

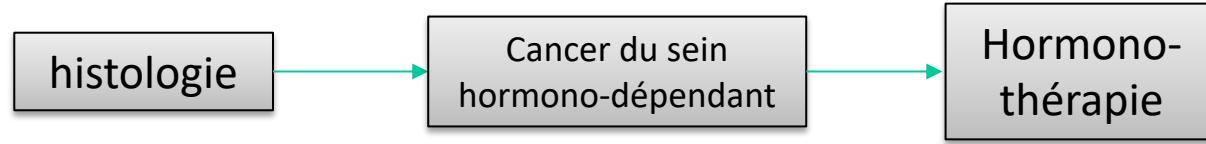
# Réalisation d'un Pipeline d'analyse d'exome

<https://github.com/gustaveroussy/ifsbm-bigdata>

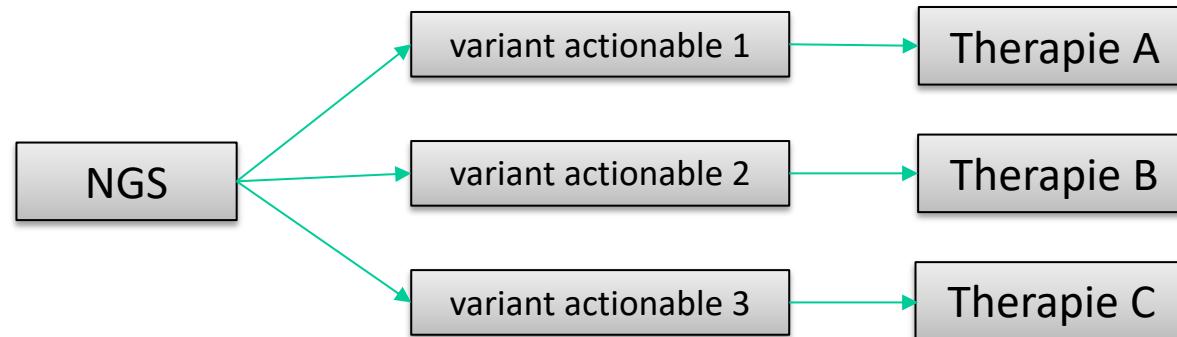
# Pour quoi faire?

# Rappel: Thérapie systémique vs. de précision

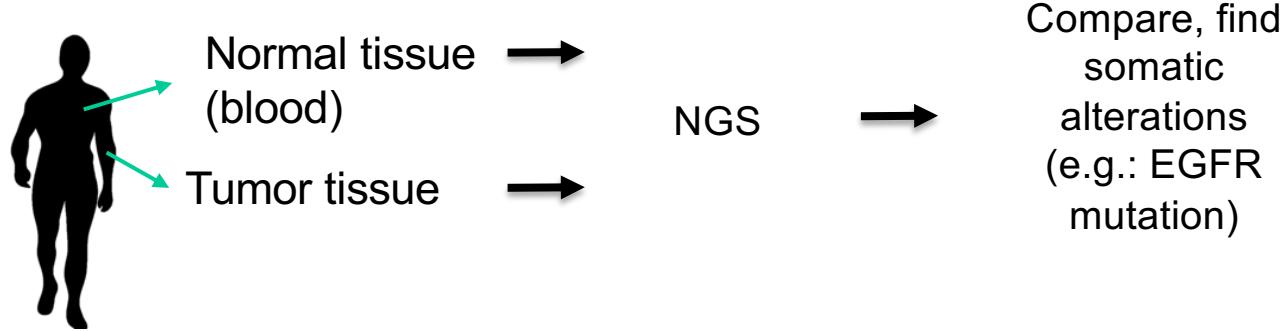
Chimio-thérapie  
Systémique



Chimio-thérapie  
de précision

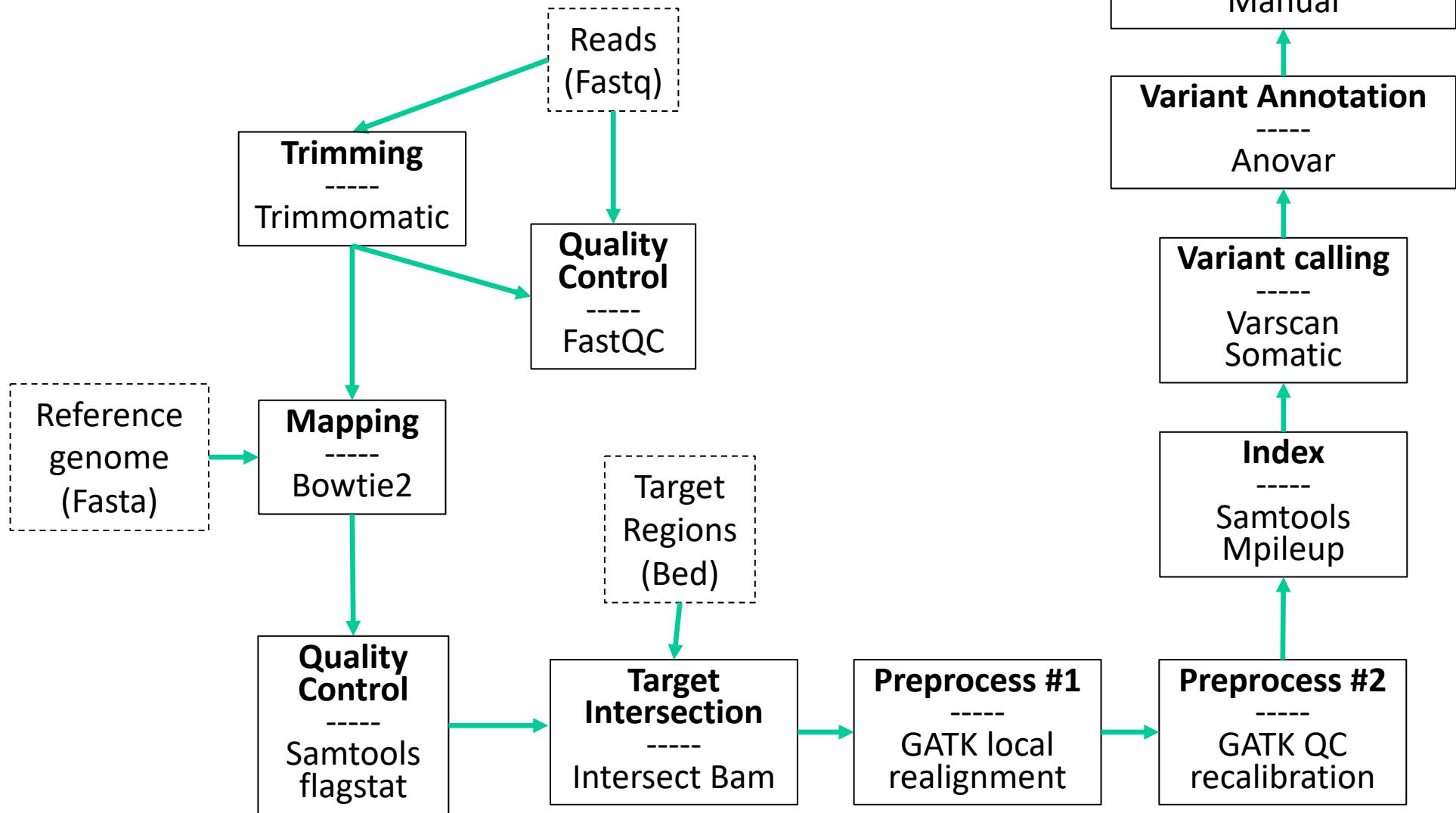


# Looking for cancer mutations

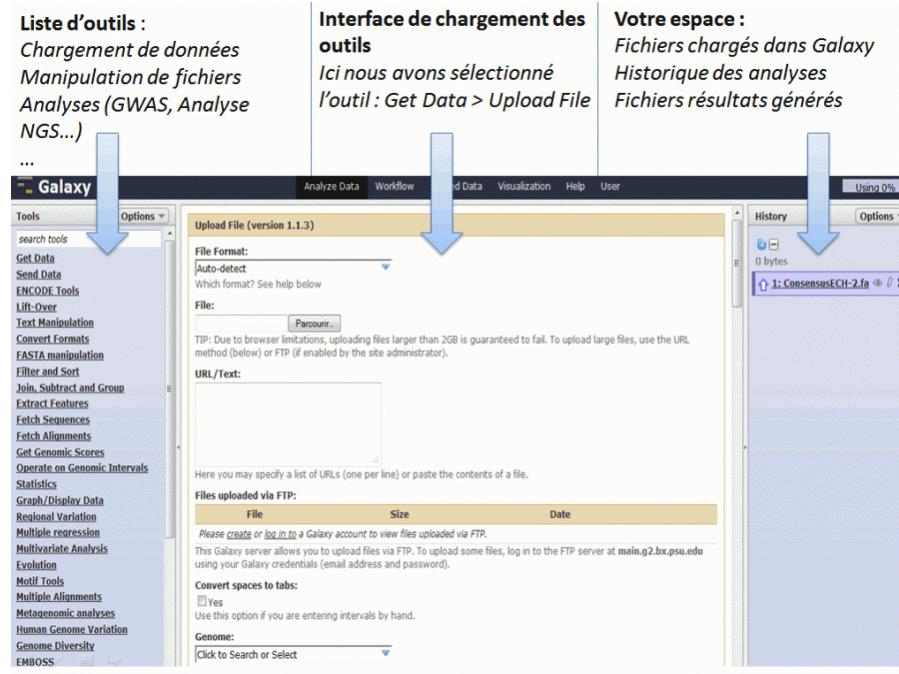


# Déployer un pipeline Bioinformatique

# Un pipeline « NGS Variant » utilisé à Gustave Roussy



# Galaxy: user-friendly interface to NGS pipelines



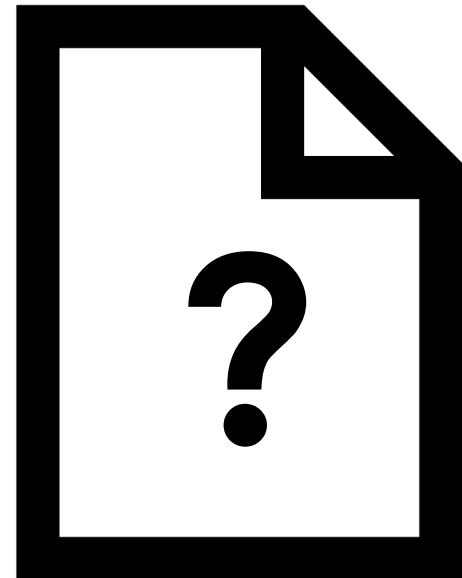
Credit: Biorigami

- Interest: avoiding Unix command line + traçability
- But: running NGS workflow on real human data often requires a computer cluster (will not run on a single-node Galaxy server)

# Open access Galaxy servers

- <https://usegalaxy.eu> ← Start with this one
- <https://usegalaxy.fr>

# Les données NGS



# FASTQ Format

```
@NB501949:31:H2NWHBGX3:1:11101:13085:7526 1:N:0:AGTTCC
CCAGACTCAGCCGAGACAGATCAGGACCGTGAGGATGGGGTCATGGGTCTCCACTGCCCTGCTGTG
+
AAAAA6AEAEAEEAEAAEEEEEE6EE6/AEEE/Eeeeeeee6eeeeeeeAEEEAEEEE6EEEEE/EEE
@NB501949:31:H2NWHBGX3:1:11101:7216:7526 1:N:0:AGTTCC
GTTTGTGTTGTTTTGAGACAGGTATTGCTCTGTCATCCAGGCCAGAGTGTAGTGGGTGATCACCACTCACTGC
+
AAAAAEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE/EA/EEE
@NB501949:31:H2NWHBGX3:1:11101:14260:7526 1:N:0:AGTTCC
GCCAGGCATAGGCTACCCAGTGGTCTCAAAGTGTCTCCTGGATCAGCAGCAGCATCACCGGGGATGGA
+
AAAAAEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE
@NB501949:31:H2NWHBGX3:1:11101:22341:7527 1:N:0:AGTTCC
CCCACCACCAAGAAATGAACAAAAGCATTACCTAAAATACACCAGCAAATGTACTCAGCTCAATCACAAAT
+
AAAAAEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE
@NB501949:31:H2NWHBGX3:1:11101:22098:7527 1:N:0:AGTTCC
GCCGAAGCCACTCCACTGTCTCAGCATTGACTGAAAAAGTCCTGCTCCAGACCTCCGTGTTAGCC
+
AAAAAEEEEEEEEEEAEEEEEEEEEEEAEEEEEEEAEEEEEEEAEEEEEEEAEEEEEEEAEEEEEEE
@NB501949:31:H2NWHBGX3:1:11101:8707:7528 1:N:0:AGTTCC
GTCTGAGGACCTCTGTATTTGTCAATTCTCCACGTTCTCGGCCTGTTCCGTAGCCTCATGAGCT
+
AAAAAEEEEEEEEEEAEEEEEEEAEEEEEEEAEEEEEEEAEEEEEEEAEEEEEEEAEEEEEEEAE
@NB501949:31:H2NWHBGX3:1:11101:4370:7528 1:N:0:AGTTCC
GGCCACTGCACCCAGTTATCGTGTGCAACTGTAAACCTTGAATAAACACCATGGGCCATACGA
+
AAAAAEEEE6EEEEEEAEEEEEEEAEEEEEEEAEEEEEEEAEEEEEEEAEEEEEEEAE
```

# Fastq format quality

@SEQ\_ID1  
CCAATGCT  
+  
8ASR/2@B

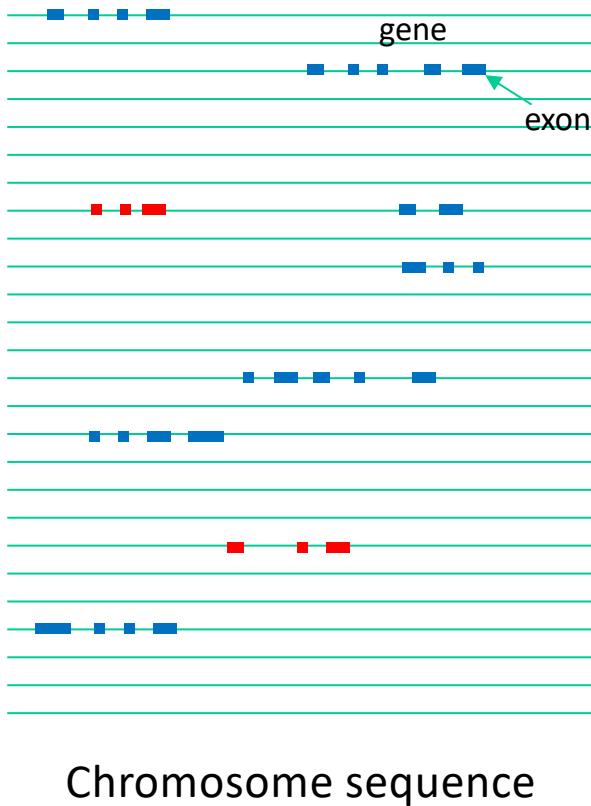
@SEQ\_ID1  
CCAATGCT  
+  
8ASR/2@B  
  
16, 25, 43, 42, 7, ...



ASCII Code

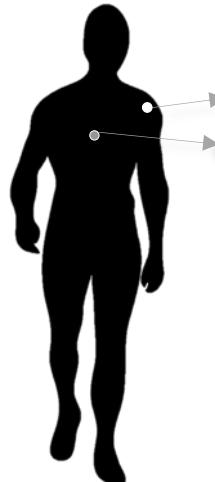
8	32	64	96	'	128	ç	160	á	192	l	224	ó
1	33 !	65 A	97 a	129 ü	161 í	193 Ł	193	ł	225	ø		
2	34 "	66 B	98 b	130 é	162 ó	194 T	194	t	226	ö		
3	35 #	67 C	99 c	131 á	163 ú	195 Þ	195	þ	227	ð		
4	36 \$	68 D	100 d	132 à	164 ñ	196 —	196	—	228	ñ		
5	37 %	69 E	101 e	133 à	165 Ñ	197 þ	197	þ	229	ð		
6	38 &	70 F	102 f	134 á	166 ø	198 å	198	å	230	ø		
	39 '	71 G	103 g	135 ç	167 œ	199 Á	199	Á	231	é		
	40 <	72 H	104 h	136 è	168 Ł	200 Ł	200	ł	232	ę		
	41 >	73 I	105 i	137 ē	169 ®	201 Í	201	í	233	ú		
	42 *	74 J	106 j	138 è	170 Í	202 Í	202	í	234	ú		
11	43 +	75 K	107 k	139 ï	171 Ł	203 Ł	203	ł	235	ó		
12	44 -	76 L	108 l	140 í	172 Ł	204 Ł	204	ł	236	ø		
13	45 .	77 M	109 m	141 á	173 ð	205 ð	205	ð	237	ð		
14	46 ,	78 N	110 n	142 à	174 «	206 «	206	«	238	—		
15	47 /	79 O	111 o	143 á	175 »	207 »	207	»	239	—		
16	48 0	80 P	112 p	144 È	176 Ø	208 Ø	208	Ø	240	—		
17	49 1	81 Q	113 q	145 æ	177 Ø	209 Ø	209	Ø	241	±		
18	50 2	82 R	114 r	146 È	178 Ø	210 Ø	210	Ø	242	=		
19	51 3	83 S	115 s	147 ô	179 —	211 —	211	—	243	—		
20	52 4	84 T	116 t	148 ö	180 —	212 —	212	—	244	—		
21	53 5	85 U	117 u	149 ö	181 á	213 á	213	á	245	§		
22	54 6	86 V	118 v	150 û	182 á	214 í	214	í	246	÷		
23	55 7	87 W	119 w	151 ù	183 á	215 í	215	í	247	ö		
24	56 8	88 X	120 x	152 ü	184 ®	216 Í	216	Í	248	—		
25	57 9	89 Y	121 y	153 ö	185 ®	217 Í	217	Í	249	—		
26	58 :	90 Z	122 z	154 ü	186	218 —	218	—	250	—		
27	59 ;	91 [	123 {	155 ø	187 —	219 ■	219	■	251	—		
28	60 <	92 \	124 :	156 £	188 —	220 ■	220	■	252	—		
29	61 =	93 ]	125 >	157 Ø	189 ¢	221 —	221	—	253	—		
30	62 >	94 ^	126 ~	158 ×	190 ¥	222 —	222	—	254	■		
31	63 ?	95 _	127 △	159 f	191 ı	223 ■	223	■	255			

# Rappel: DNA sequencing types



- WGS= Whole genome (3Gb)
- WES: whole exome (50Mb)
- Panel: selected genes (200kb)

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

# Récupérer les fichiers

- [https://drive.google.com/drive/folders/1TggAAFH9Ao0MS8WuerrGEEIQxttnUJOY?usp=share\\_link](https://drive.google.com/drive/folders/1TggAAFH9Ao0MS8WuerrGEEIQxttnUJOY?usp=share_link)
  - normal\_R1.fastq.gz
  - normal\_R2.fastq.gz
  - tumor\_R1.fastq.gz
  - tumor\_R2.fastq.gz
  - exome\_regions.bed

# Chargez les 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 History. The main content area features a quote by Prof. Stephen Hawking: "Anyone, anywhere in the world should have free, unhindered access to not just my research, but to the research of every great and enquiring mind across the spectrum of human understanding." Below the quote, there are sections for News and Events.

**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

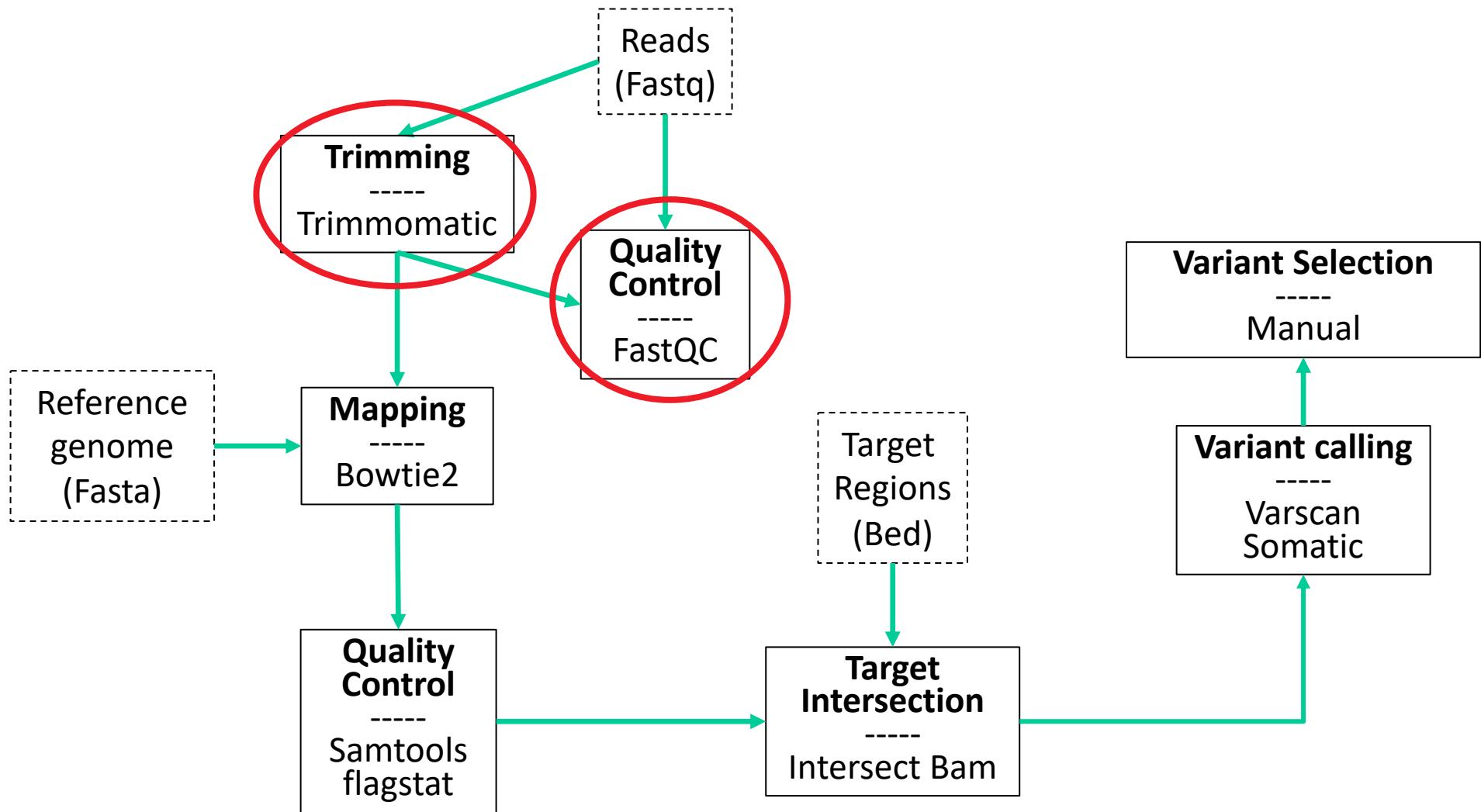
**History:** The History panel shows a single entry named "exome test 2" which is currently empty. A red arrow points to a tooltip in this panel: "Cet historique est vide. You can Charger vos propres données or Charger des données depuis une source externe".

**File List:** On the right, a file list displays several files with green icons and edit buttons:

- 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:** This text is overlaid at the bottom right of the image.

# A simplified Variant Pipeline



# fastqc

Vérifiez les 4 fichiers fastq en mode multi-files

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 Read Quality reports (Galaxy Version 0.71)

Short read data 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 specifying 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

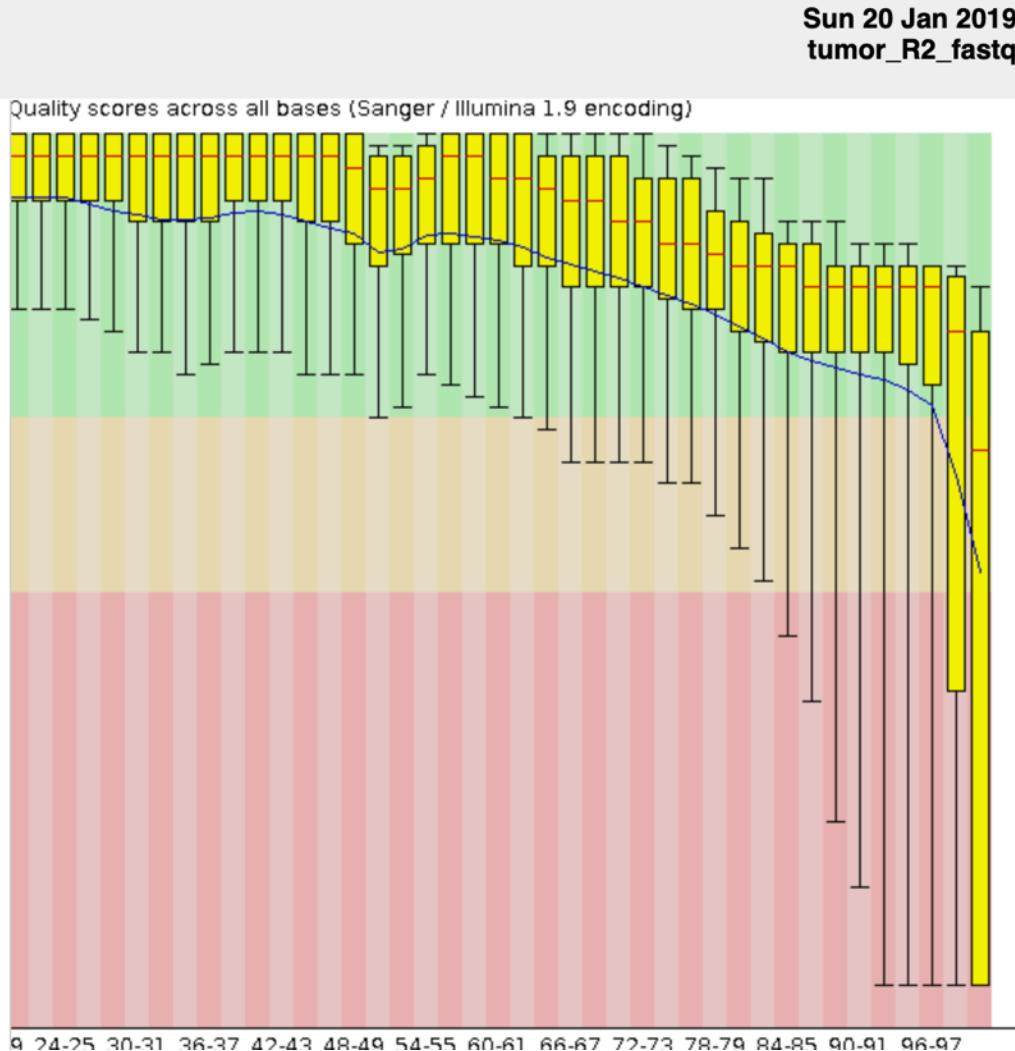
1: tumor\_R1.fastq

# Fastqc results

## FastQC Report

### 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](#)



History	Rechercher des données
exome test 2	14 shown
49.67 MB	<input checked="" type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
<a href="#">14: FastQC on data 4: Ra wData</a>	  
<a href="#">13: FastQC on data 4: We bpage</a>	  
<a href="#">12: FastQC on data 3: Ra wData</a>	  
<a href="#">11: FastQC on data 3: We bpage</a>	  
<a href="#">10: FastQC on data 2: Ra wData</a>	  
<a href="#">9: FastQC on data 2: We bpage</a>	  
<a href="#">8: FastQC on data 1: Raw Data</a>	  
<a href="#">7: FastQC on data 1: Web page</a>	  
<a href="#">6: exome_regions.bed</a>	  
<a href="#">5: known_sites_regions.vcf</a>	  

- 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

trimmomatic

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)

Versions Options

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

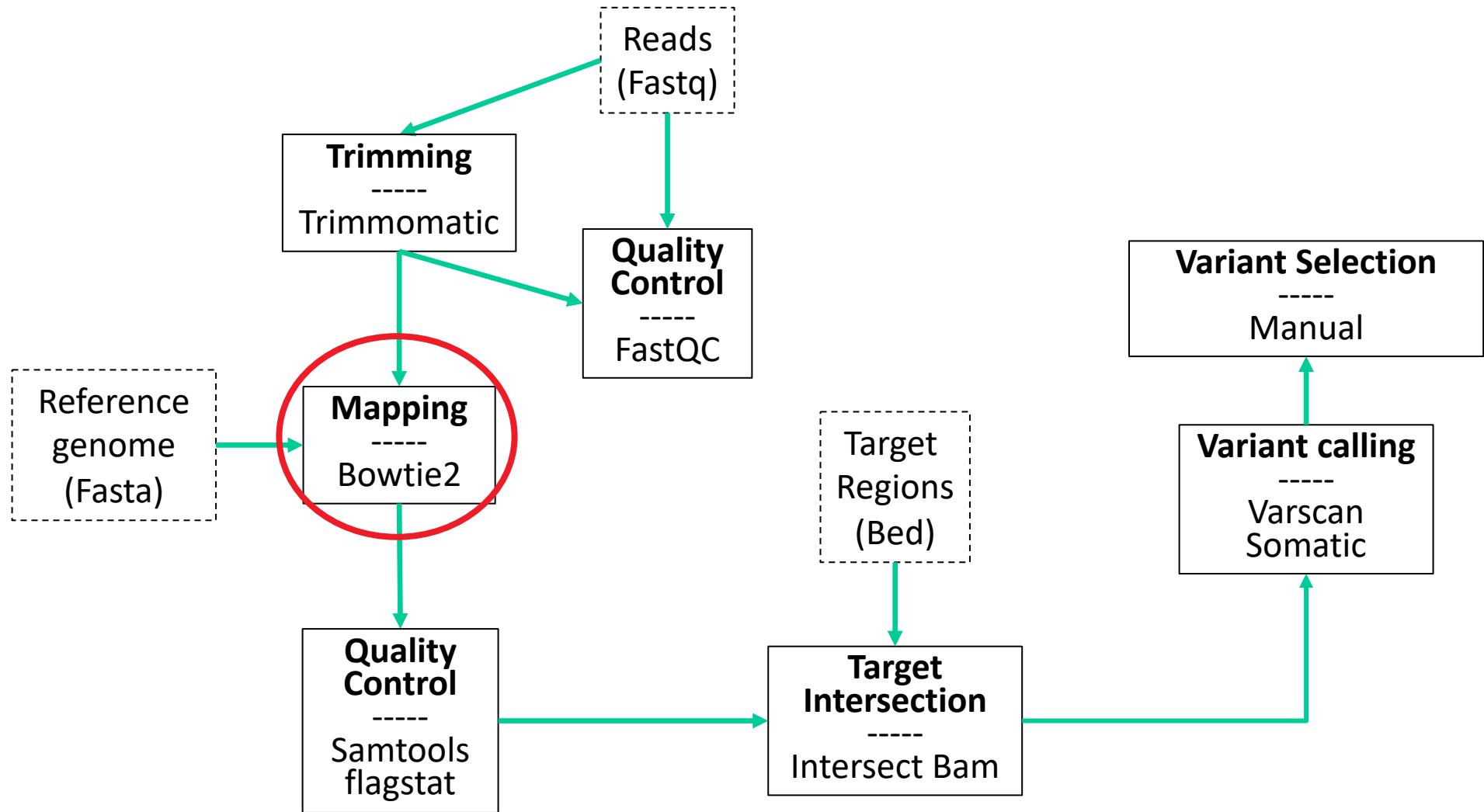
2: tumor\_R2.fastq

Vérifiez à nouveau les fichiers corrigés avec fastqc

# Trimmomatic (fin)

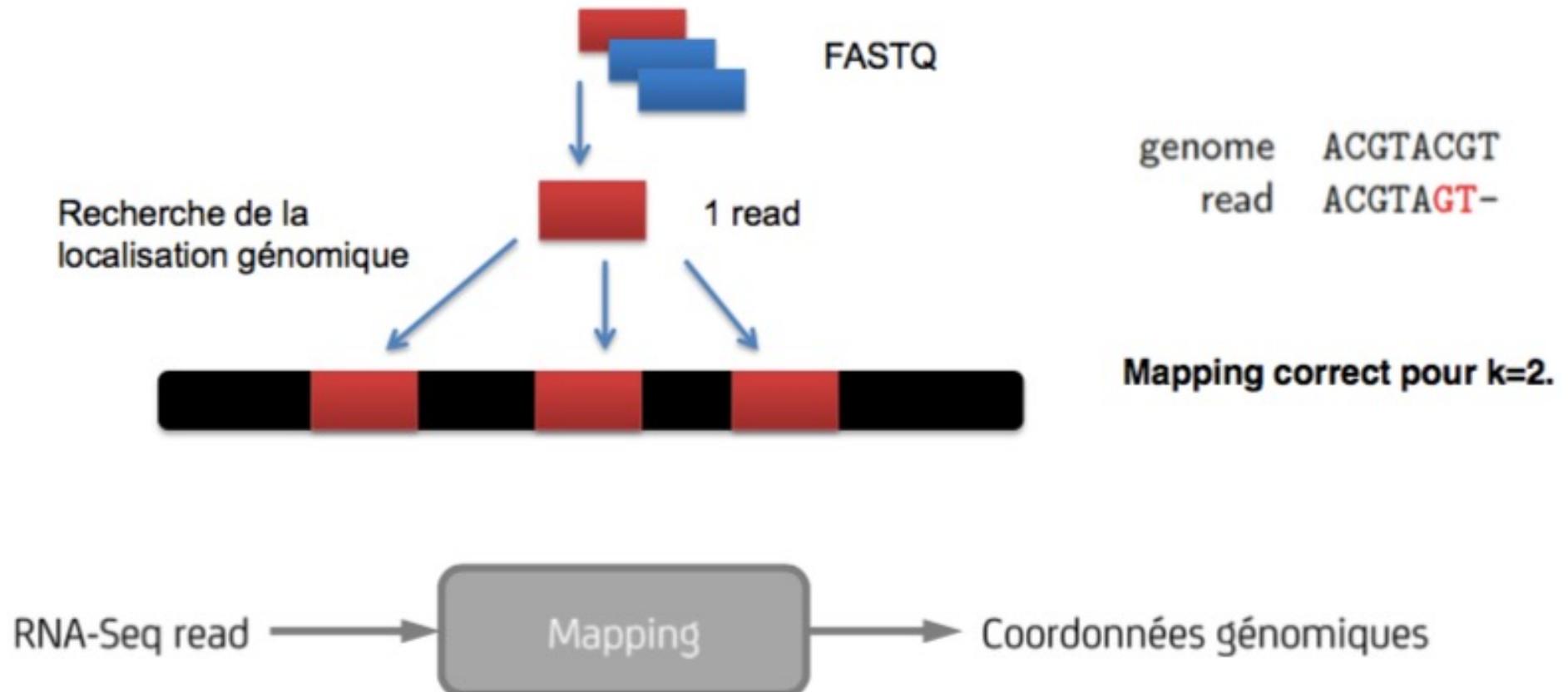
- Vérifiez le gain de qualité (faites fastqc d'un fastq)
- Eliminez les données « unpaired »

# Mapping/alignement



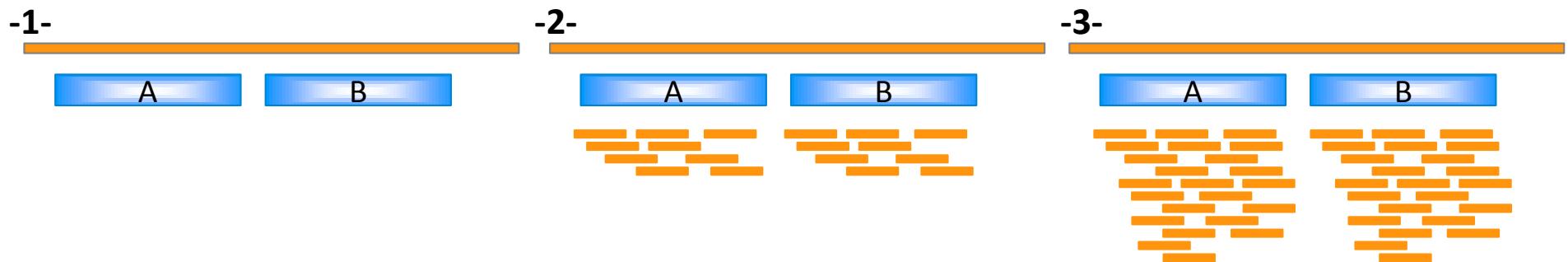
# Mapping/alignement

Mapper=trouver tous les endroits où le read est présent à k erreurs près.



# Alignment key parameters – Repeats – 3 strategies

- 1- Report only unique alignment
- 2- Report best alignments and randomly assign reads across equally good loci
- 3- Report all (best) alignments



Treangen T.J. and Salzberg S.L. 2012. Nature review Genetics 13, 36-46

# Intérêt du paired-end pour les régions répétées

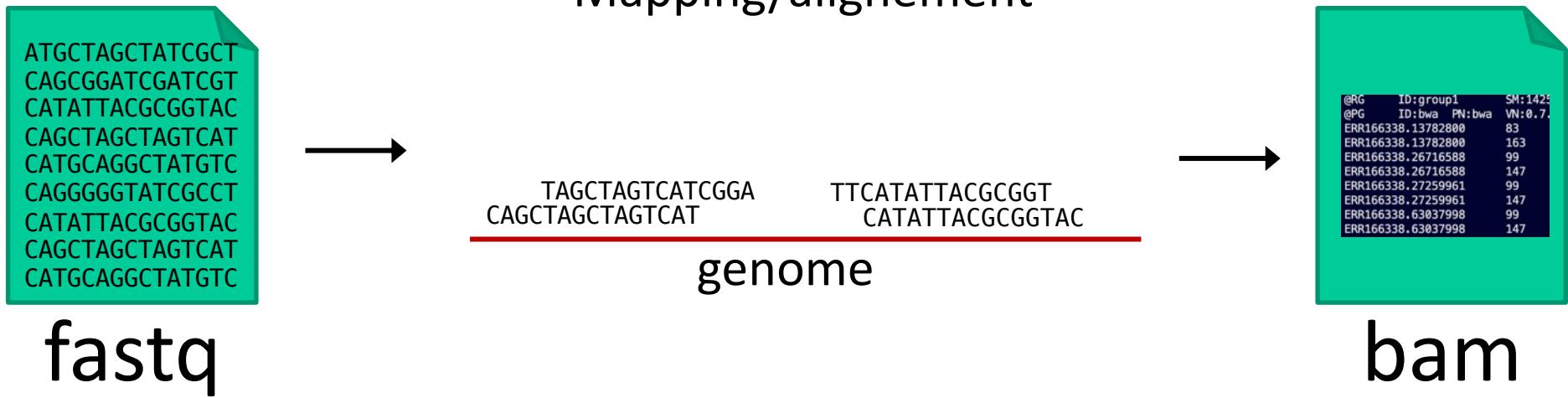
- Single-end alignment – repeated sequence



- Paired-end alignment – unique sequence

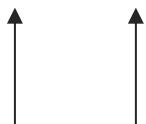


# Mapping: Fastq > BAM



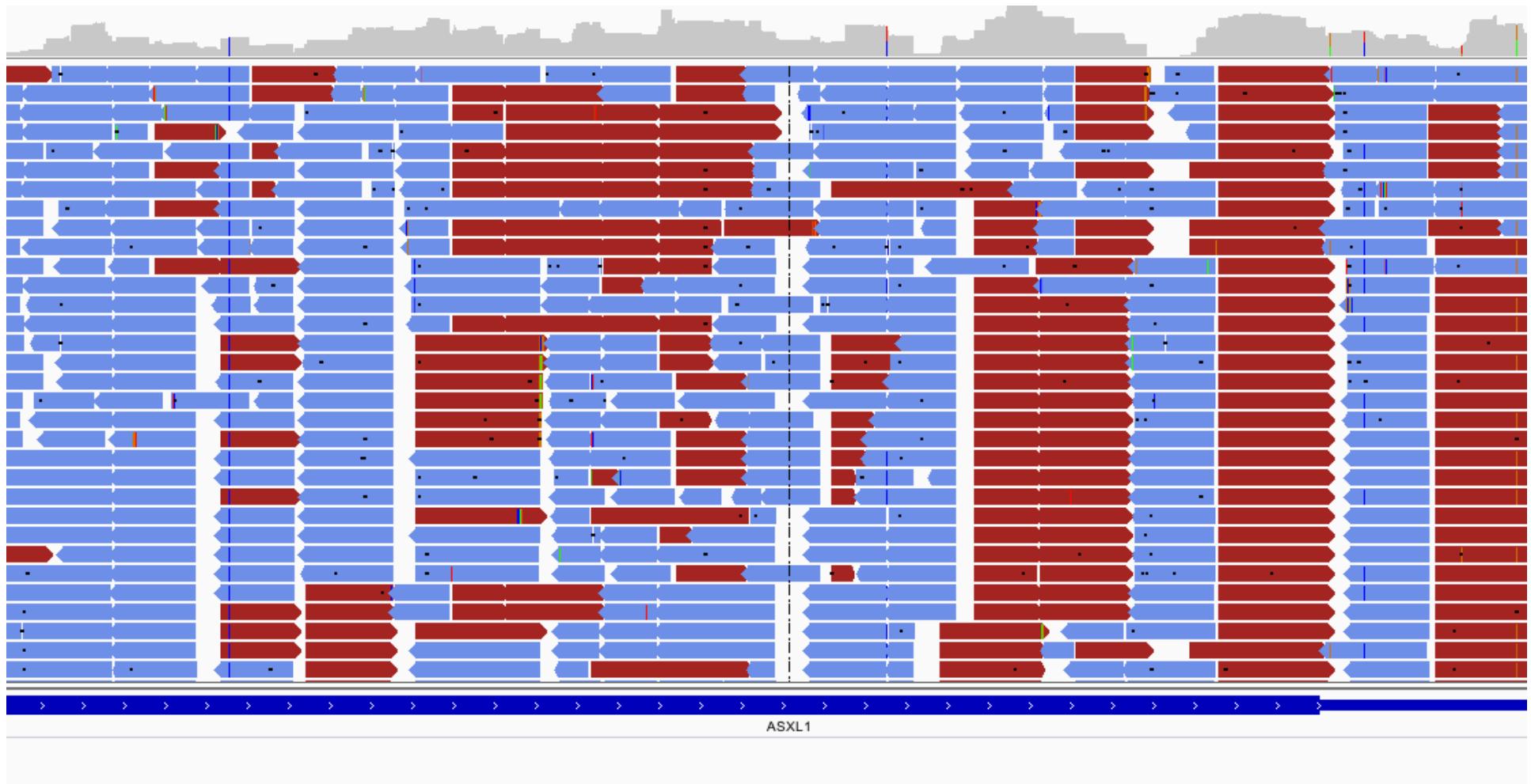
# BAM format

```
@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 GGGACTGAATTAGAACAAATTTCAGCGCTT
ERR166338.13782800 163 chr13 32890343 60 75M = 32890449 207 CACTAGCCACGTTCGAGTGCTTAATGTGGCTAGTGGC
ERR166338.26716588 99 chr13 32890406 60 101M = 32890553 222 AATGTTCCCACCTCACAGTAAGCTGTTACCGTCCAG
ERR166338.26716588 147 chr13 32890553 60 75M = 32890406 -222 TTGCAGACTTACCAAGCATTGGAGGAATATCGTA
ERR166338.27259961 99 chr13 32890496 60 101M = 32890558 137 ACCTCAGTCACATAATAAGGAATGCATCCCTGTGTAAG
ERR166338.27259961 147 chr13 32890558 60 75M = 32890496 -137 GACTTACCAAGCATTGGAGGAATATCGTAGGTA
ERR166338.63037998 99 chr13 32890496 60 101M = 32890558 137 ACCTCAGTCACATAATAAGGAATGCATCCCTGTGTAAG
ERR166338.63037998 147 chr13 32890558 60 75M = 32890496 -137 GACTTACCAAGCATTGGAGGAATATCGTAGGTA
```



Position du  
début du read

# Reads alignés: le format BAM/SAM



# Bowtie

Galaxy / Europe      Analyse de données   Workflow   Visualize ▾   Données partagées ▾   Aide ▾   Utilisateur ▾   Grid

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)**

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

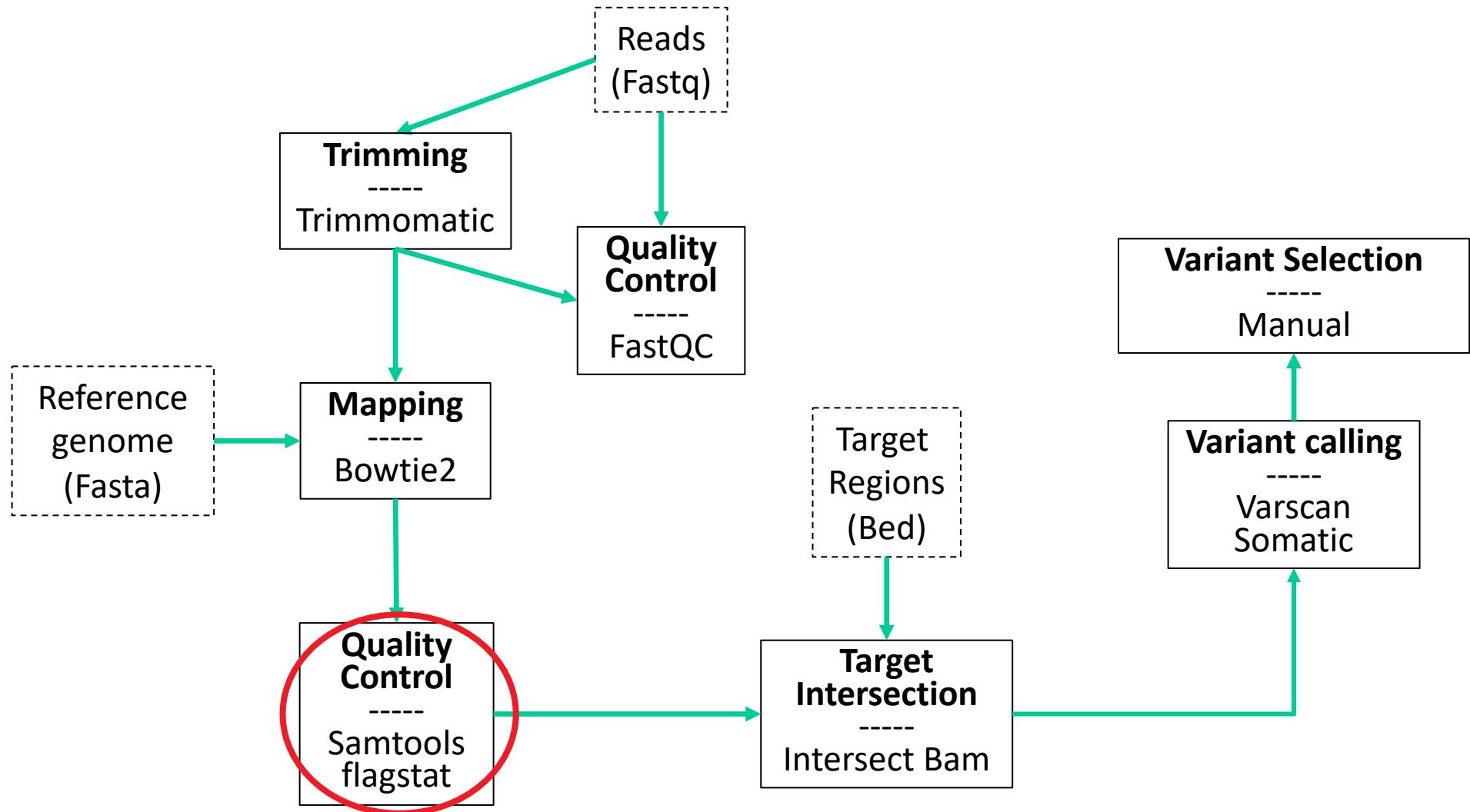
41: Samtools sort BWA normal

40: Samtools sort Bowtie TUMOR

39: Samtools sort Bowtie NORMAL

Check Bowtie result: what type of file is it?

# Contrôle qualité sur alignement



# 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)

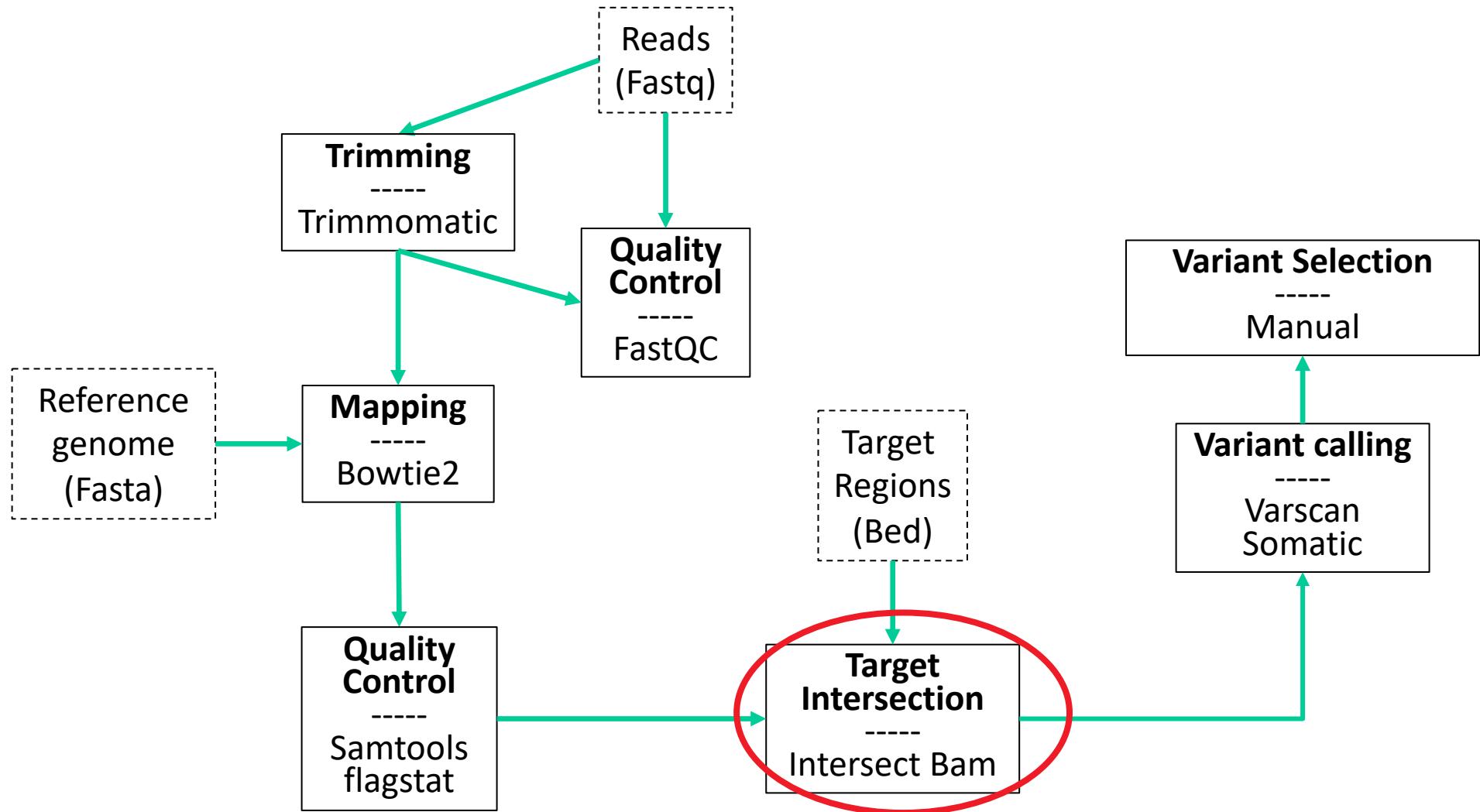
BAM file

21: Bowtie2 on data 14 and data 13: alignments

resultat

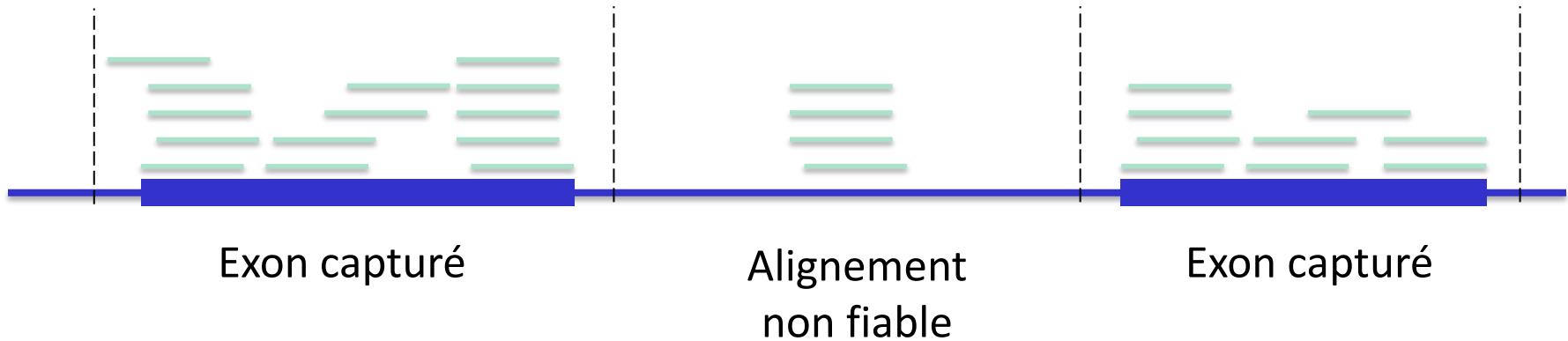
```
# This file was produced by samtools stats (1.13+htslib-1.13) and can be plotted using plot-bamstats
# This file contains statistics for all reads.
# The command line was: stats -@ 0 infile
# CHK, Checksum [2]Read Names [3]Sequences [4]Qualities
# CHK, CRC32 of reads which passed filtering followed 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 secondary
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 + both ends
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



# Target intersection

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



# Bedtools intersect intervals

The screenshot shows the Galaxy Europe interface with the 'bedtools intersect' tool selected. The tool configuration is as follows:

- File A to intersect with B:** 39: Samtools sort Bowtie NORMAL (BAM/bed,bedgraph,gff,vcf format)
- File(s) B to intersect with A:** 6: exome\_regions.bed (BAM/bed,bedgraph,gff,vcf format)
- Combined or separate output files:** One output file per 'input B' file
- 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
- Required overlap:** Default: 1bp
- Report only those alignments that \*\*do not\*\* overlap with file(s) B:** Yes

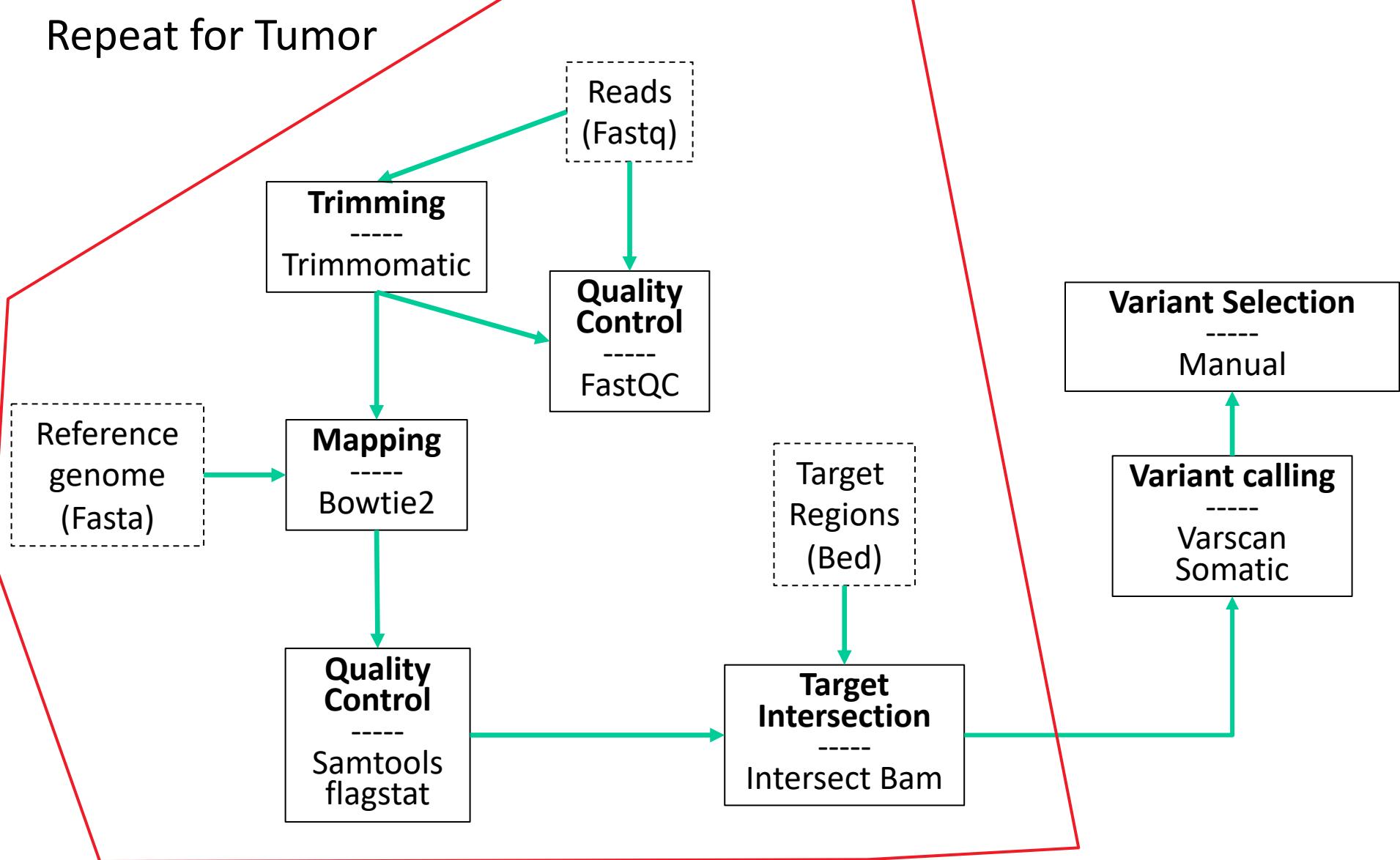
The right side of the interface shows a history of previous analyses, including:

- 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 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

Vérifiez la réduction de taille du fichier BAM

# Répéter le workflow

Repeat for Tumor



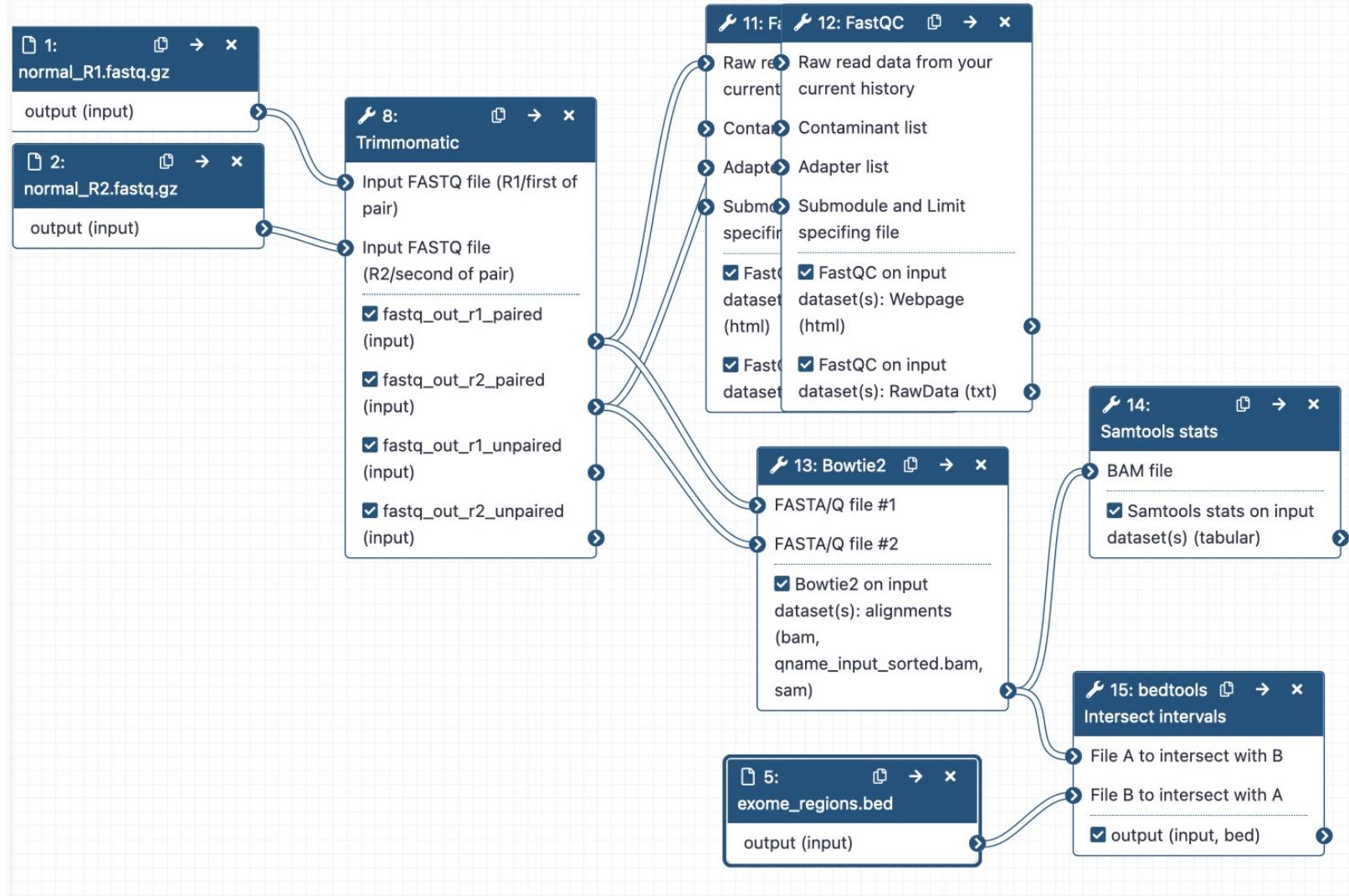
# Extraire un workflow

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 » plutot que « normal »)
- Puis save workflow

## align\_and\_check\_1\_sample



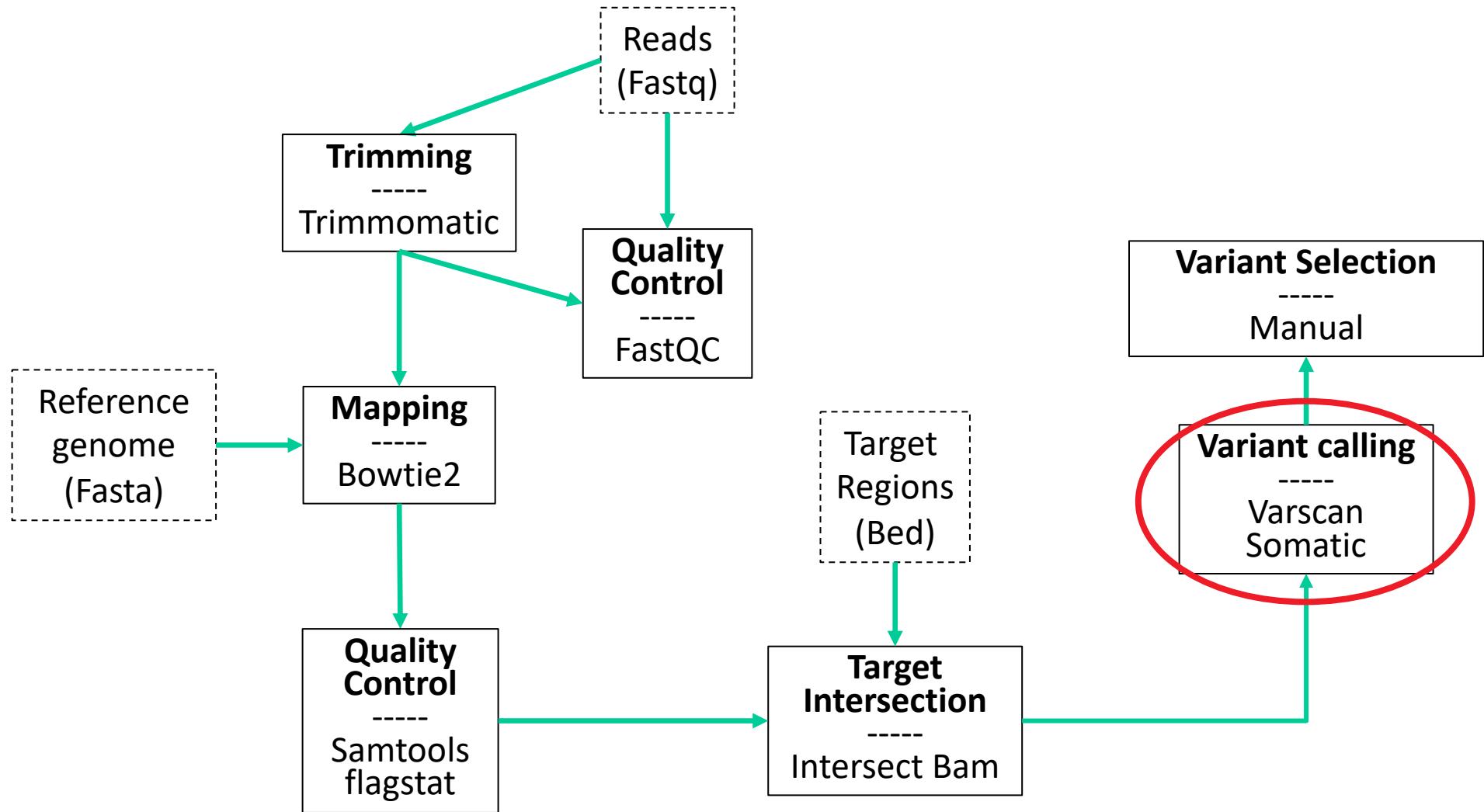
# Maintenant lancez le workflow sur les données Tumor (run workflow)

The screenshot shows the Galaxy web interface with the following details:

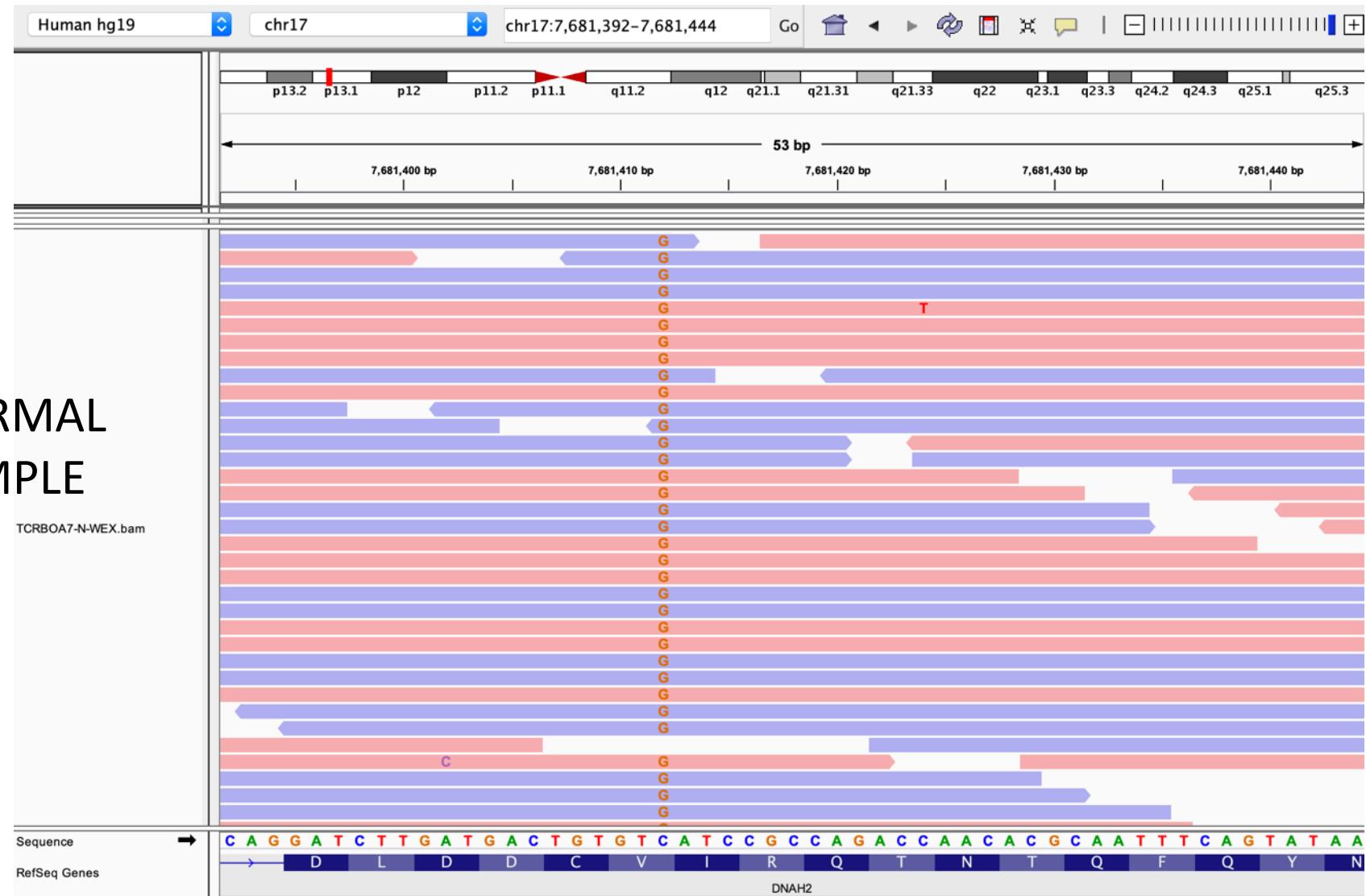
- Header:** Analyse de données, Workflow, Visualize, Données partagées, Aide, Utilisateur.
- Left Sidebar (META TOOLS):** tools, formats, Operations, TEXT TOOLS, Manipulation, Sort, Extract and Group, S, NGS Features, Enrichments, Genomic Intervals, FASTQ manipulation, Alignments, FASTQ manipulation, Control, Calling, Editing, Tools.
- Workflow Title:** Workflow: OneSample
- Run Workflow Button:** ✓ Run workflow
- History Options:** Send results to a new history (Yes or No).
- Workflow Steps:**
  - 1: Fastq R1 (Input: 1: tumor\_R1.fastq)
  - 2: Fastq R2 (Input: 2: tumor\_R2.fastq)
  - 3: exome\_regions.bed (Input: 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)
- Right Sidebar (History):** Pipeline1, 25 shown, 73 deleted, 151.47 MB, displaying log entries related to fastq files and trimming.

Three red arrows point to the input fields for steps 1, 2, and 3, indicating where to select the tumor fastq and bed files.

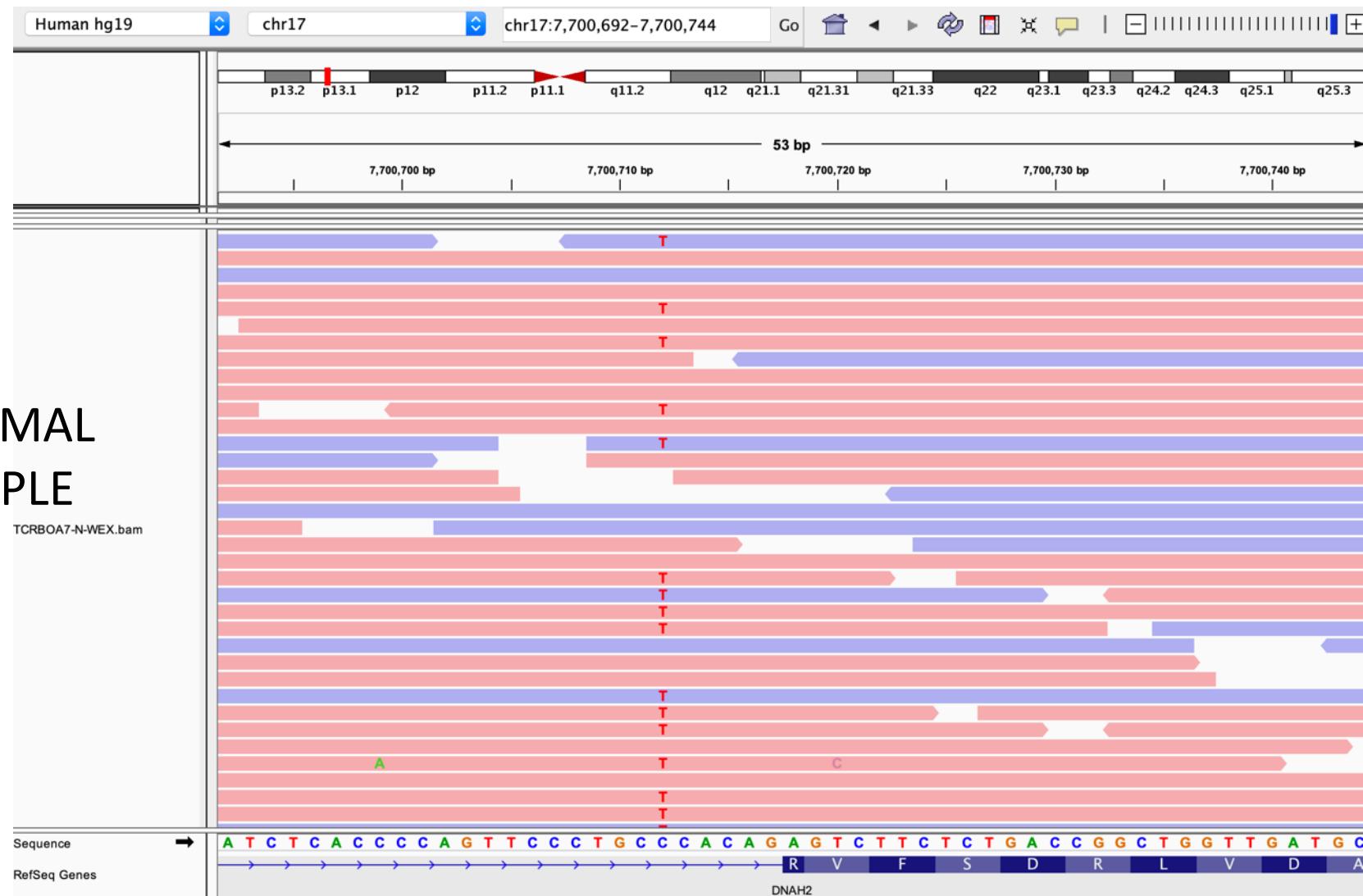
# Variant calling



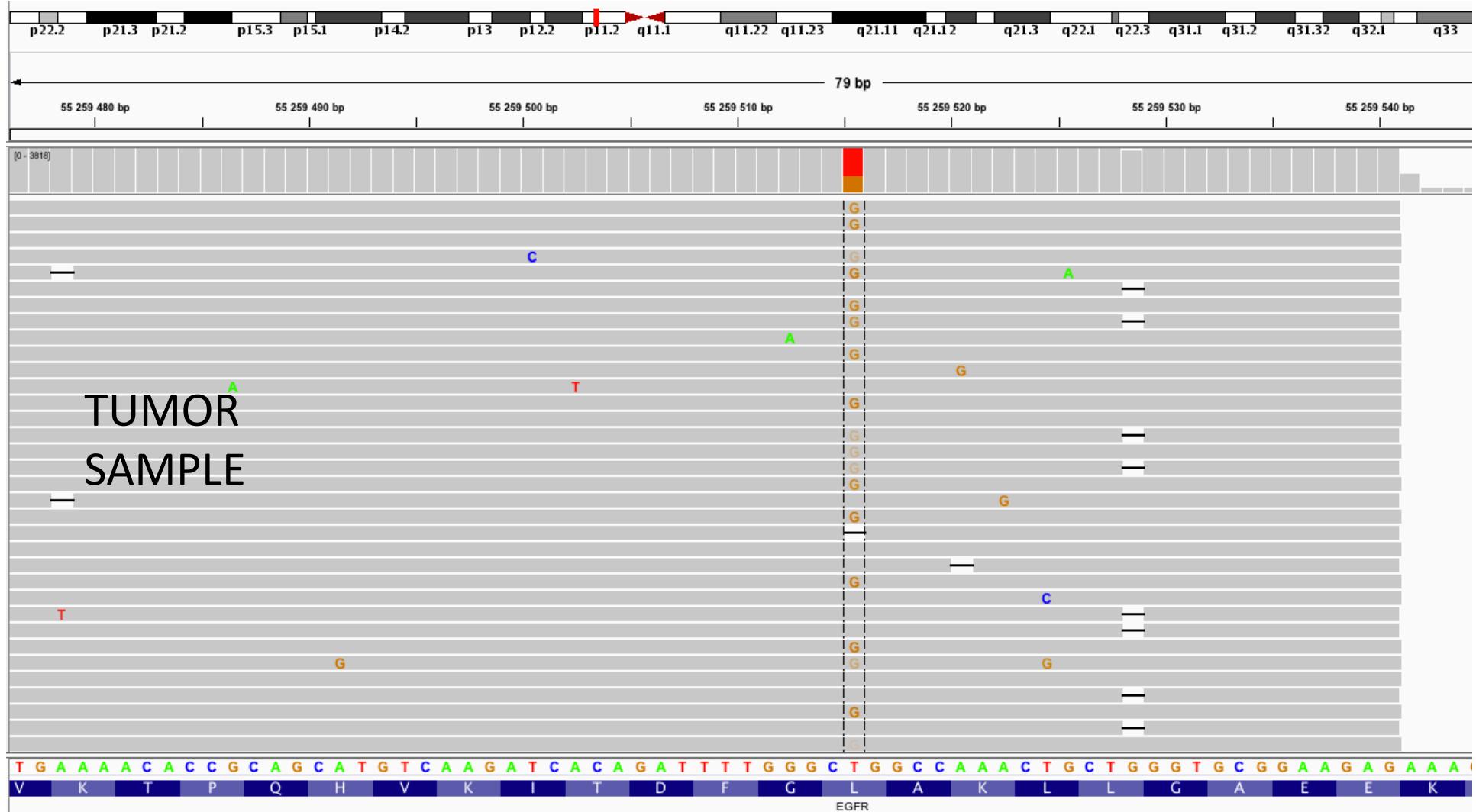
# Un polymorphisme (SNP) homozygote



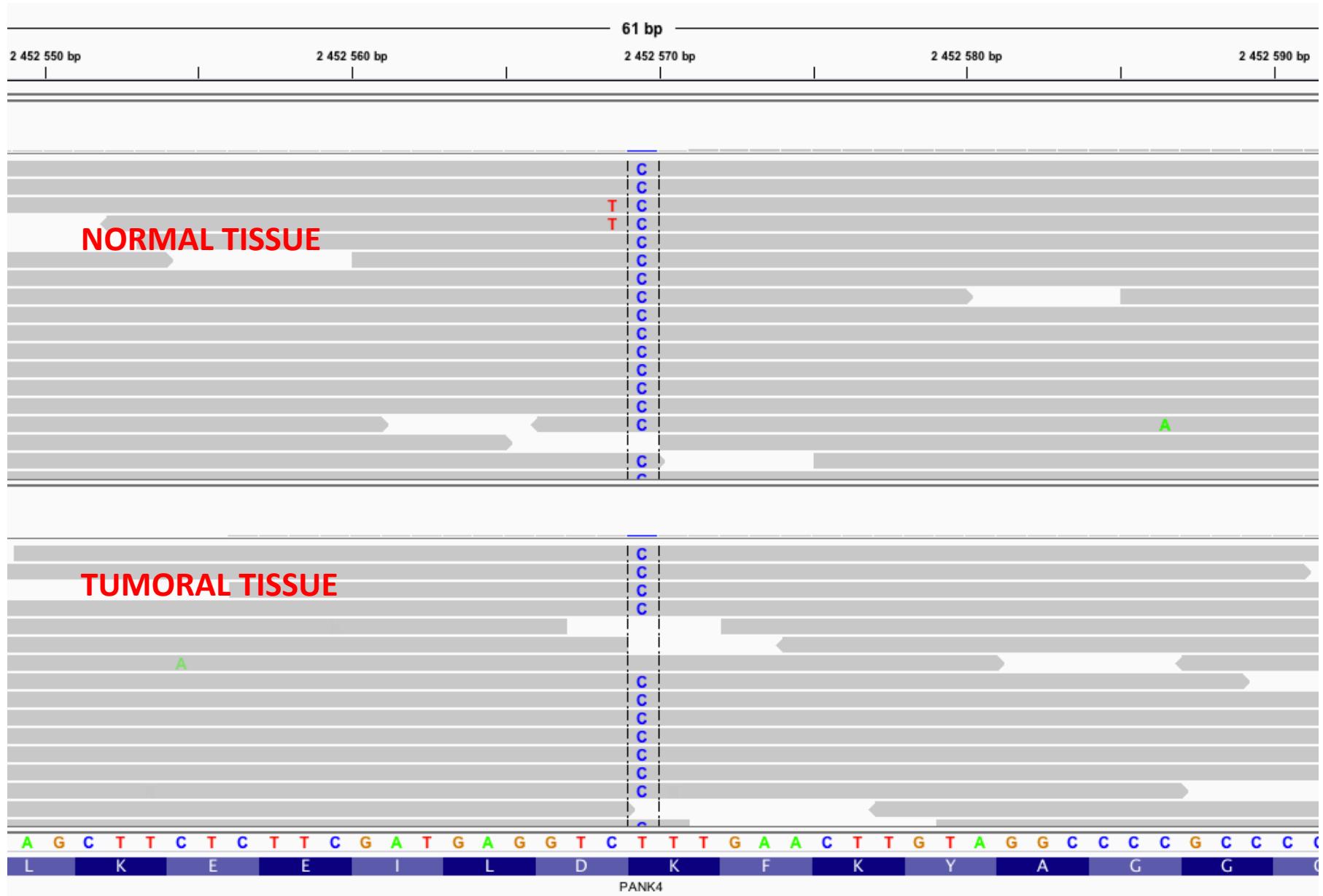
# Un polymorphisme (SNP) hétérozygote



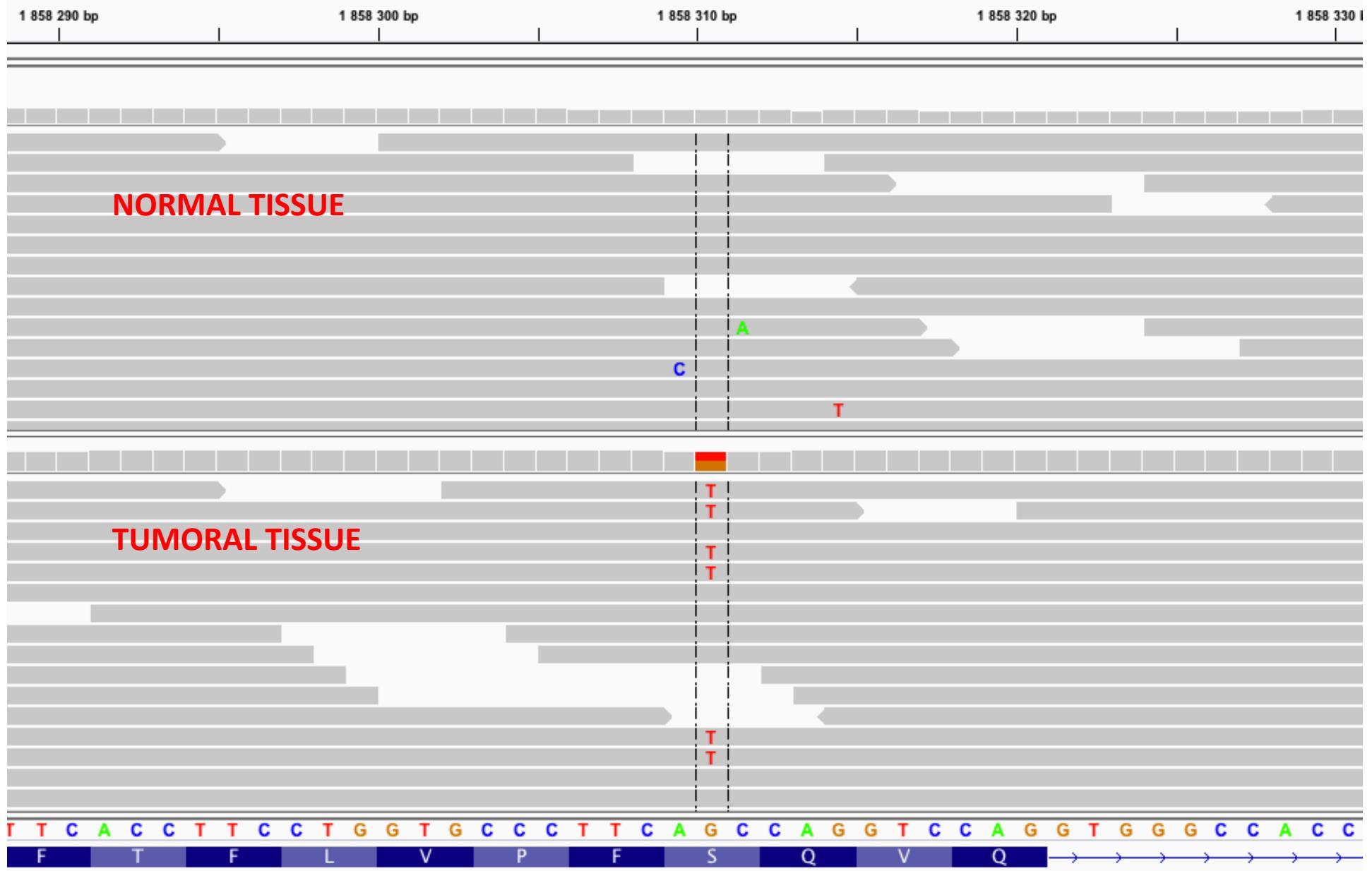
# Un variant tumoral: polymorphisme ou mutation?



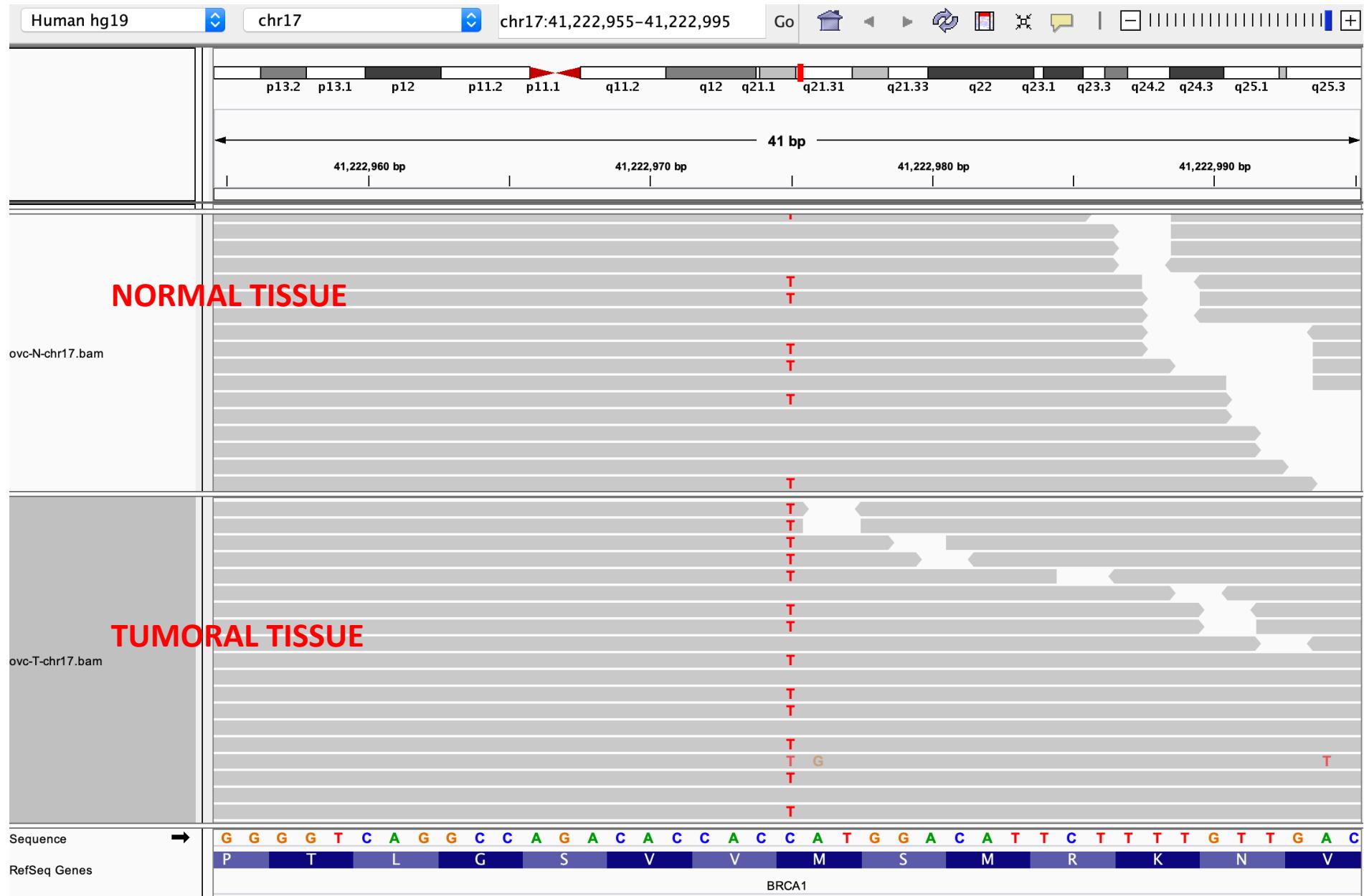
# Un polymorphisme vu dans N et T



# Une mutation somatique

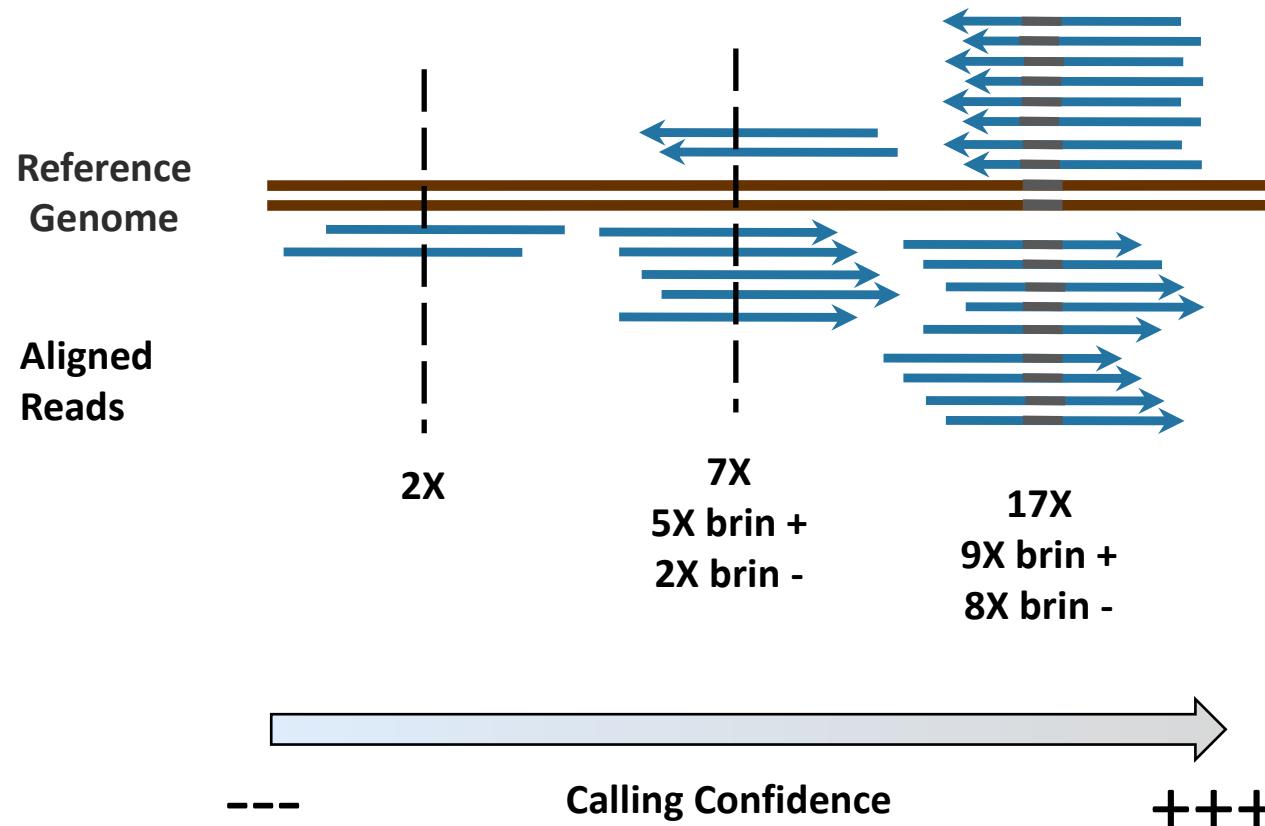


# Une LOH (loss of heterozygosity)



# Variant Calling Quality

Depth of Coverage = number of reads supporting one position ex: 1X, 5X, 100X... >1000X





- Mutation caller written in **Java** (portable)
- Somatic Mode: input=**Tumor/Normal pairs**:
  - Somatic, germline, LOH events
  - Somatic copy number alterations (CNAs)

# Varscan's Somatic P-value

## Variant Calling and Comparison

At every position where both normal and tumor have sufficient coverage, a comparison is made. First, normal and tumor are called independently using the germline consensus calling functionality. Then, their genotypes are compared by the following algorithm:

Calculate significance of allele frequency difference by Fisher's Exact Test

**If difference is significant (p-value < threshold):**

If normal matches reference

==> Call Somatic

Else If normal is heterozygous

==> Call LOH

Else normal and tumor are variant, but different

==> Call IndelFilter or Unknown

**If difference is not significant:**

==> Call Germline

The diagram illustrates the logic for determining if a variant is somatic or LOH based on allele counts in normal (N) and tumor (T) samples.

**Alleles**

	Ref	Var
N	8	0
T	6	7

A red arrow points from this table to the text "Somatic".

**LOH**

	Ref	Var
N	4	4
T	8	1

A red arrow points from this table to the text "LOH".

# VarScan Native Format (default Varscan output)

	chrom	position	ref	var	normal_reads1	normal_reads2	normal_var_freq	normal_gt	tumor_reads1
1	chr16	64458	G	A	8	38	82.61%	A	
2	chr16	64480	G	C	32	8	20%	S	
3	chr16	64735	C	T	22	12	35.29%	Y	
4	chr16	65561	G	C	4	13	76.47%	C	
5	chr16	66888	A	G	10	9	47.37%	R	
6	chr16	67135	T	C	37	14	27.45%	Y	
7	chr16	67254	A	C	6	12	66.67%	M	
8	chr16	67832	A	G	12	1	7.69%	A	
9	chr16	79609	C	G	24	14	36.84%	S	
10	chr16	79962	C	G	14	7	33.33%	S	
11	chr16	81639	C	A	11	15	57.69%	M	
12	chr16	81779	G	T	106	32	23.19%	K	

**Cons** : Consensus Genotype of Variant Called (IUPAC code):

M -> A or C	Y -> C or T	D -> A or G or T	W -> A or T	V -> A or C or G
R -> A or G	K -> G or T	B -> C or G or T	S -> C or G	H -> A or C or T

# Fields in Varscan "native format"

Native Output Field	VCF Field	Description
chrom	CHROM	Chromosome or reference name
position	POS	Position from pileup (1-based)
ref	REF	Reference base at this position
var	ALT	Variant base seen in tumor
normal_reads1	RD (col 10)	Reads supporting reference in normal
normal_reads2	AD (col 10)	Reads supporting variant in normal
normal_var_freq	FREQ (col 10)	Variant allele frequency in normal
normal_gt	GT (col 10)	Consensus genotype call in normal
tumor_reads1	RD (col 11)	Reads supporting reference in tumor
tumor_reads2	AD (col 11)	Reads supporting variant in tumor
tumor_var_freq	FREQ (col 11)	Variant allele frequency in tumor
tumor_gt	GT (col 11)	Consensus genotype call in tumor
somatic_status	SS (col 8)	Somatic status call (Germline, Somatic, LOH, or Unknown)
variant_p_value	GPV (col 8)	Variant p-value for Germline events
somatic_p_value	SPV (col 8)	Somatic p-value for Somatic/LOH events
tumor_reads1_plus	DP4 (col 11)	Tumor reference-supporting reads on + strand
tumor_reads1_minus	DP4 (col 11)	Tumor reference-supporting reads on - strand
tumor_reads2_plus	DP4 (col 11)	Tumor variant-supporting reads on + strand
tumor_reads2_minus	DP4 (col 11)	Tumor variant-supporting reads on - strand
normal_reads1_plus	DP4 (col 10)	Normal reference-supporting reads on + strand
normal_reads1_minus	DP4 (col 10)	Normal reference-supporting reads on - strand
normal_reads2_plus	DP4 (col 10)	Normal variant-supporting reads on + strand
normal_reads2_minus	DP4 (col 10)	Normal variant-supporting reads on - strand

# Somatic variant calling: Varscan

Attention: étape de 10-30min!

The screenshot shows the Galaxy web interface with the VarScan somatic Call tool selected. The tool configuration window is open, displaying the following settings:

- reference genome:** Human (Homo sapiens): hg19
- 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
- Settings for Variant Calling:** Use default values
- Settings for Posterior Variant Filtering:** Do not perform posterior filtering

The right side of the interface shows a history panel with the following entries:

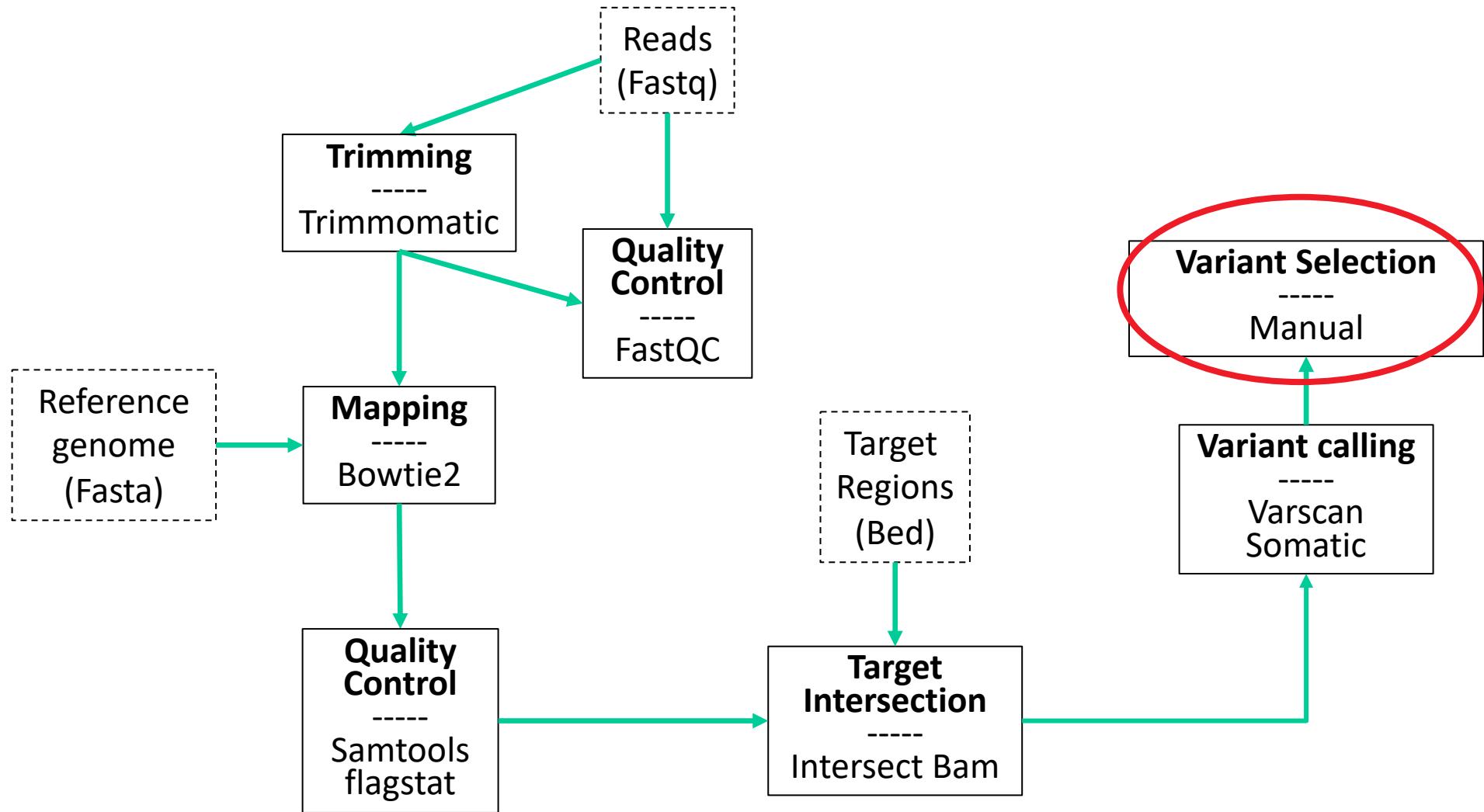
- exome test 2 (29 shown, 26 deleted, 170.08 MB)
- 55: Samtools flagstat on data 53
- 54: TUMOR BAM intersect (5.5 MB, format: bam, génome de référence: hg19)
- display at UCSC main, display at Ensembl Current, display with IGV local Human hg19, display in IGB View
- Binary bam alignments file
- 53: 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	chr17 18874685 . C CGGT . PASS DP=32;SS=3;SSC=16;GPV=1;SPV=0.022989;INDEL	Pipeline1
<a href="#">Get Data</a>	chr17 18874720 . C G . PASS DP=33;SS=1;SSC=0;GPV=1.3852e-19;SPV=1	24 shown, 72 deleted
<a href="#">Send Data</a>	chr17 18882991 . T A . PASS DP=60;SS=1;SSC=0;GPV=1.035e-35;SPV=1	151.47 MB
<a href="#">Convert Formats</a>	chr17 41256074 . C CA . PASS DP=81;SS=1;SSC=1;GPV=0.0015196;SPV=0.63343;INDEL	  
<a href="#">Collection Operations</a>	chr17 73759304 . G T . PASS DP=36;SS=1;SSC=0;GPV=2.2598e-21;SPV=1	<u>88: VarScan somatic calls</u> <u>data 82 and data 80</u>
GENERAL TEXT TOOLS	chr19 6374813 . T C . PASS DP=33;SS=1;SSC=0;GPV=2.8029e-05;SPV=0.8425	153 lines, 113 comments
<a href="#">Text Manipulation</a>	chr19 7550844 . G A . PASS DP=44;SS=1;SSC=4;GPV=2.3358e-10;SPV=0.35332	format: vcf, génome de référence:
<a href="#">Filter and Sort</a>	chr19 36504365 . C T . PASS DP=34;SS=1;SSC=1;GPV=5.1914e-07;SPV=0.63966	hg19
<a href="#">Join, Subtract and Group</a>	chr1 10596341 . C T . PASS DP=44;SS=1;SSC=2;GPV=7.4746e-10;SPV=0.53262	Starting variant calling ..
GENOMICS, NGS	chr1 160251792 . A G . PASS DP=37;SS=1;SSC=0;GPV=5.1339e-06;SPV=0.87856	Calling variants for contig: chr10
<a href="#">Extract Features</a>	chr1 167082869 . G A . PASS DP=71;SS=1;SSC=8;GPV=2.0173e-19;SPV=0.13252	Contig chr10 finished.
<a href="#">BED Tools</a>	chr1 167095163 . G C . PASS DP=52;SS=1;SSC=5;GPV=6.8522e-13;SPV=0.28624	Calling variants for contig: chr11
<a href="#">Fetch Alignments</a>	chr1 167097739 . C A . PASS DP=64;SS=1;SSC=3;GPV=4.3049e-14;SPV=0.44587	Contig chr11 finished.
<a href="#">Operate on Genomic Intervals</a>	chr1 214788427 . C T . PASS DP=45;SS=1;SSC=1;GPV=8.5784e-10;SPV=0.66234	Calling variants for contig: chr12
<a href="#">FASTA/FASTQ manipulation</a>	chr1 214802553 . CT C . PASS DP=83;SOMATIC;SS=2;SSC=18;GPV=1;SPV=0.015148;INDEL	Contig chr12 finish
<a href="#">Multiple Alignments</a>	chr1 214803969 . G C . PASS DP=111;SOMATIC;SS=2;SSC=35;GPV=1;SPV=0.00029013	    
<a href="#">FASTA/FASTQ manipulation</a>	chr1 214804041 . C A . PASS DP=65;SS=1;SSC=0;GPV=2.7963e-08;SPV=0.9934	display at UCSC main
<a href="#">Picard</a>	chr1 214811174 . G A . PASS DP=76;SS=1;SSC=0;GPV=3.6183e-12;SPV=0.99124	display with IGV local Human hg19
<a href="#">Quality Control</a>	chr1 214811244 . C G . PASS DP=120;SS=1;SSC=0;GPV=1.7875e-19;SPV=0.92629	display at RVViewer main
<a href="#">Assembly</a>	chr1 214813487 . A G . PASS DP=291;SS=1;SSC=3;GPV=1.3526e-38;SPV=0.47444	
<a href="#">Mapping</a>	chr1 214813782 . A G . PASS DP=108;SS=1;SSC=0;GPV=1.7692e-19;SPV=0.98472	
<a href="#">Variant Calling</a>	chr1 214813941 . C G . PASS DP=86;SS=1;SSC=4;GPV=8.038e-16;SPV=0.34707	
<a href="#">Genome editing</a>	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	

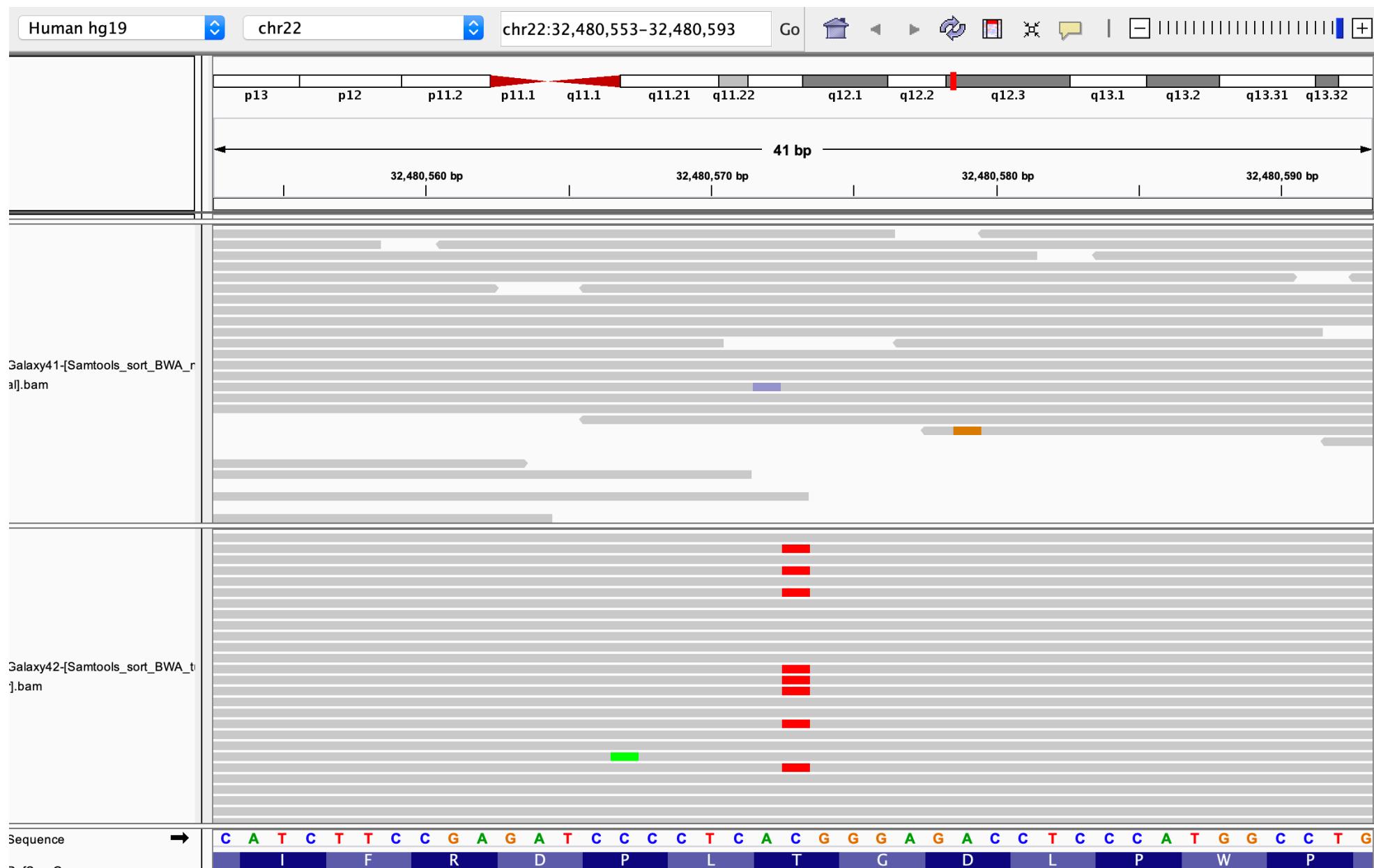
- Vous pouvez utiliser la fonction « grep » pour filter les lignes avec somatic ou LOH
- Vérifiez la somatic P-value (SPV), les comptages
- Regardez les sous IGV



# 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 » or « track / local file »)
  - In Galaxy, click on « display with IGV local ». (will automatically connect with your local IGV session)
- Visualize somatic events (next slide)

# IGV view



# IGV web

- Lancez IGV à partir du site « IGV web »
- Choisissez hg19 comme génome de référence (si ce n'est pas déjà le cas)
- Zoomez sur une région chromosomique au hasard. Voyez les annotations de genes (exons, introns, isoformes)
- Chargez les 2 fichiers bam Normal et Tumor (track menu / local file) (sur IGV web: selectionner les BAM et les BAI en même temps)
- Recherchez des événements somatiques vus dans votre sortie Varscan

# Annexes

# 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

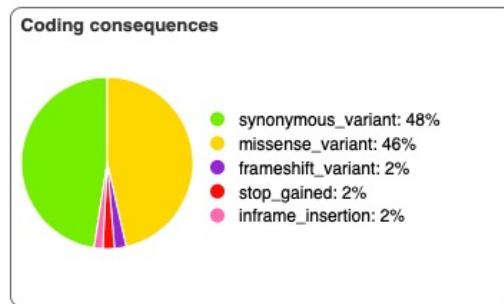
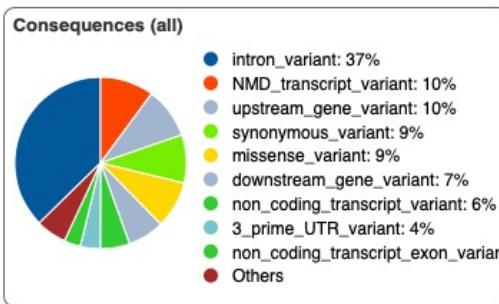
Note: the highest impact variants are not necessarily somatic!

## 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)
 Filters
 Download

Show: [1](#) [5](#) [10](#) [50](#) [All](#) variants
Uploaded variant  
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 position
.	<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="#">ENSG0000173230</a>	Transcript	<a href="#">ENST00000340645</a>	protein_coding	13/22	-	3173
.	<a href="#">3:121416308-121416308</a>	T	stop_gained	HIGH	GOLGB1	<a href="#">ENSG0000173230</a>	Transcript	<a href="#">ENST00000393667</a>	protein_coding	13/22	-	3173
.	<a href="#">3:121416308-121416308</a>	T	stop_gained	HIGH	GOLGB1	<a href="#">ENSG0000173230</a>	Transcript	<a href="#">ENST00000489400</a>	protein_coding	9/9	-	2659

# 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 "+"

The screenshot shows a user interface for managing genomic data. On the left, a modal window titled "About this History" is open, featuring a "Switch to this history" button and a prominent "+" button with a red arrow pointing to it. On the right, a list of files is displayed in a table format:

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			

A text overlay at the bottom right reads "Fera apparaître:".

# Samtools mpileup sur fichier intersect bed

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