

Vérifiez la réduction de taille du fichier BAM


Repeat for Tumor



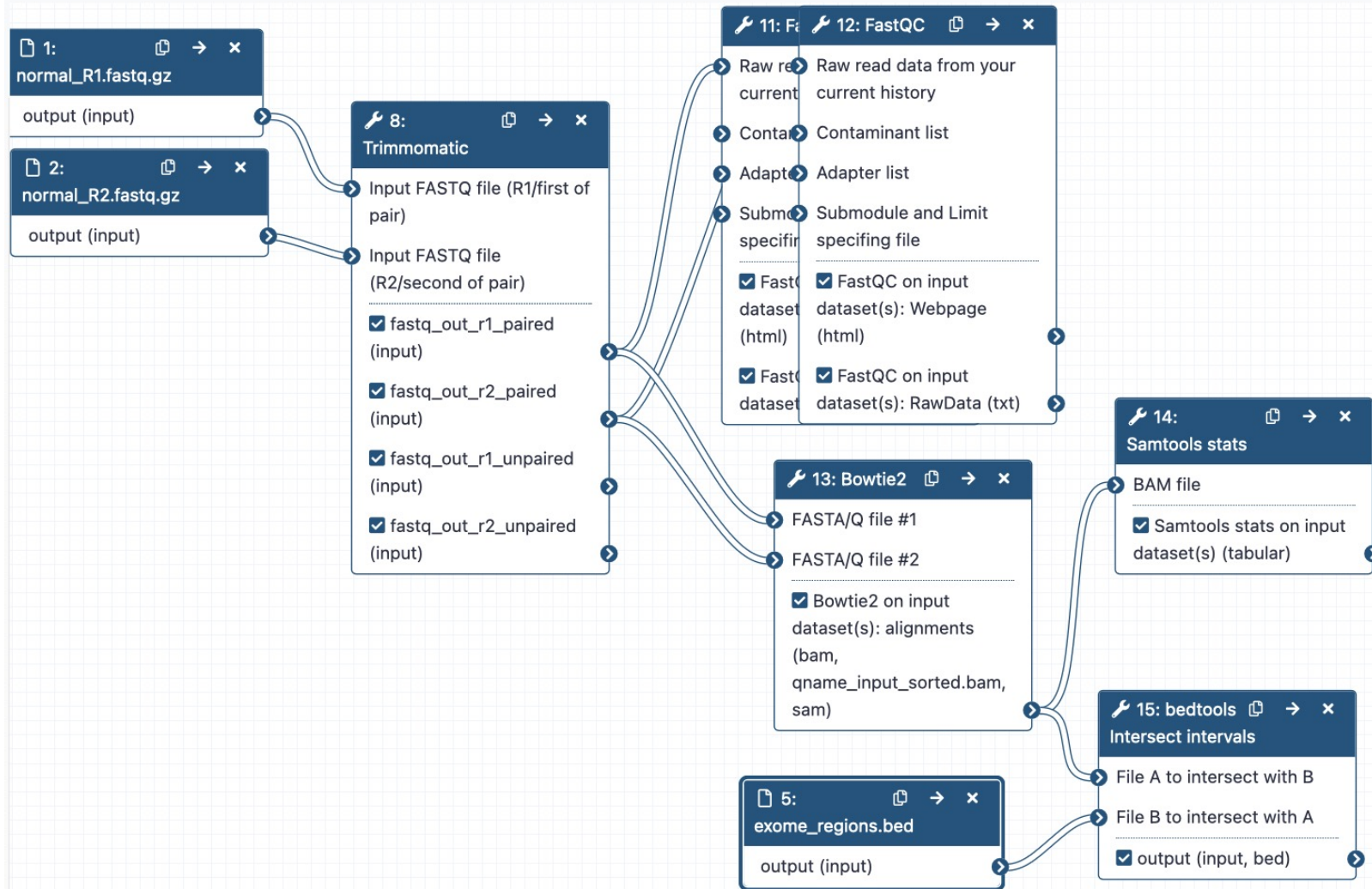
# Extraire un workflow

History



- 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 » plutôt que « normal »)
- Puis  save workflow

## align\_and\_check\_1\_sample



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

Galaxy / Europe

Analyse de données Workflow Visualize Données partagées Aide Utilisateur

Workflow: OneSample ✓ Run workflow

History Options

Send results to a new history

Yes No

1: Fastq R1

1: tumor\_R1.fastq

2: Fastq R2

2: tumor\_R2.fastq

3: exome\_regions.bed

6: exome\_regions.bed

4: Trimmomatic (Galaxy Version 0.36.0)

5: Bowtie2 (Galaxy Version 2.3.4.2)

6: bedtools Intersect intervals (Galaxy Version 2.27.1)

7: Samtools flagstat (Galaxy Version 2.0.2)

History

Rechercher d

Pipeline1

25 shown, 73 del

151.47 MB

Warning: skip

'ERR160124.40

length (1) <=

Warning: skip

'ERR160124.40

was < 2 charac

43398 reads; c

43398 (100.00

display at UCSC

display at Enser

display with IGV

display in IGB V

Binary bam ali

76: Trimmoma

or\_R2.fastq (R2

75: Trimmoma

or\_R1.fastq (R1

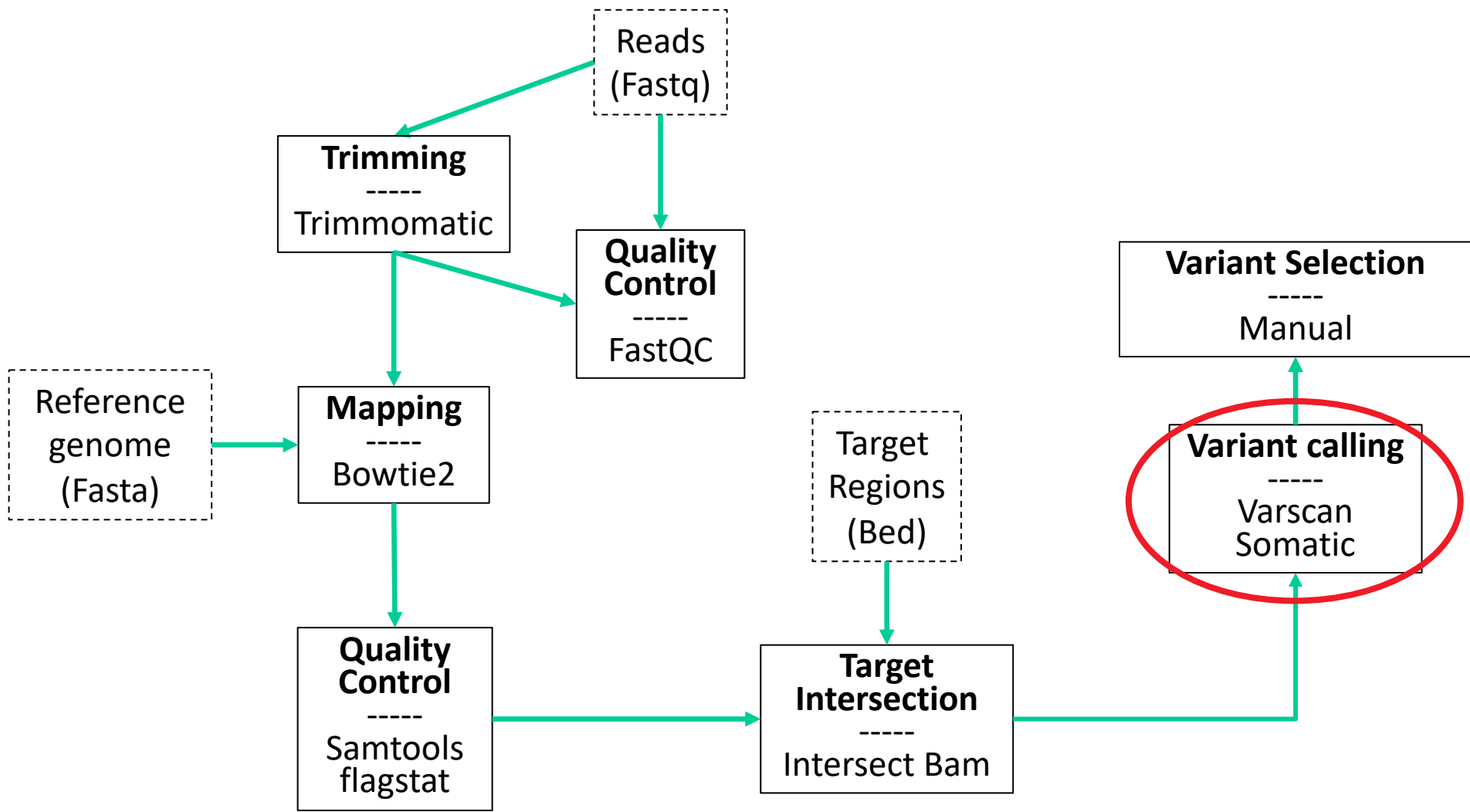
72: Trimmoma

mal\_R2.fastq (f

71: Trimmoma

mal\_R1.fastq (f

6: exome\_regio





# Somatic variant calling: Varscan

Attention: étape de 10-30min!

**Varscan somatic Call germline/somatic and LOH variants from tumor-normal sample pairs (Galaxy Version 2.4.3.3)**

Will you select a reference genome from your history or use a built-in genome?  
Use a built-in genome

**reference genome**  
Human (Homo sapiens): hg19

The fasta reference genome that variants should be called against.

**aligned reads from normal sample**  
44: NORMAL BAM intersect

**aligned reads from tumor sample**  
54: TUMOR BAM intersect

**Estimated purity (non-tumor content) of normal sample**  
1

(--normal-purity)

**Estimated purity (tumor content) of tumor sample**  
1

(--tumor-purity)

**Generate separate output datasets for SNP and indel calls?**  
Yes No

**Settings for Variant Calling**  
Use default values

**Settings for Posterior Variant Filtering**  
Do not perform posterior filtering

**Execute**

**Varscan Overview**  
Varscan performs variant detection for massively parallel sequencing data, such as exome, WGS, and transcriptome data. Full documentation of the command line package is available [here](#).  
The Varscan Somatic tool for Galaxy

**History**  
Rechercher des données  
exome test 2  
29 shown, 26 deleted  
170.08 MB  
55: Samtools flagstat on data 53  
54: TUMOR BAM intersect  
5.5 MB  
format: bam, génome de référence: hg19  
53: Bowtie2 on data 46 and data 45: aligned reads (BAM)  
52: FastQC on data 46: RawData  
51: FastQC on data 46: Webpage  
50: FastQC on data 45: RawData  
49: FastQC on data 45: Webpage  
48: Trimmomatic on tumor R2.fastq (R2 unpaired)

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](#)

[Genome editing](#)

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	DP=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

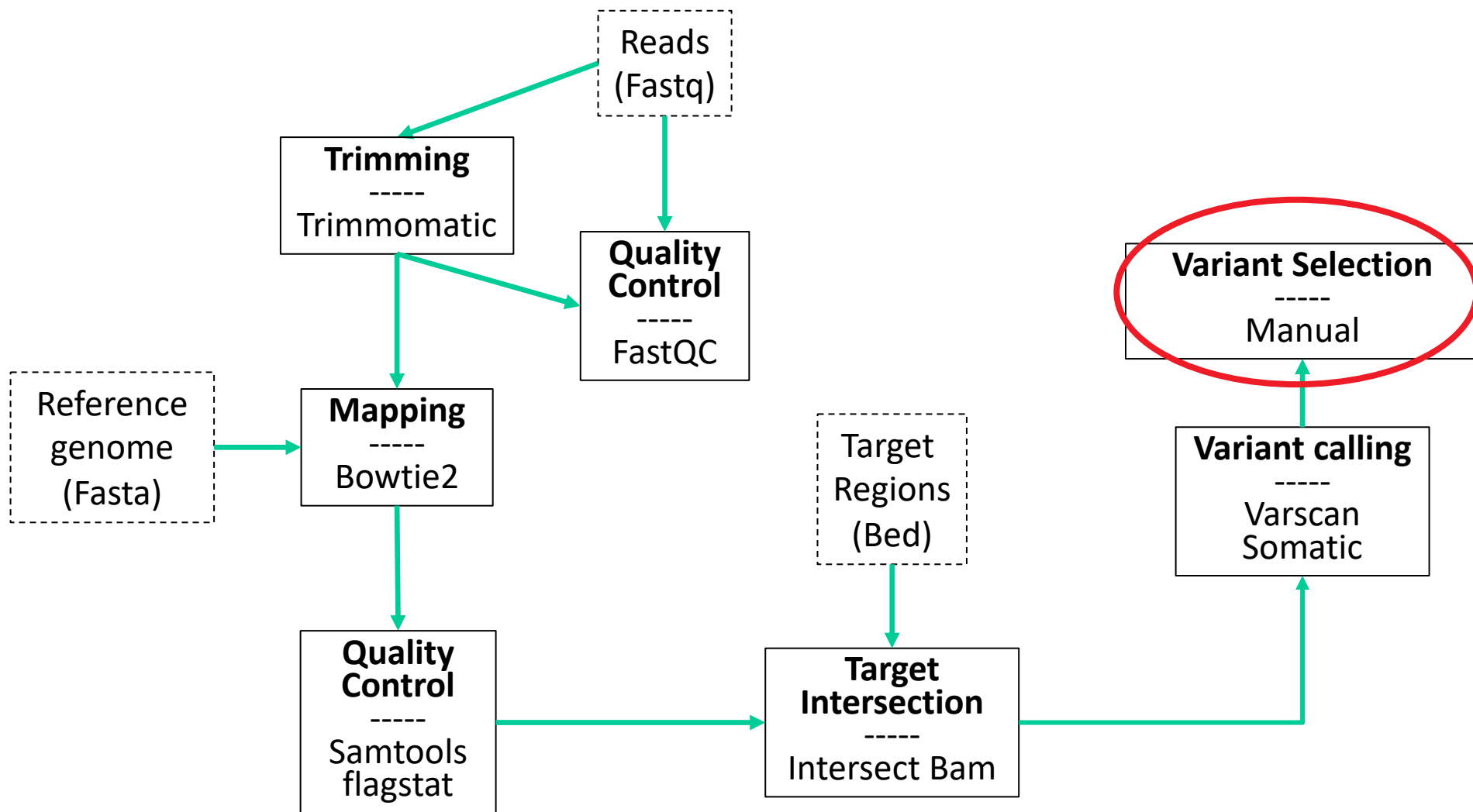
Contig chr12 finish

display at UCSC [main](#)

display with IGV [local](#) [Human](#) [hg19](#)

display at RViewer [main](#)

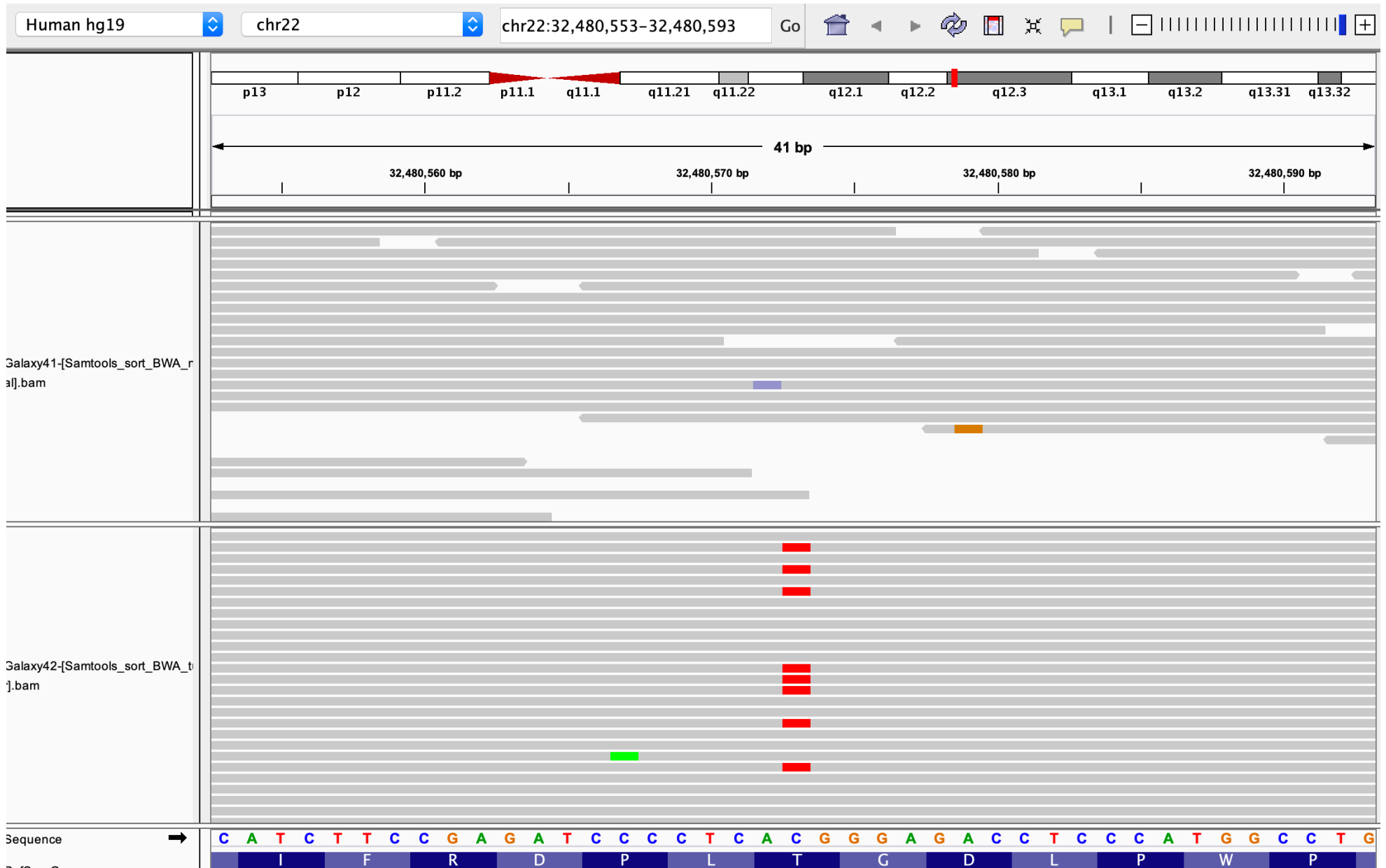
- Vous pouvez utiliser la fonction « grep » pour filter les lignes avec somatic ou LOH
- Vérifiez la somatic P-value (SPV), les comptages
- Regardez les sous IGV



# 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



# 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

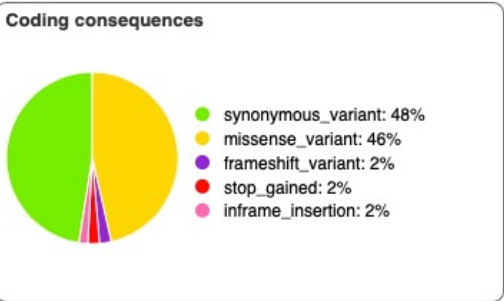
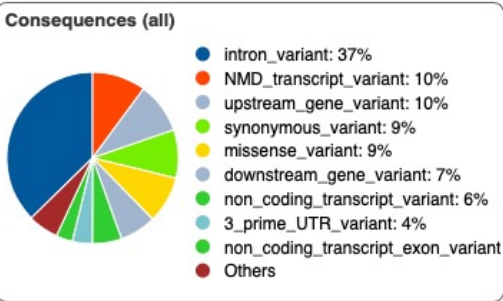
Note: the highest impact variants are not necessarily somatic!

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
.	<a href="#">1:248059779-248059779</a>	A	frameshift_variant	HIGH	OR2W3	<a href="#">ENSG00000238243</a>	Transcript	<a href="#">ENST00000360358</a>	protein_coding	1/1	-	891-8
.	<a href="#">1:248059779-248059779</a>	A	frameshift_variant	HIGH	OR2W3	<a href="#">ENSG00000238243</a>	Transcript	<a href="#">ENST00000537741</a>	protein_coding	3/3	-	1148-
.	<a href="#">3:121416308-121416308</a>	T	stop_gained	HIGH	GOLGB1	<a href="#">ENSG00000173230</a>	Transcript	<a href="#">ENST00000340645</a>	protein_coding	13/22	-	3173
.	<a href="#">3:121416308-121416308</a>	T	stop_gained	HIGH	GOLGB1	<a href="#">ENSG00000173230</a>	Transcript	<a href="#">ENST00000393667</a>	protein_coding	13/22	-	3173
.	<a href="#">3:121416308-121416308</a>	T	stop_gained	HIGH	GOLGB1	<a href="#">ENSG00000173230</a>	Transcript	<a href="#">ENST00000489400</a>	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 "+"



6: <u>exome_regions.bed</u>			
5: <u>known_sites_regions.vcf</u>			
4: <u>normal_R1.fastq</u>			
3: <u>normal_R2.fastq</u>			
2: <u>tumor_R2.fastq</u>			
1: <u>tumor_R1.fastq</u>			

Fera apparaître:

# Samtools mpileup sur fichier intersect bed

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