

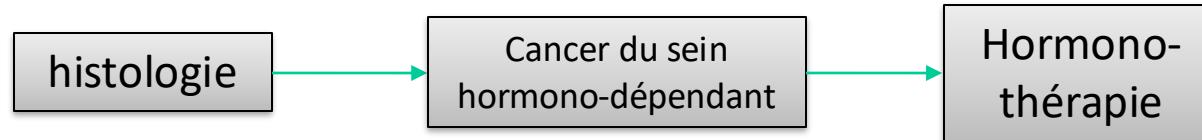
Réalisation d'un Pipeline d'analyse d'exome

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

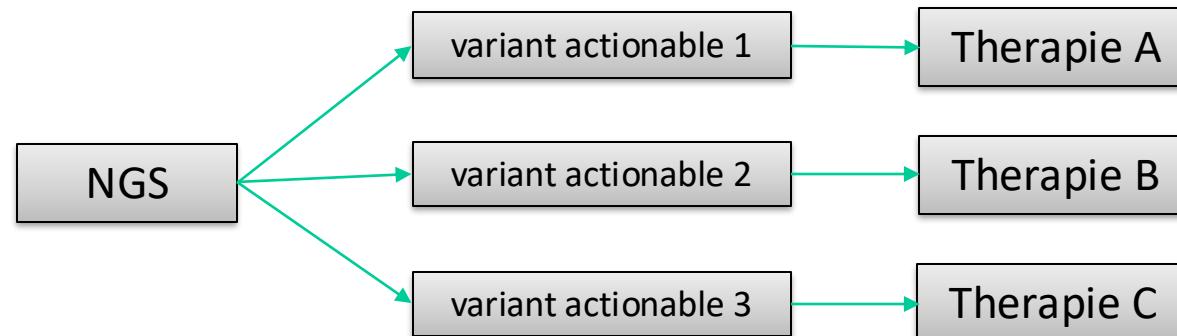
Pour quoi faire?

Rappel: Thérapie systémique vs. de précision

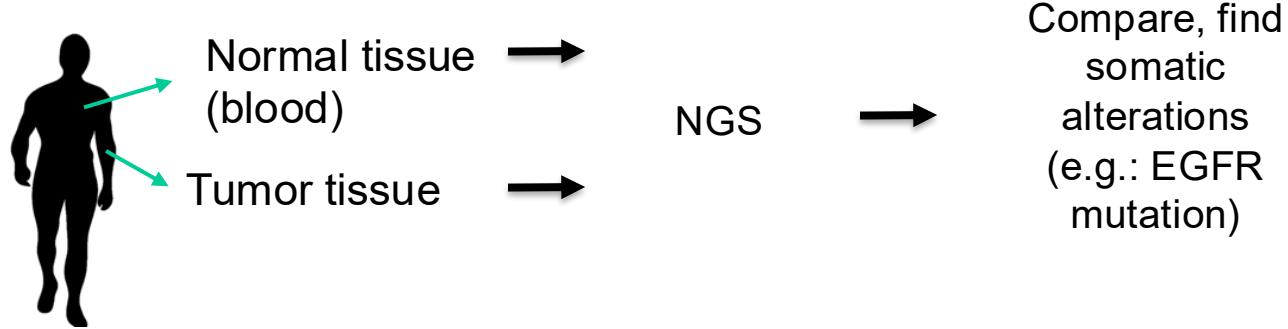
Chimio-thérapie
Systémique



Chimio-thérapie
de précision

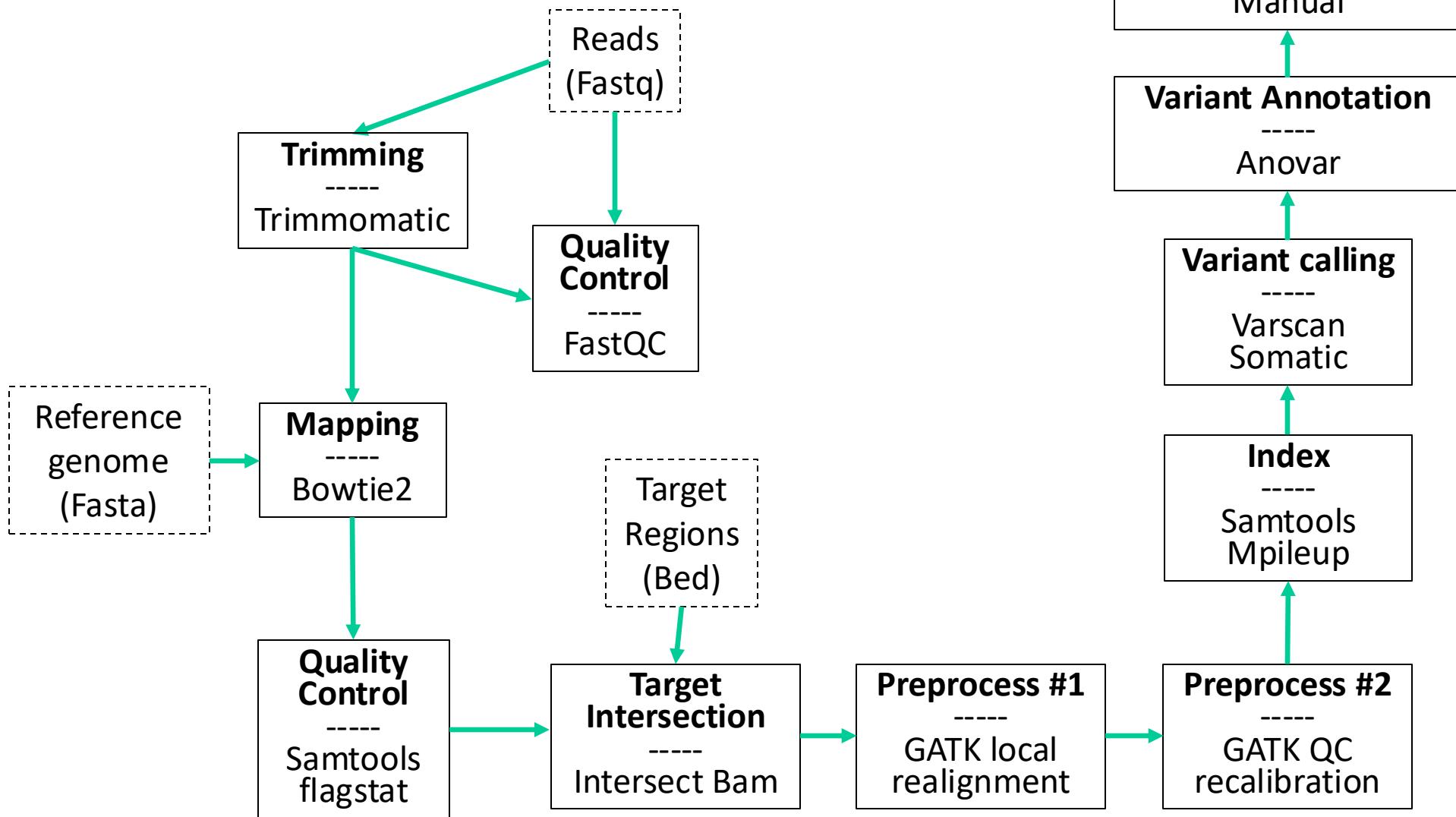


From sample to DNA alteration

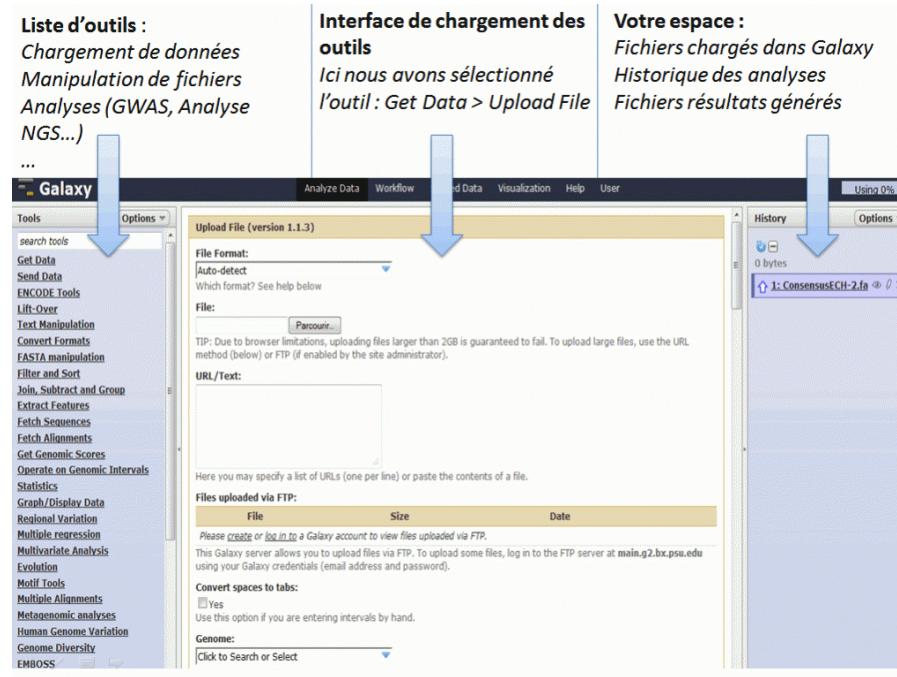


Déployer un pipeline Bioinformatique

Un pipeline « NGS Variant » utilisé à Gustave Roussy



Galaxy: user-friendly interface to NGS pipelines



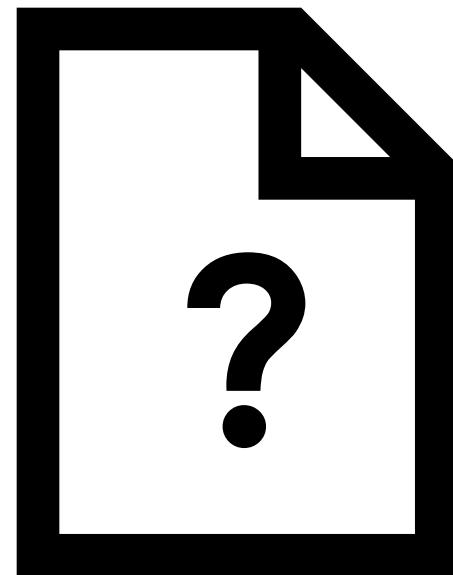
Credit: Biorigami

- 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

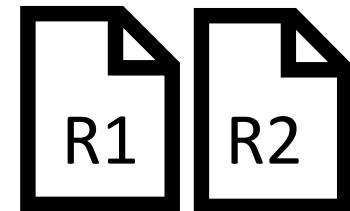


Paired end sequencing



DNA fragment
(typically 250nt)

2 fastq files



FASTQ Format

Fastq format quality

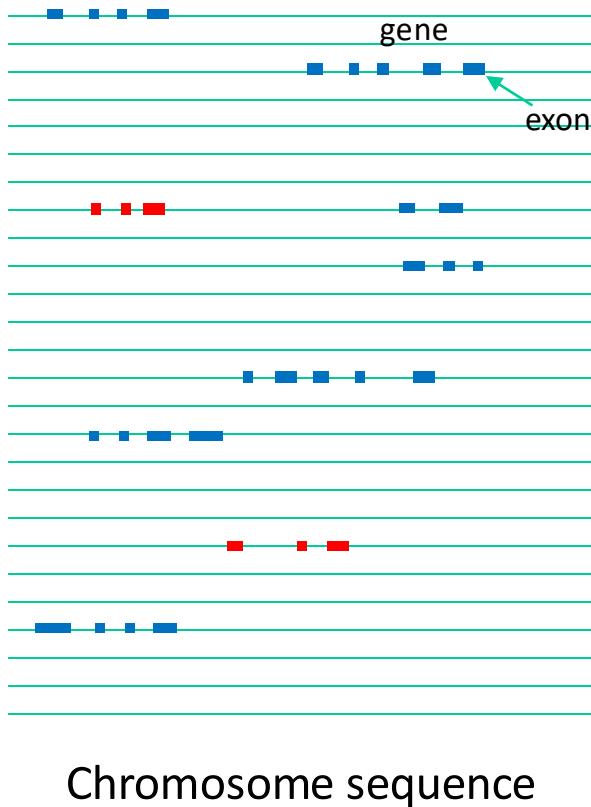
@SEQ_ID1
CCAATGCT
+
8ASR/2@B

@SEQ_ID1
CCAATGCT
+
8ASR/2@B

16, 25, 43, 42, 7, ...

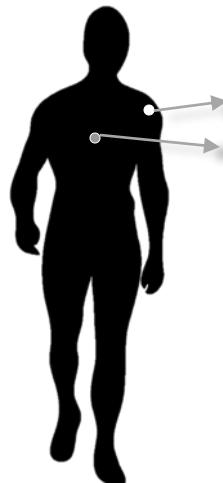
ASCII Code												
0	32	64	96	128	160	192	224	256	288	320	352	384
1	33 !	65 A	97 a	129 ü	161 í	193 Ł	225 Ø	257 Ø	289 Ø	321 Ø	353 Ø	385 Ø
2	34 "	66 B	98 b	130 é	162 ö	194 T	226 Ø	258 Ø	290 Ø	322 Ø	354 Ø	386 Ø
3	35 #	67 C	99 c	131 á	163 ú	195 T	227 Ø	259 Ø	291 Ø	323 Ø	355 Ø	387 Ø
4	36 \$	68 D	100 d	132 à	164 ñ	196 -	228 Ø	260 Ø	292 Ø	324 Ø	356 Ø	388 Ø
5	37 %	69 E	101 e	133 à	165 Ñ	197 þ	229 Ø	261 Ø	293 Ø	325 Ø	357 Ø	389 Ø
6	38 &	70 F	102 f	134 a	166 ø	198 å	230 µ	262 µ	294 µ	326 µ	358 µ	390 µ
	39 ,	71 G	103 g	135 ç	167 ø	199 Å	231 ¶	263 ¶	295 ¶	327 ¶	359 ¶	391 ¶
	40 <	72 H	104 h	136 è	168 å	200 L	232 ¶	264 ¶	296 ¶	328 ¶	360 ¶	392 ¶
	41 >	73 I	105 i	137 è	169 ®	201 H	233 ¶	265 ¶	297 ¶	329 ¶	361 ¶	393 ¶
	42 *	74 J	106 j	138 è	170 -	202 H	234 ¶	266 ¶	298 ¶	330 ¶	362 ¶	394 ¶
11	6	43 +	75 K	107 k	139 ï	171 %	203 ¶	235 ¶	267 ¶	299 ¶	332 ¶	364 ¶
12	♀	44 -	76 L	108 l	140 í	172 %	204 ¶	236 ¶	268 ¶	300 ¶	333 ¶	365 ¶
13	45 :	77 M	109 m	141 í	173 %	205 ¶	237 ¶	269 ¶	301 ¶	334 ¶	366 ¶	395 ¶
14	♂	46 .	78 N	110 n	142 à	174 «	206 ¶	238 ¶	270 ¶	302 ¶	335 ¶	367 ¶
15	*	47 /	79 O	111 o	143 å	175 »	207 ¶	239 ¶	271 ¶	303 ¶	336 ¶	368 ¶
16	▶	48 @	80 P	112 p	144 È	176 ¶	208 ¶	240 ¶	272 ¶	304 ¶	337 ¶	369 ¶
17	◀	49 1	81 Q	113 q	145 æ	177 ¶	209 ¶	241 ¶	273 ¶	305 ¶	338 ¶	370 ¶
18	‡	50 2	82 R	114 r	146 Æ	178 ¶	210 ¶	242 ¶	274 ¶	306 ¶	339 ¶	371 ¶
19	::	51 3	83 S	115 s	147 ô	179 ¶	211 ¶	243 ¶	275 ¶	307 ¶	340 ¶	372 ¶
20	¶¶	52 4	84 T	116 t	148 ö	180 ¶	212 ¶	244 ¶	276 ¶	308 ¶	341 ¶	373 ¶
21	§§	53 5	85 U	117 u	149 ö	181 å	213 ¶	245 ¶	277 ¶	309 ¶	342 ¶	374 ¶
22	-	54 6	86 V	118 v	150 û	182 å	214 ¶	246 ¶	278 ¶	310 ¶	343 ¶	375 ¶
23	55 7	87 W	119 w	151 ù	183 å	215 ¶	247 ¶	279 ¶	311 ¶	344 ¶	376 ¶	
24	↑	56 8	88 X	120 x	152 ü	184 ®	216 ¶	248 ¶	280 ¶	312 ¶	345 ¶	377 ¶
25	↓	57 9	89 Y	121 y	153 ö	185 ®	217 ¶	249 ¶	281 ¶	313 ¶	346 ¶	378 ¶
26	→	58 :	90 Z	122 z	154 Ü	186 ¶	218 ¶	250 ¶	282 ¶	314 ¶	347 ¶	379 ¶
27	←	59 ;	91 [123 ç	155 ø	187 ¶	219 ¶	251 ¶	283 ¶	315 ¶	348 ¶	380 ¶
28	„	60 <	92 \	124 :	156 £	188 ¶	220 ¶	252 ¶	284 ¶	316 ¶	349 ¶	381 ¶
29	“	61 =	93 J	125 >	157 Ø	189 ¢	221 ¶	253 ¶	285 ¶	317 ¶	350 ¶	382 ¶
30	▲	62 >	94 ^	126 ~	158 ×	190 ¥	222 ¶	254 ¶	286 ¶	318 ¶	351 ¶	383 ¶
31	▼	63 ?	95 _	127 △	159 f	191 ɿ	223 ¶	255 ¶	287 ¶	319 ¶	352 ¶	384 ¶

Rappel: DNA sequencing types



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

Nos 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**

Récupérer les fichiers

- <https://doi.org/10.5281/zenodo.17929063>
 - normal_R1.fastq
 - normal_R2.fastq
 - tumor_R1.fastq
 - tumor_R2.fastq
 - 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

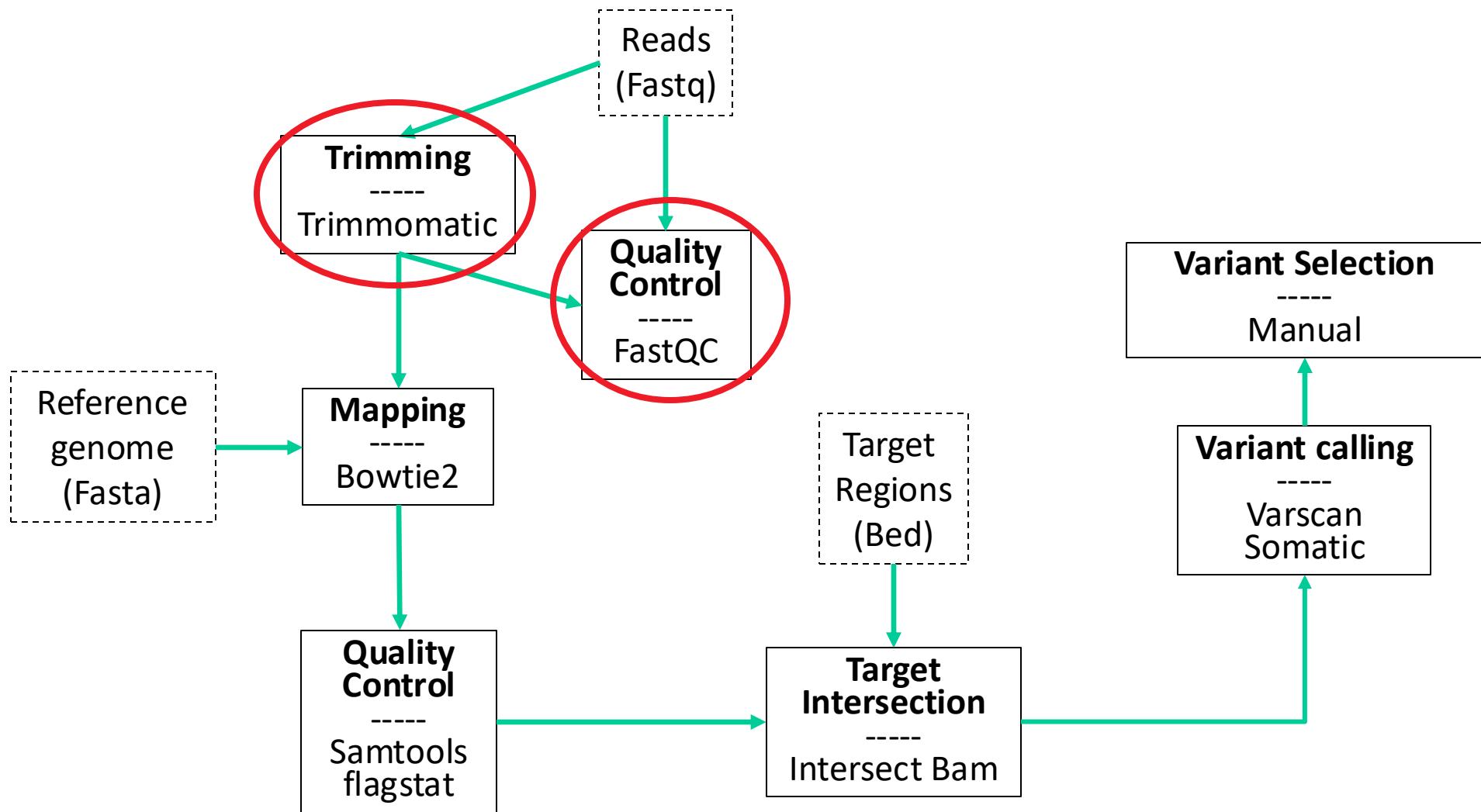
History: The History panel shows a single entry named "exome test 2" which is currently empty. A red arrow points to a tooltip in this panel: "Cet historique est vide. You can Charger vos propres données or Charger des données depuis une source externe".

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

- 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

Text at bottom: Fera apparaître:

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

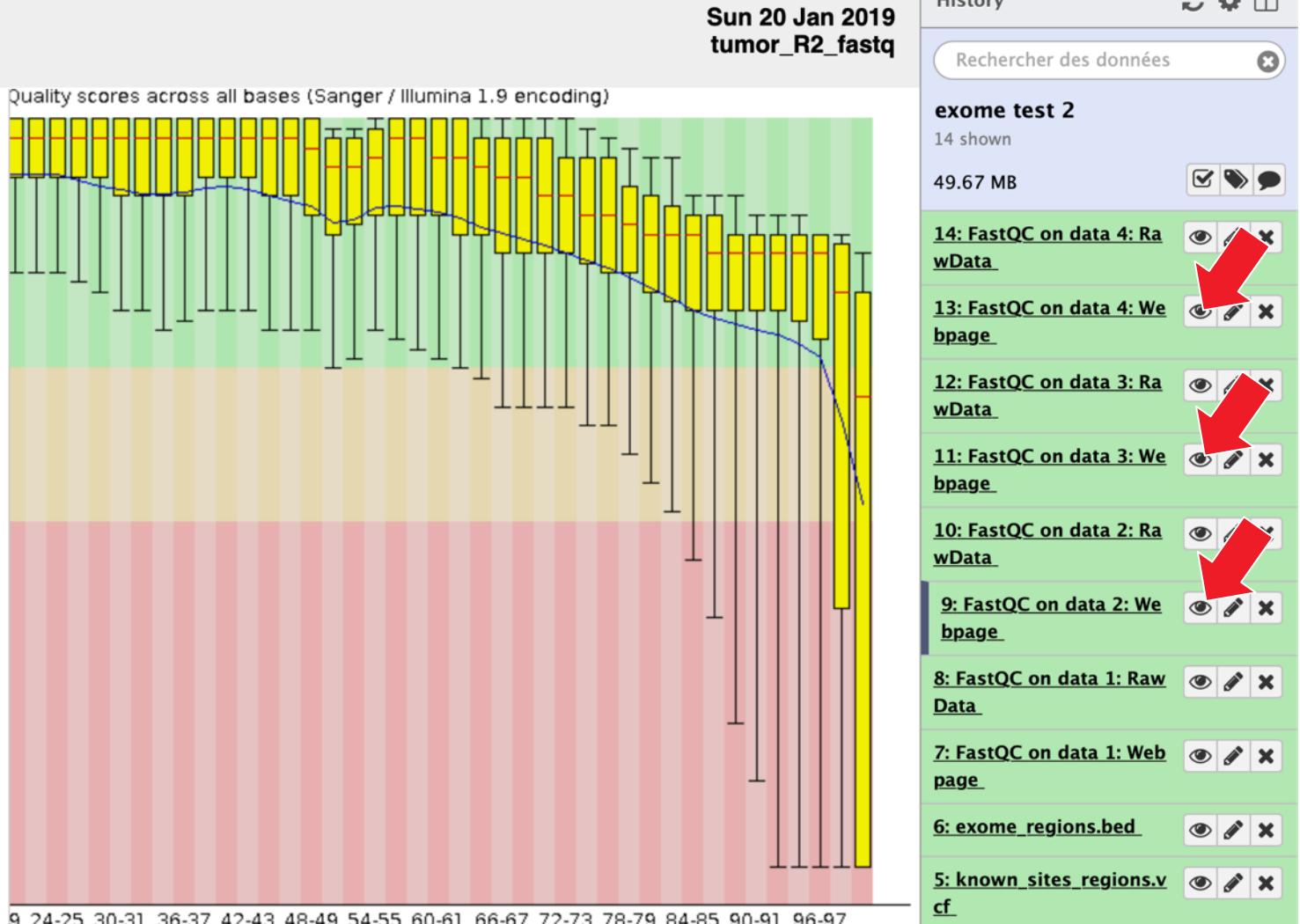
1: tumor_R1.fastq

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](#)



- 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

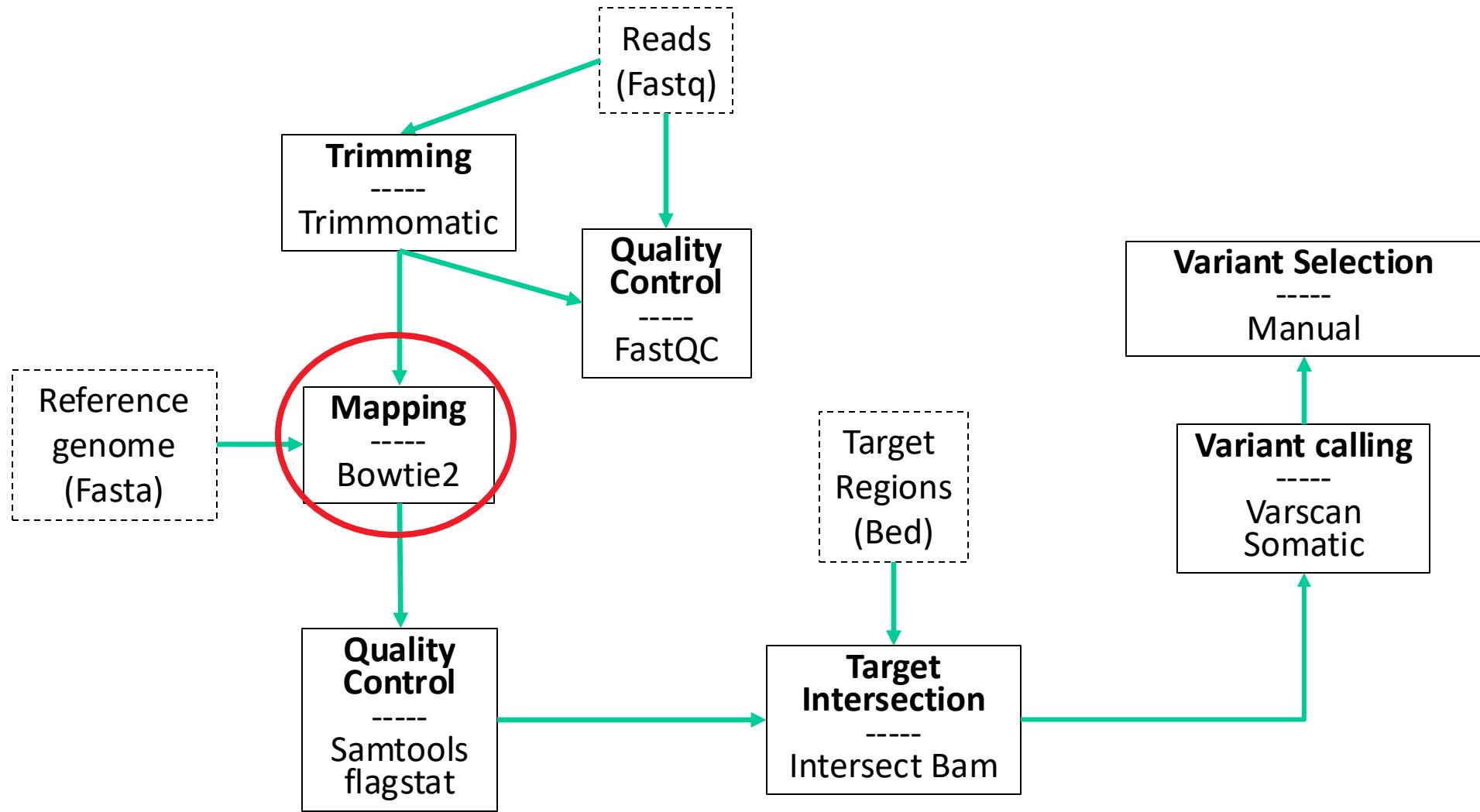
3: normal_R2.fastq

2: tumor_R2.fastq

Trimmomatic (fin)

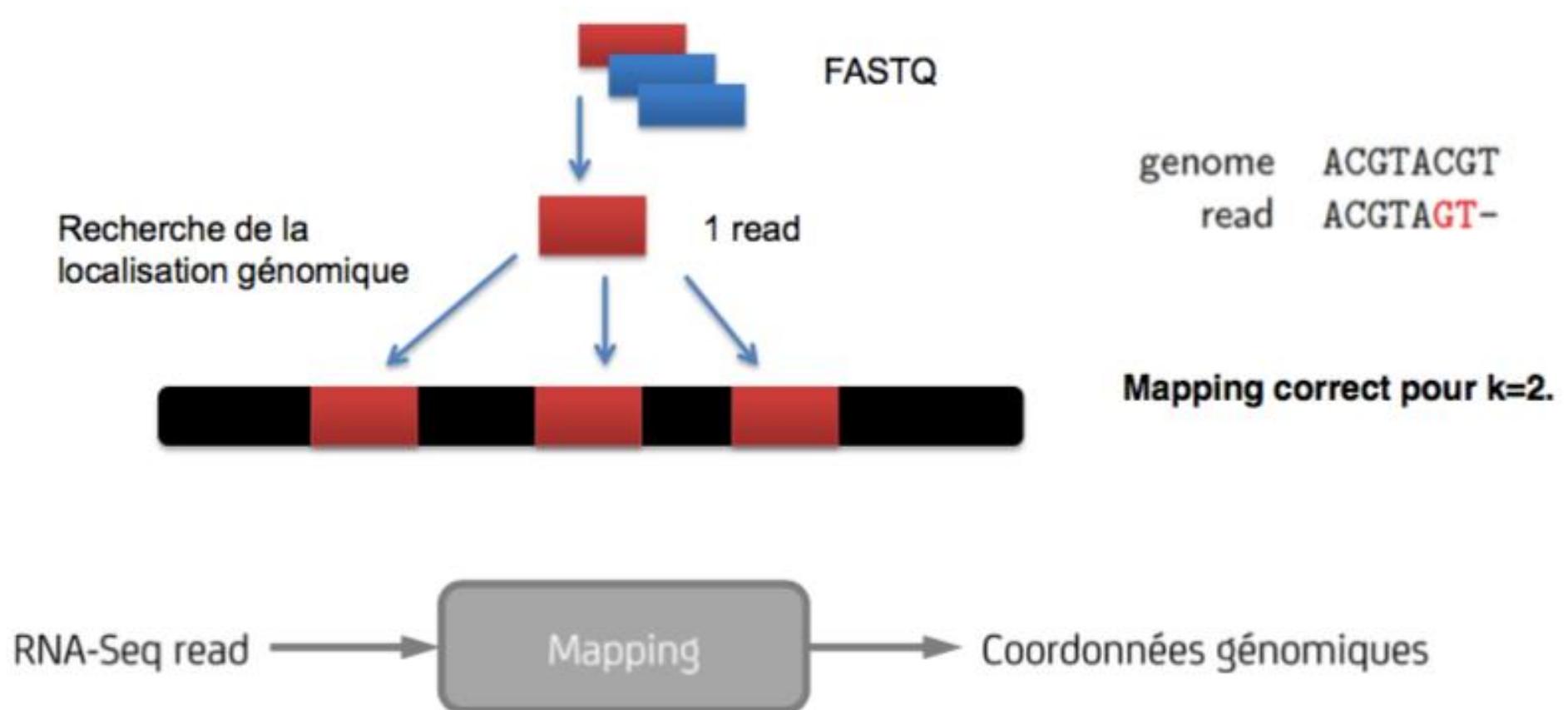
- Vérifiez le gain de qualité (faites fastqc d'un fastq)
- Eliminez les données « unpaired »

Mapping/alignement



Mapping/alignement

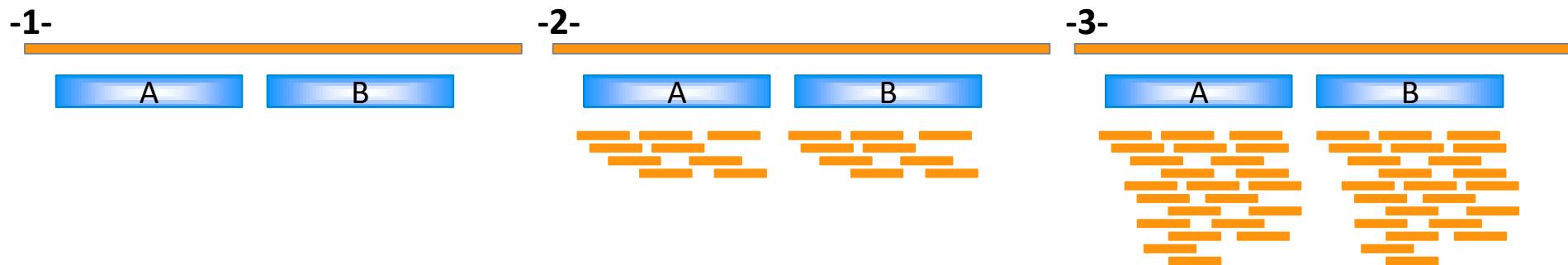
Mapper=trouver tous les endroits où le read est présent à k erreurs près.



Alignment 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

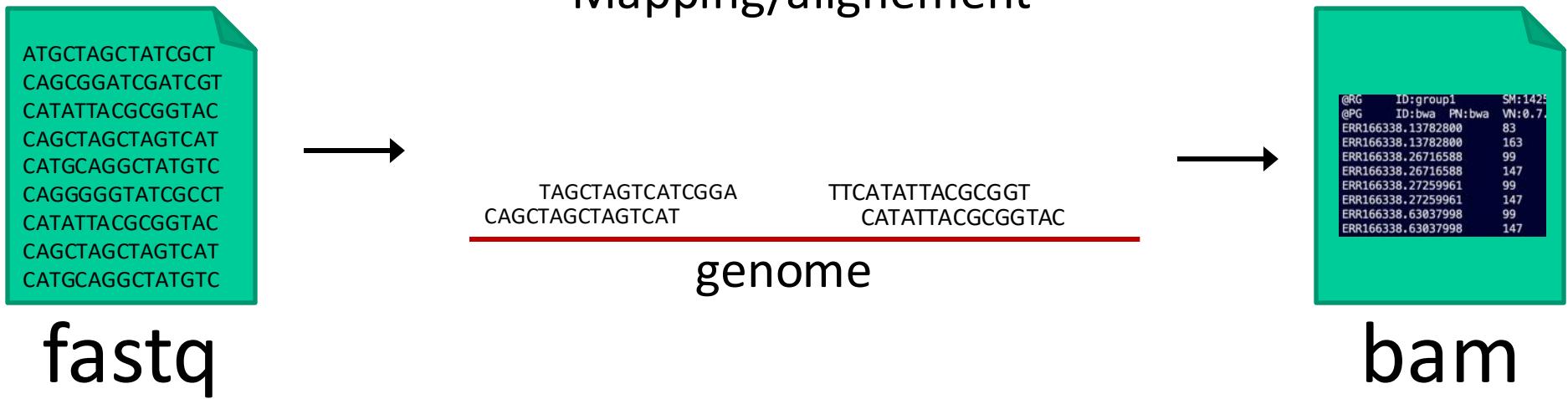
- Single-end alignment – repeated sequence



- Paired-end alignment – unique sequence



Mapping: Fastq > BAM



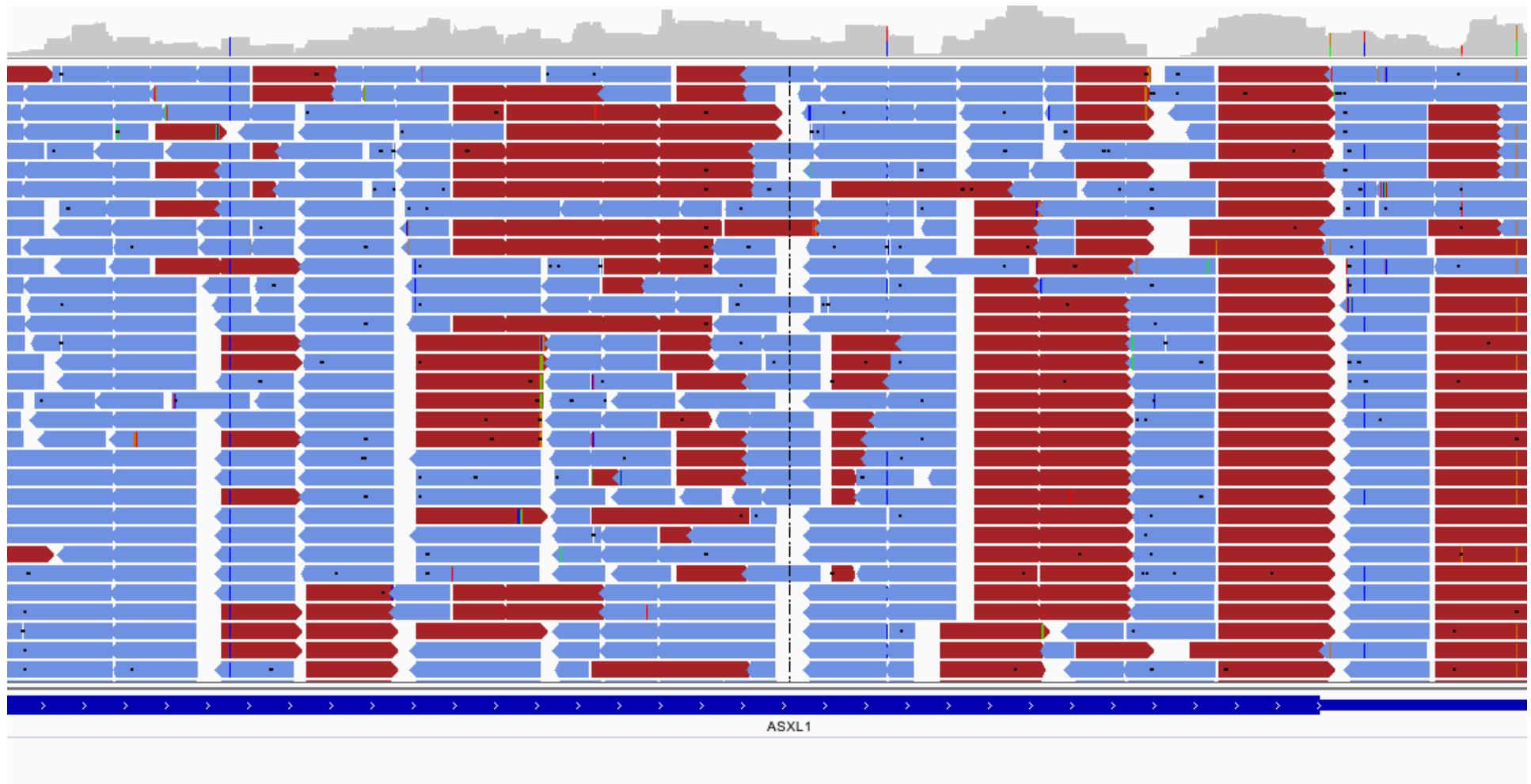
BAM format

```
@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 TTGCAGACTTACCAAGCATTGGAGGAATATCGTAAG
ERR166338.27259961 99 chr13 32890496 60 101M = 32890558 137 ACCTCAGTCACATAATAAGGAATGCATCCCTGTGTAAG
ERR166338.27259961 147 chr13 32890558 60 75M = 32890496 -137 GACTTACCAAGCATTGGAGGAATATCGTAGGTAAG
ERR166338.63037998 99 chr13 32890496 60 101M = 32890558 137 ACCTCAGTCACATAATAAGGAATGCATCCCTGTGTAAG
ERR166338.63037998 147 chr13 32890558 60 75M = 32890496 -137 GACTTACCAAGCATTGGAGGAATATCGTAGGTAAG
```



Position du
début du read

Reads alignés: le format BAM/SAM



Attention: étape de ~10min!

Bowtie2

V.2.5.3 or older

Galaxy / Europe

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

Analyse de données Workflow Visualize ▾ Données partagées ▾ Aide ▾ Utilisateur ▾

Using 0%

Using 0%

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

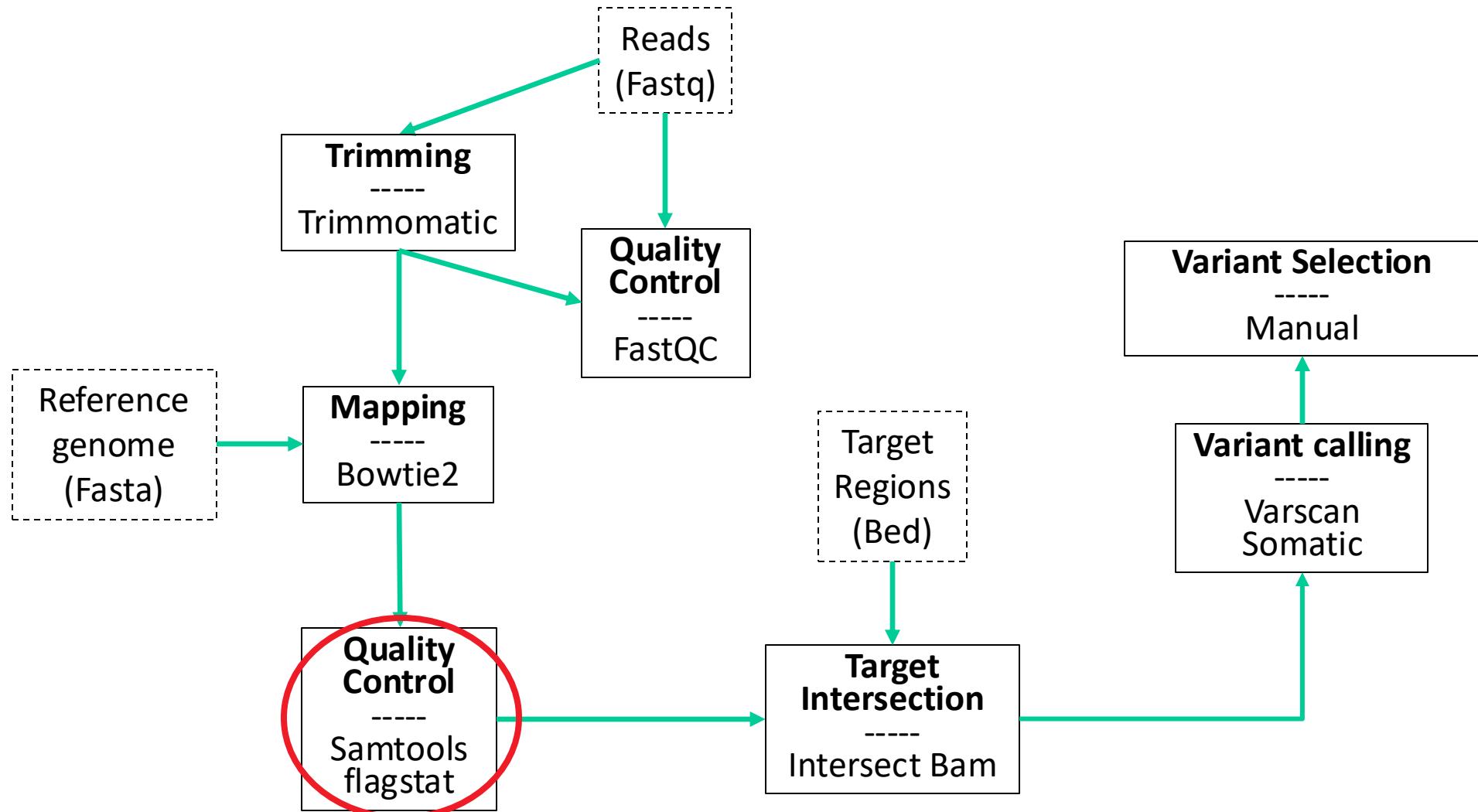
Set read groups information?
Do not set
Specifying read group information can greatly simplify your downstream analyses by allowing combining multiple datasets.

Switch to 2.5.4+galaxy0
Selected 2.5.3+galaxy1
Switch to 2.5.3+galaxy0
Switch to 2.5.0+galaxy0
Switch to 2.4.5+galaxy1
Switch to 2.4.5+galaxy0
Switch to 2.4.2+galaxy0
Switch to 2.3.4.3+galaxy0
Switch to 2.3.4.3
Switch to 2.3.4.2

Using 0%
des données
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)

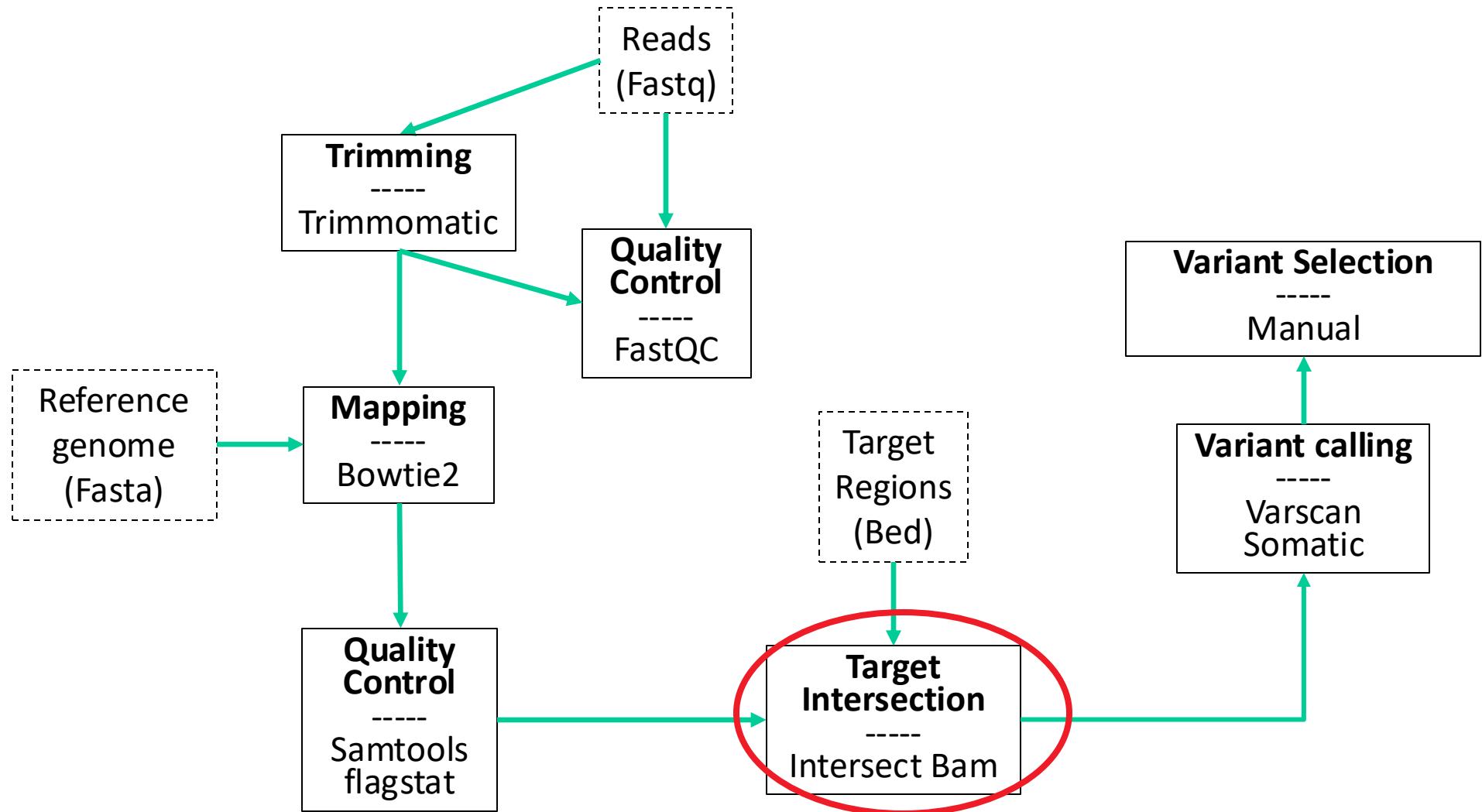
BAM file

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
```

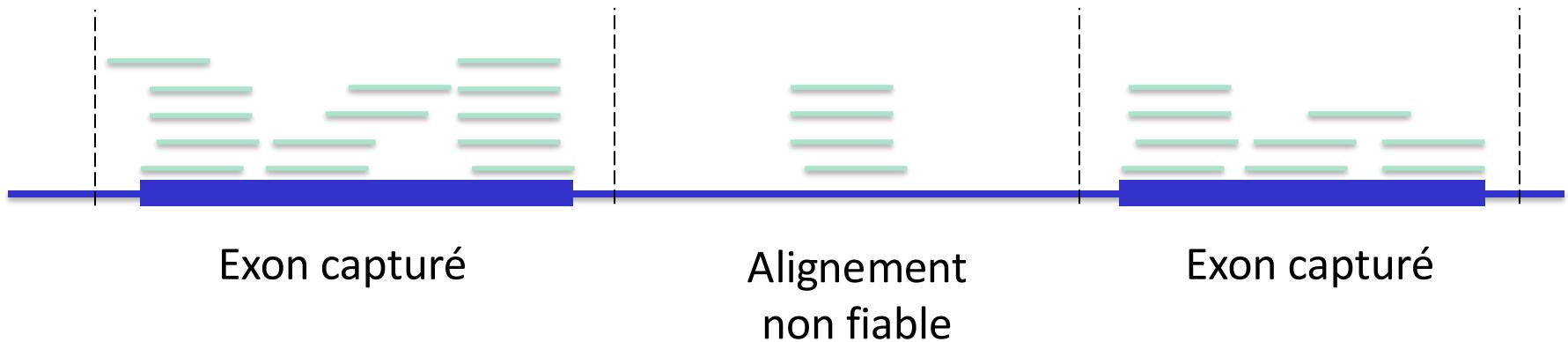
Suggestion:
renommer vos
fichiers BAM etc
avec des noms
plus simples



Target intersection

Etape non indispensable,
peut être sautée si temps
limité

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



Bedtools intersect intervals

Galaxy / Europe Analyze Data Workflow Visualize Shared Data Help User Versions Options

Using 0%

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 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

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)

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

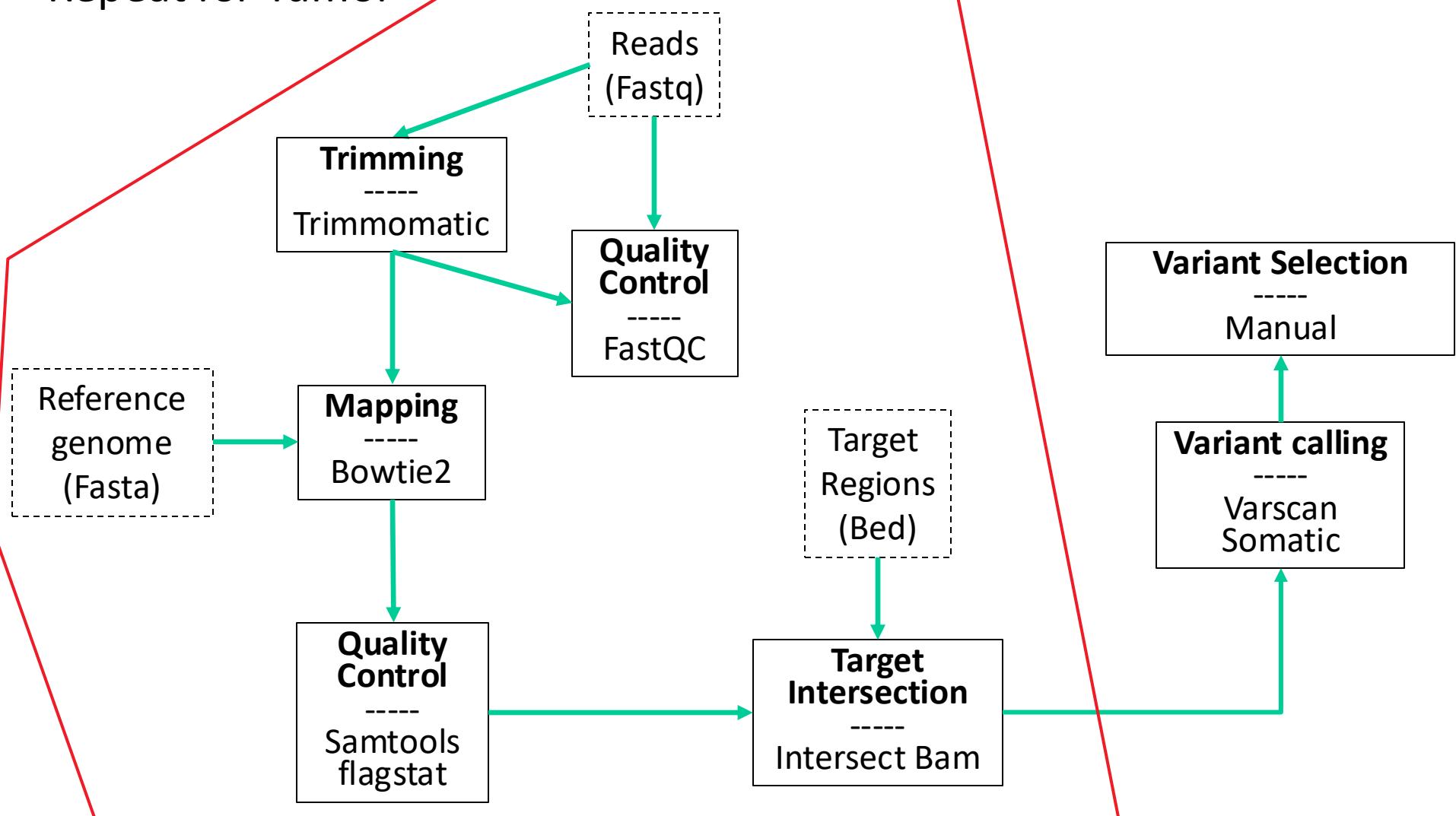
Report only those alignments that **do not** overlap with file(s) B

Yes No

Vérifiez la réduction de taille du fichier BAM

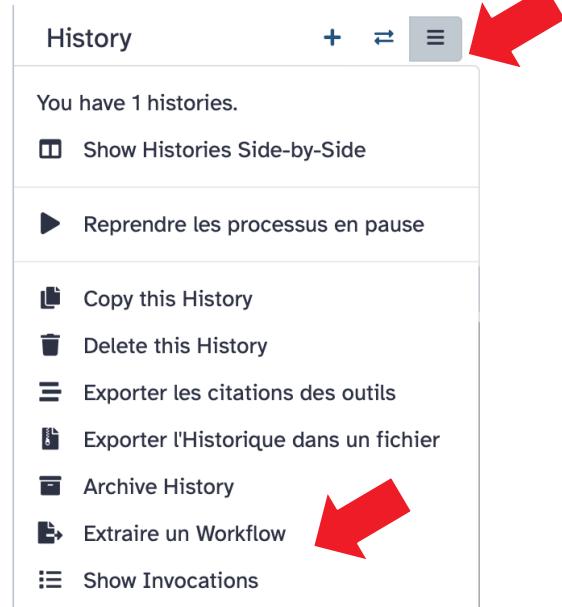
Répéter le workflow

Repeat for Tumor

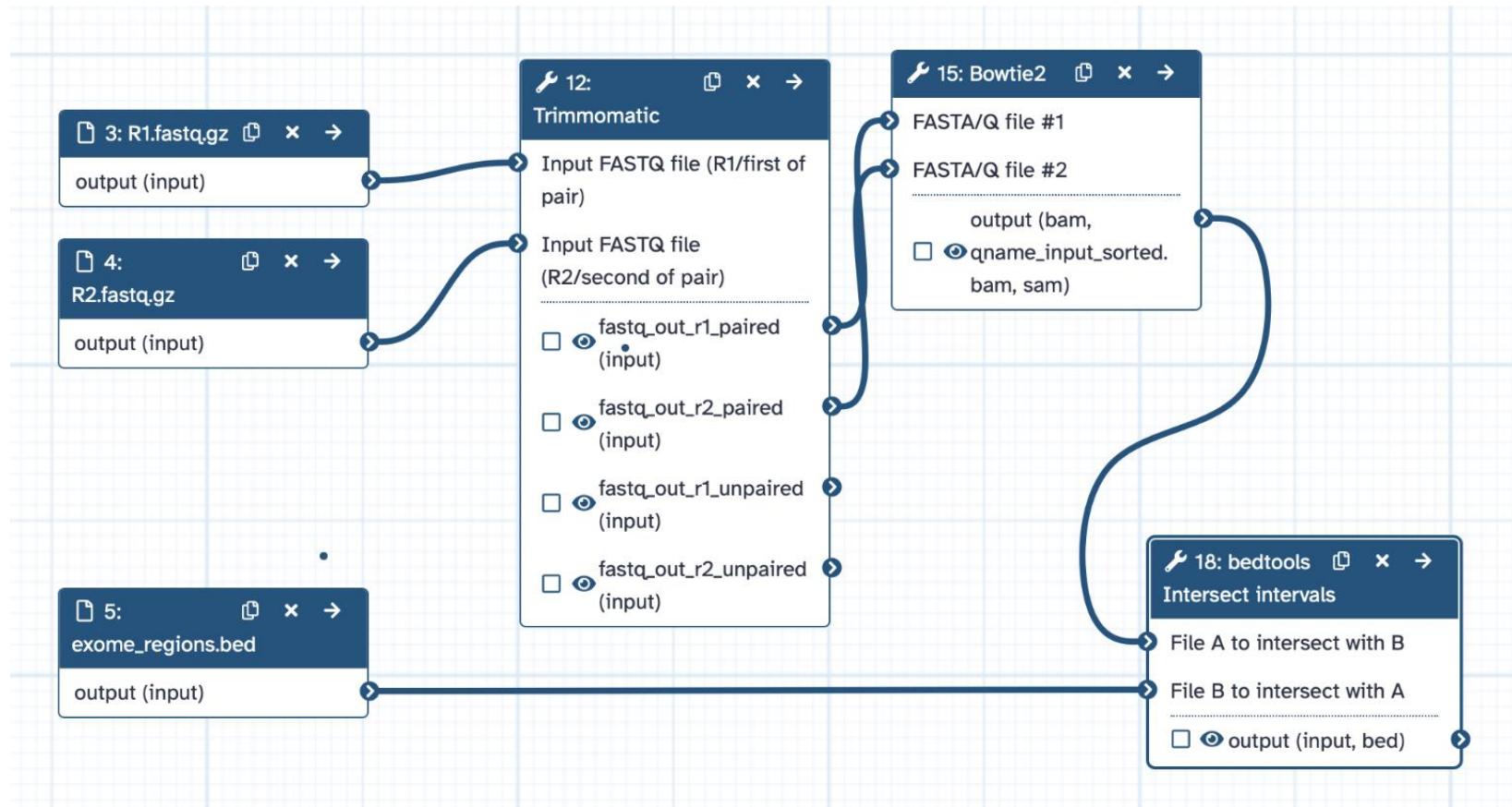


Extraire un workflow

- Extraire workflow
- Le nommer + créer
- Editer le workflow
- Choisir les données pertinentes
 - Juste 2 fastq et regions.bed
 - Enlever les data inutilisées (par ex. fastq tumor)
- Choisir les étapes de Trimmomatic à Intersect bed
 - Enlever les étapes non essentielles (fastqc)
- Renommer les objets de façon générique (par ex « sample.R1 » plutot que « normal.R1 ») (cliquer sur élément du workflow + changer nom à droite)
- Puis  save workflow



Un workflow simple

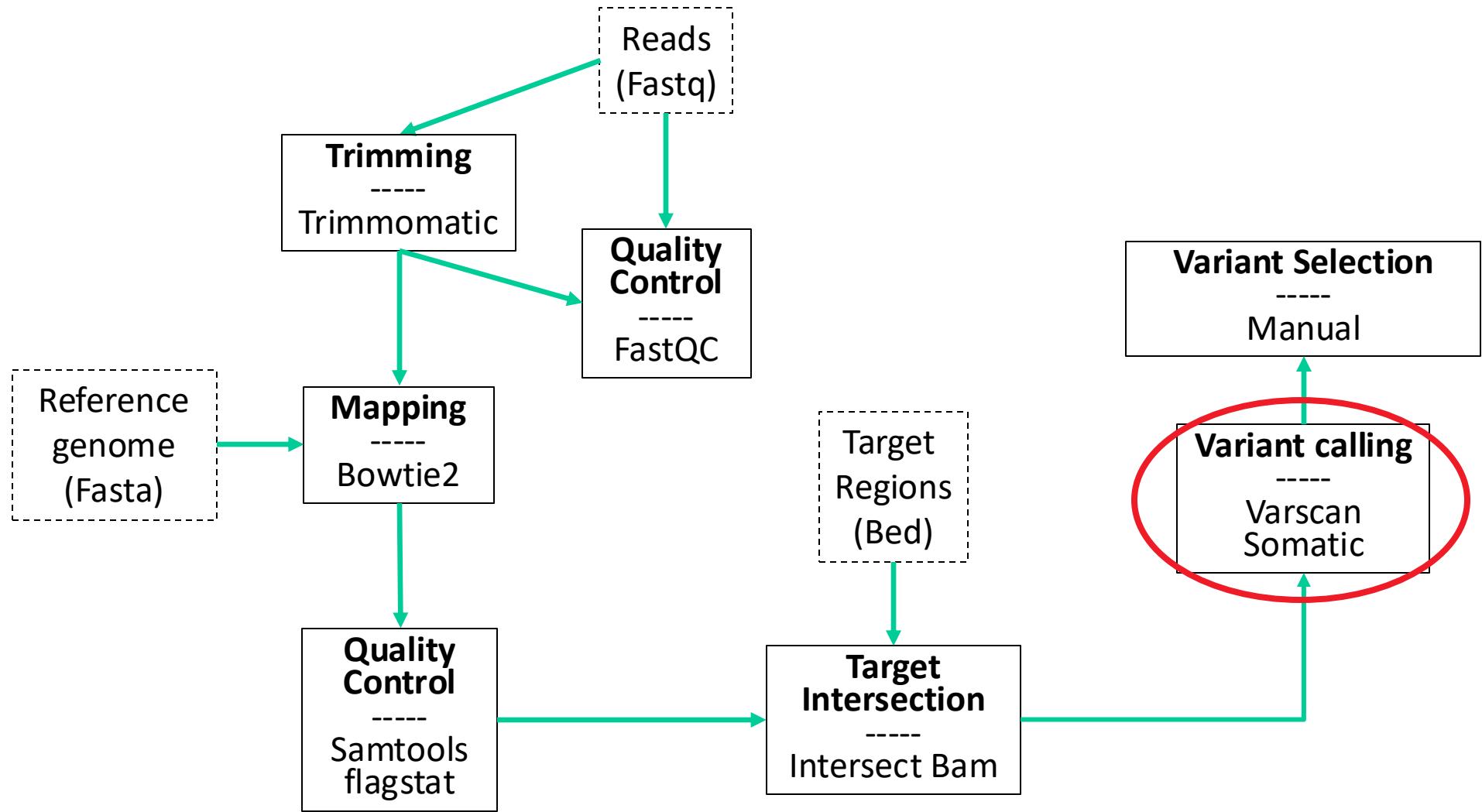


Maintenant lancez le workflow sur les données Tumor (▶ run workflow)

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
- Workflow Steps (highlighted by red arrows):**
 - 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)
- History Panel:** Pipeline1, 25 shown, 73 deleted, 151.47 MB, displaying log entries related to FASTQ files and bedtools operations.

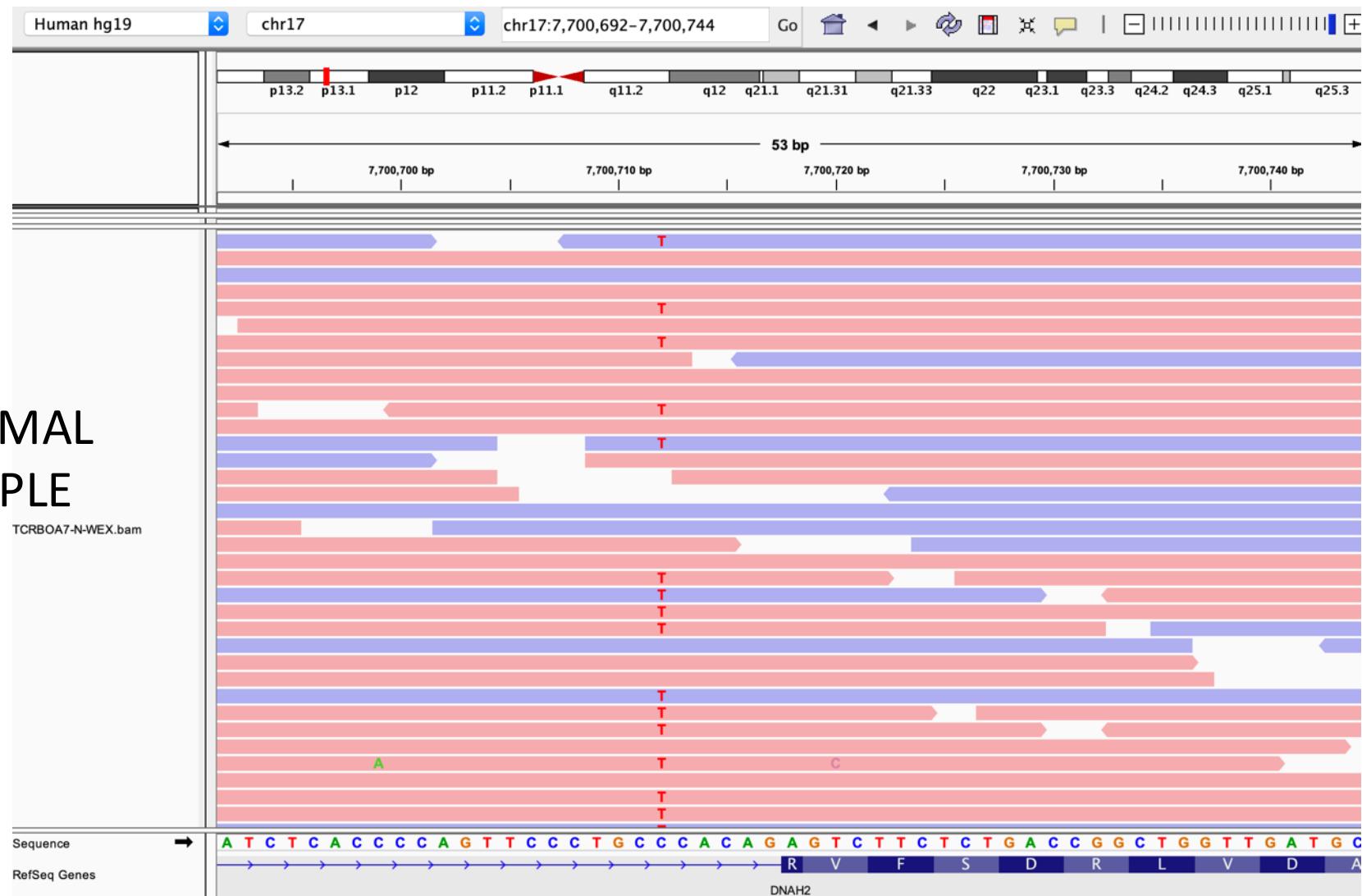
Variant calling



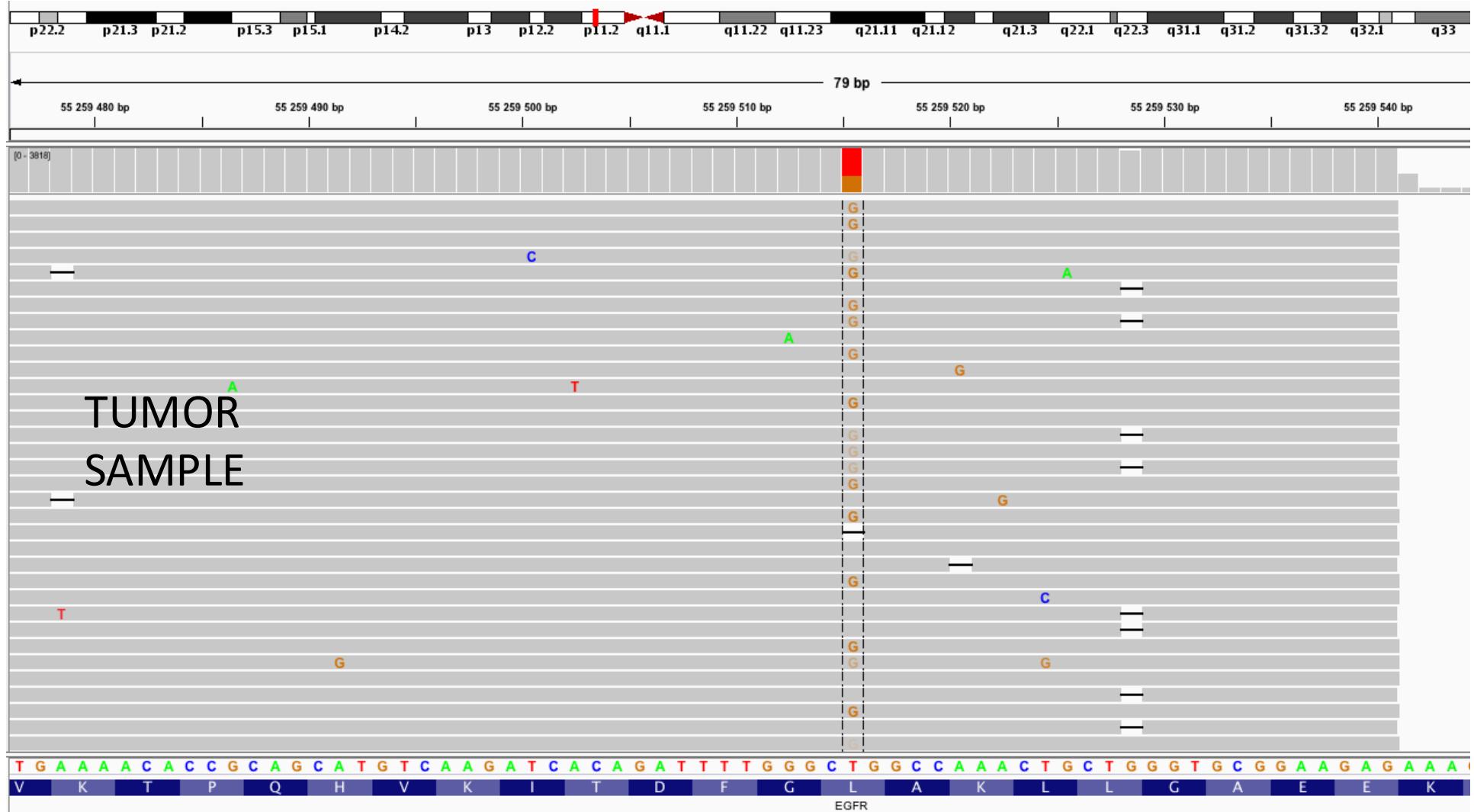
Un polymorphisme (SNP) homozygote



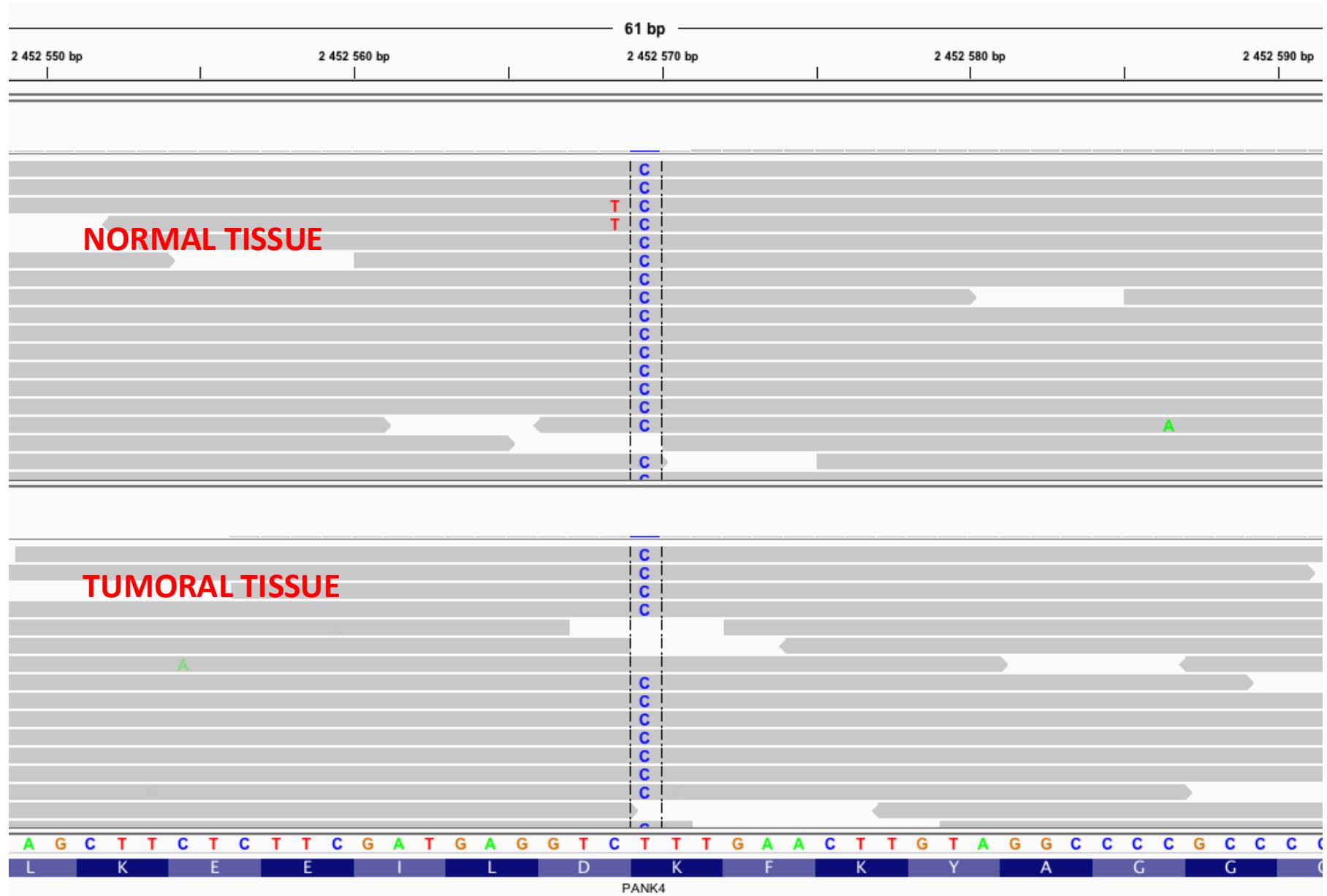
Un polymorphisme (SNP) hétérozygote



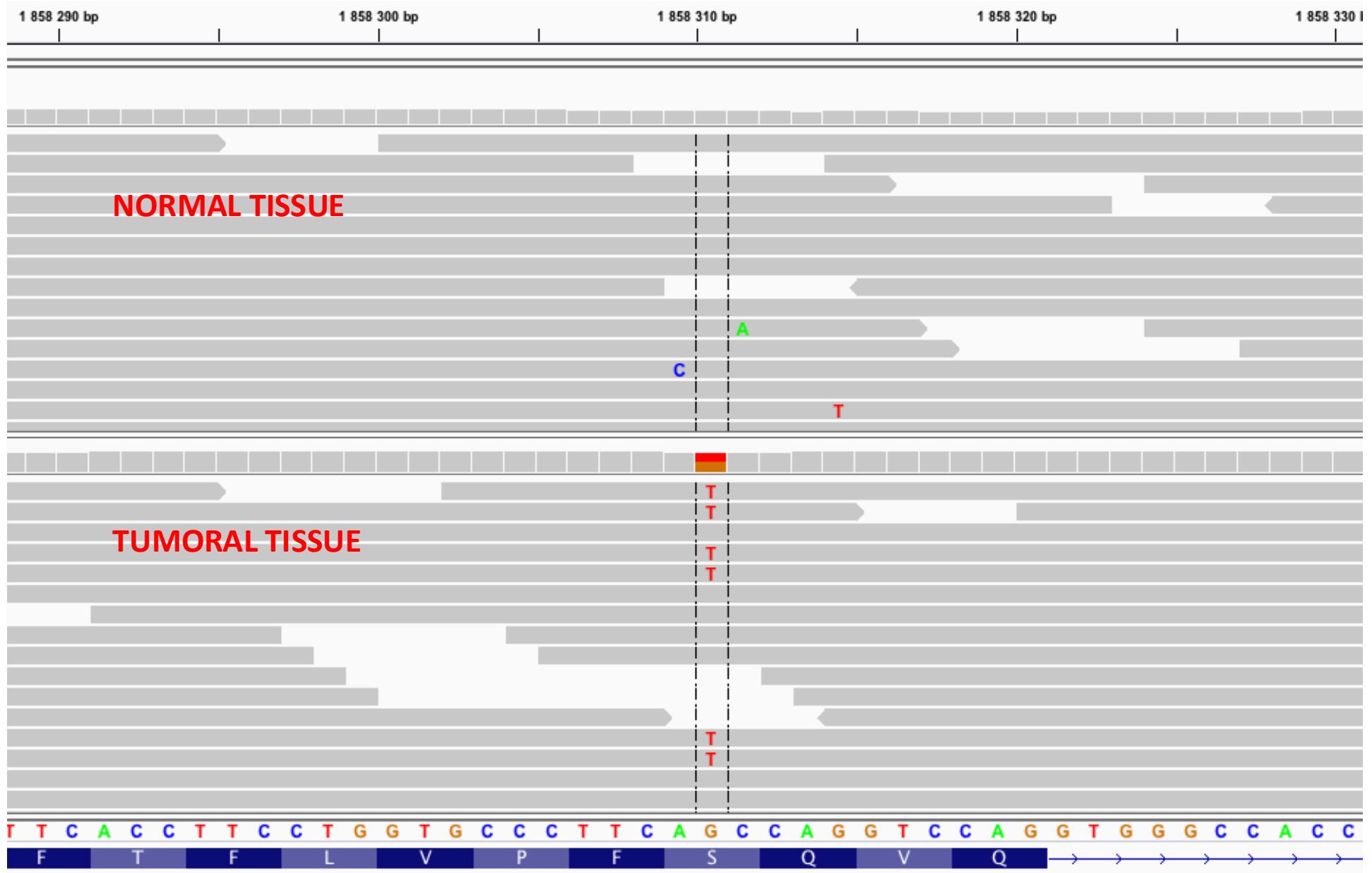
Un variant tumoral: polymorphisme ou mutation?



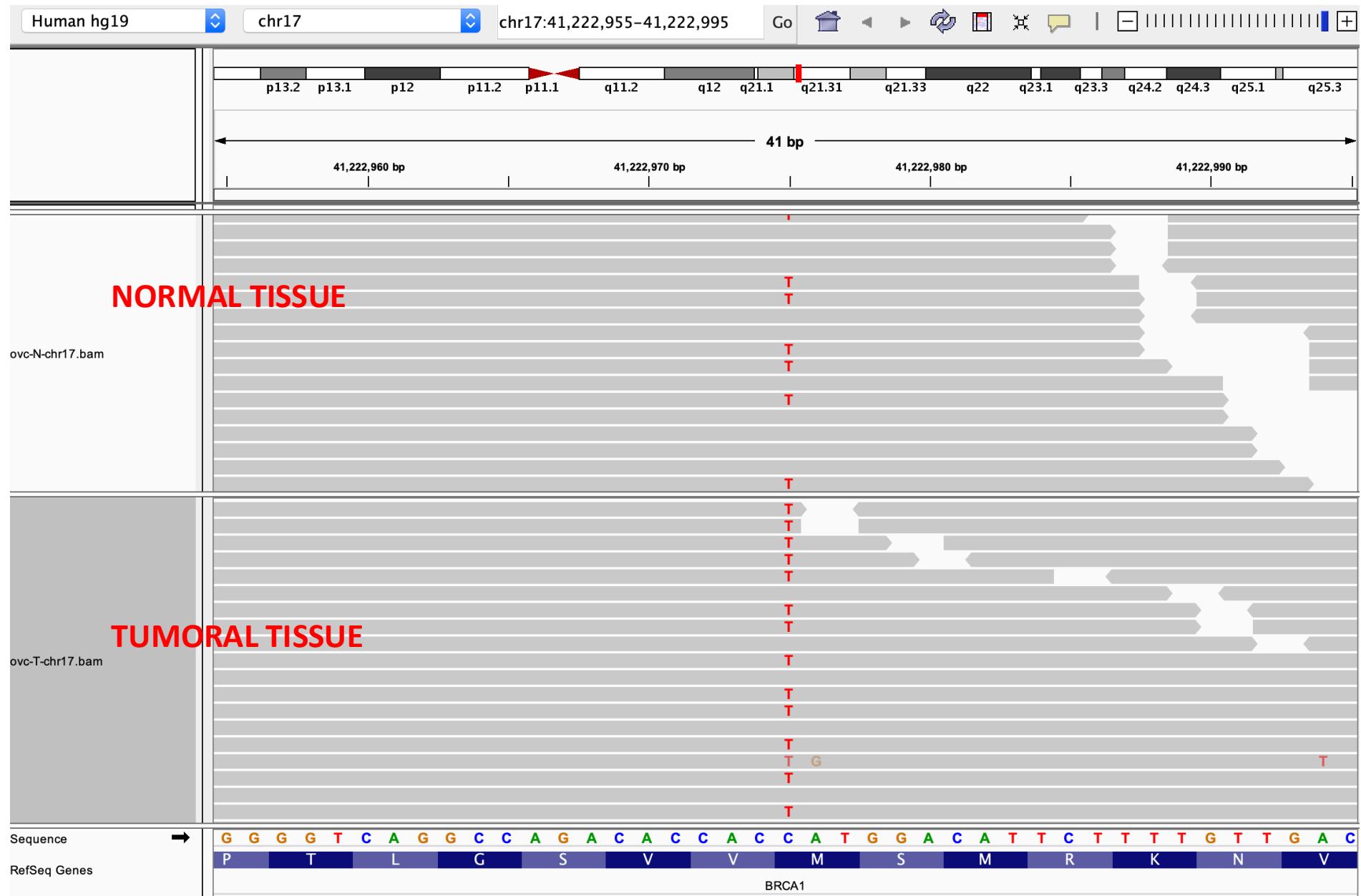
Un polymorphisme vu dans N et T



Une mutation somatique



Une LOH (loss of heterozygosity)

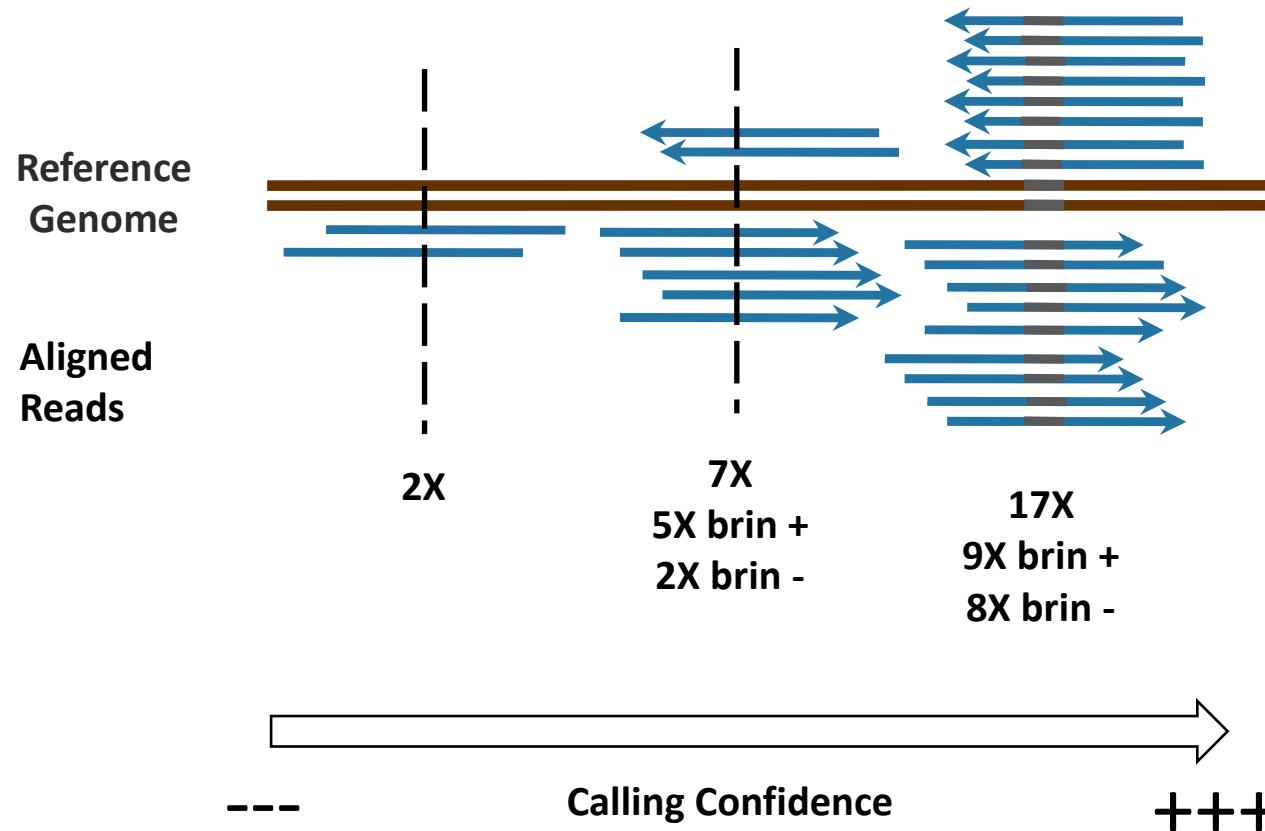




- Mutation caller written in **Java** (portable)
- Somatic Mode: input=**Tumor/Normal pairs**:
 - Somatic, germline, LOH events
 - Somatic copy number alterations (CNAs)

Variant Calling Quality

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



Varscan's Somatic P-value

Normal



17 positions
= reference

Tumor

10 positions
= reference,
9 positions
=variant

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

Alleles
Ref Var

N	17	0
T	10	9

Somatic

Le Format VCF

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT
chr11	26692481	.	C	T	.	PASS	DP=25;SS=1;SSC=3;GPV=0.00031821;SPV=0.46602	GT:GQ:DP:AD:ADF:>
chr11	26692594	.	T	TTGTGTGTG	.	.	PASS DP=118;SS=1;SSC=8;GPV=3.3217e-07;SPV=0.13917;INDEL	>
chr11	26692631	.	A	G	.	PASS	DP=163;SS=1;SSC=7;GPV=4.5508e-25;SPV=0.16668	GT:GQ:DP:AD:ADF:>
chr11	26692733	.	T	C	.	PASS	DP=205;SS=1;SSC=16;GPV=1;SPV=0.020219	GT:GQ:DP:AD:ADF:ADR >
chr11	26692742	.	G	A	.	PASS	DP=196;SS=1;SSC=13;GPV=1;SPV=0.041646	GT:GQ:DP:AD:ADF:ADR >
chr11	26692811	.	C	G	.	PASS	DP=120;SS=1;SSC=16;GPV=1;SPV=0.020385	GT:GQ:DP:AD:ADF:ADR >
chr11	26692933	.	A	T	.	PASS	DP=26;SS=1;SSC=0;GPV=1.7241e-06;SPV=0.83112	GT:GQ:DP:AD:ADF:>
chr11	26694979	.	A	G	.	PASS	DP=131;SS=1;SSC=0;GPV=1.1116e-22;SPV=0.86166	GT:GQ:DP:AD:ADF:>
chr11	26702664	.	C	T	.	PASS	DP=37;SS=1;SSC=0;GPV=5.7269e-22;SPV=1	GT:GQ:DP:AD:ADF:ADR >
chr11	26724995	.	C	A	.	PASS	DP=35;SS=1;SSC=0;GPV=4.6376e-07;SPV=0.80442	GT:GQ:DP:AD:ADF:>
chr11	26732923	.	A	G	.	PASS	DP=37;SS=1;SSC=1;GPV=2.432e-07;SPV=0.67888	GT:GQ:DP:AD:ADF:>
chr11	26734209	.	C	T	.	PASS	DP=459;SS=1;SSC=0;GPV=0;SPV=1	GT:GQ:DP:AD:ADF:ADR 1/1::18>
chr11	43905440	.	C	G	.	PASS	DP=112;SS=1;SSC=0;GPV=1.0892e-20;SPV=0.9295	GT:GQ:DP:AD:ADF:>
chr12	25378594	.	C	T	.	PASS	DP=194;SOMATIC;SS=2;SSC=31;GPV=1;SPV=0.00072541	GT:GQ:DP:AD:ADF:
chr12	25398285	.	C	A	.	PASS	DP=42;SOMATIC;SS=2;SSC=6;GPV=1;SPV=0.20035	GT:GQ:DP:AD:ADF:
chr12	109530278	.	G	GCA	.	PASS	DP=24;SS=1;SSC=1;GPV=2.4726e-07;SPV=0.66321;INDEL	GT:GQ:DP
chr13	32893197	.	A	AT	.	PASS	DP=99;SOMATIC;SS=2;SSC=29;GPV=1;SPV=0.0010034;INDEL	GT:GQ:DP
chr13	32906480	.	A	C	.	PASS	DP=160;SS=1;SSC=3;GPV=7.7968e-28;SPV=0.45731	GT:GQ:DP:AD:ADF:

Insertion

DP: combined depth

SS: somatic status (0:ref, 1:germline, 2:somatic)

SSC: somatic score (Phred)

GT: genotype (0=ref, 1=alt)

GQ: genotype quality (Phred)

DP: read depth

HQ: haplotype quality

Somatic variant calling: Varscan

Attention: étape de ~10min!

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

Tools
[varsan somatic](#)
Variant Calling
[VarScan somatic Call](#)
germline/somatic and LOH variants from tumor-normal sample pairs

Workflows
• All workflows

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

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](#).

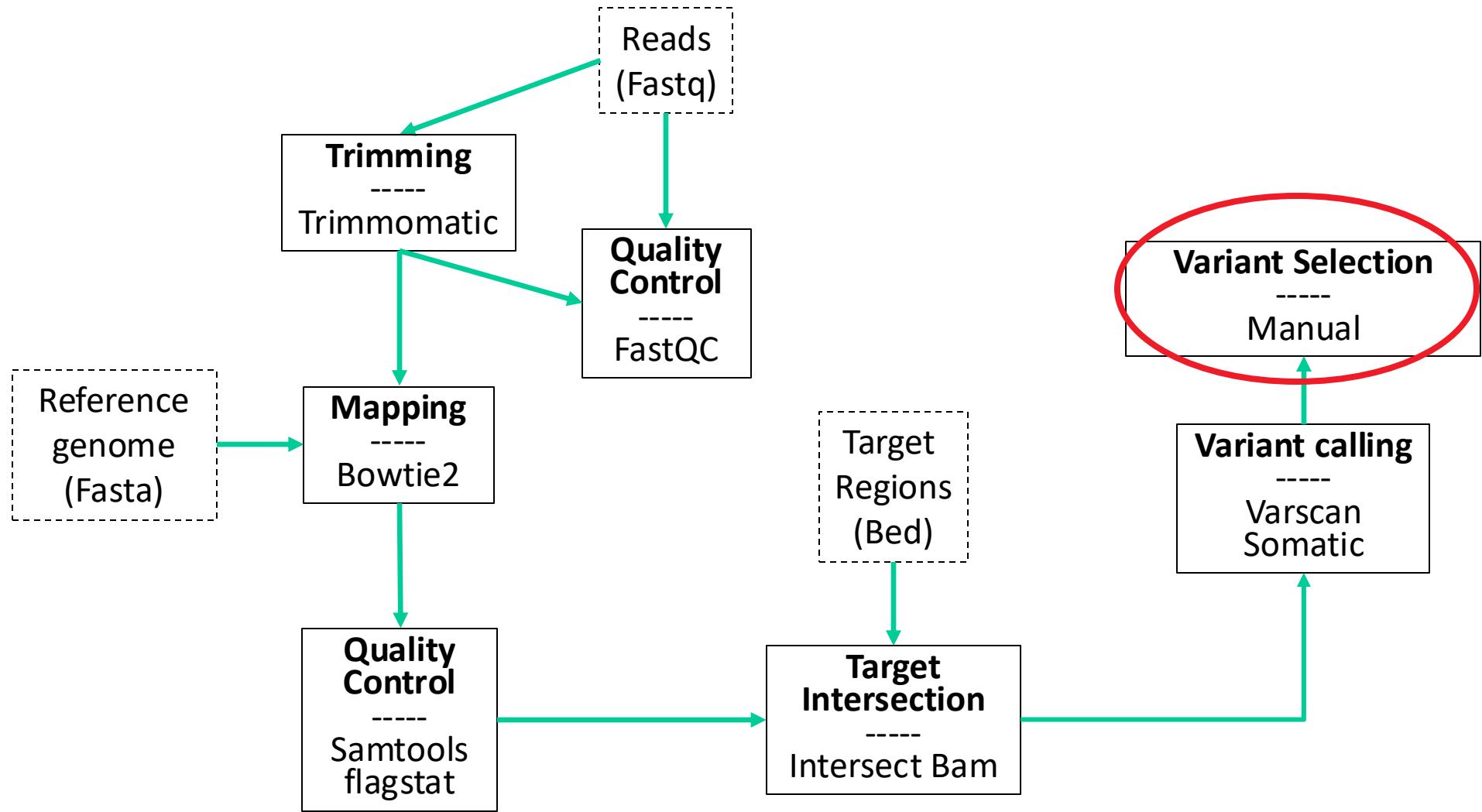
The Varscan Somatic tool for Galaxy

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
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)

Vérifiez la sortie de Varscan

FILE AND META TOOLS		Variant Call Format (VCF) Data							Pipeline1						
		Chromosome	Position	Ref	Alt	Type	Quality	DP	SS	SC	GPV	SPV	Notes	Actions	
Get Data		chr17	18874685	.	C	CGGT	.	PASS	DP=32;SS=3;SSC=16;GPV=1;SPV=0.022989;INDEL				24 shown, 72 deleted		
Send Data		chr17	18874720	.	C	G	.	PASS	DP=33;SS=1;SSC=0;GPV=1.3852e-19;SPV=1				151.47 MB		
Convert Formats		chr17	18882991	.	T	A	.	PASS	DP=60;SS=1;SSC=0;GPV=1.035e-35;SPV=1						
Collection Operations		chr17	41256074	.	C	CA	.	PASS	DP=81;SS=1;SSC=1;GPV=0.0015196;SPV=0.63343;INDEL						
GENERAL TEXT TOOLS		chr17	73759304	.	G	T	.	PASS	DP=36;SS=1;SSC=0;GPV=2.2598e-21;SPV=1						
Text Manipulation		chr19	6374813	.	T	C	.	PASS	DP=33;SS=1;SSC=0;GPV=2.8029e-05;SPV=0.8425						
Filter and Sort		chr19	7550844	.	G	A	.	PASS	DP=44;SS=1;SSC=4;GPV=2.3358e-10;SPV=0.35332						
Join, Subtract and Group		chr19	36504365	.	C	T	.	PASS	DP=34;SS=1;SSC=1;GPV=5.1914e-07;SPV=0.63966						
GENOMICS, NGS		chr1	10596341	.	C	T	.	PASS	DP=44;SS=1;SSC=2;GPV=7.4746e-10;SPV=0.53262						
Extract Features		chr1	160251792	.	A	G	.	PASS	DP=37;SS=1;SSC=0;GPV=5.1339e-06;SPV=0.87856						
BED Tools		chr1	167082869	.	G	A	.	PASS	DP=71;SS=1;SSC=8;GPV=2.0173e-19;SPV=0.13252						
Fetch Alignments		chr1	167095163	.	G	C	.	PASS	DP=52;SS=1;SSC=5;GPV=6.8522e-13;SPV=0.28624						
Operate on Genomic Intervals		chr1	167097739	.	C	A	.	PASS	DP=64;SS=1;SSC=3;GPV=4.3049e-14;SPV=0.44587						
FASTA/FASTQ manipulation		chr1	214788427	.	C	T	.	PASS	DP=45;SS=1;SSC=1;GPV=8.5784e-10;SPV=0.66234						
Multiple Alignments		chr1	214802553	.	CT	C	.	PASS	DP=83;SOMATIC;SS=2;SSC=18;GPV=1;SPV=0.015148;INDEL						
FASTA/FASTQ manipulation		chr1	214803969	.	G	C	.	PASS	DP=111;SOMATIC;SS=2;SSC=35;GPV=1;SPV=0.00029013						
Picard		chr1	214804041	.	C	A	.	PASS	DP=65;SS=1;SSC=0;GPV=2.7963e-08;SPV=0.9934						
Quality Control		chr1	214811174	.	G	A	.	PASS	DP=76;SS=1;SSC=0;GPV=3.6183e-12;SPV=0.99124						
Assembly		chr1	214811244	.	C	G	.	PASS	DP=120;SS=1;SSC=0;GPV=1.7875e-19;SPV=0.92629						
Mapping		chr1	214813487	.	A	G	.	PASS	DP=291;SS=1;SSC=3;GPV=1.3526e-38;SPV=0.47444						
Variant Calling		chr1	214813782	.	A	G	.	PASS	DP=108;SS=1;SSC=0;GPV=1.7692e-19;SPV=0.98472						
Genome editing		chr1	214813941	.	C	G	.	PASS	DP=86;SS=1;SSC=4;GPV=8.038e-16;SPV=0.34707						
Variant Calling		chr1	214814125	.	G	A	.	PASS	DP=80;SS=1;SSC=0;GPV=1.2414e-11;SPV=0.85982						
Genome editing		chr1	214814582	.	G	A	.	PASS	DP=226;SS=1;SSC=5;GPV=3.0361e-32;SPV=0.28302						
Genome editing		chr1	214814733	.	T	G	.	PASS	DP=244;SS=1;SSC=0;GPV=2.27499e-40;SPV=0.97323						

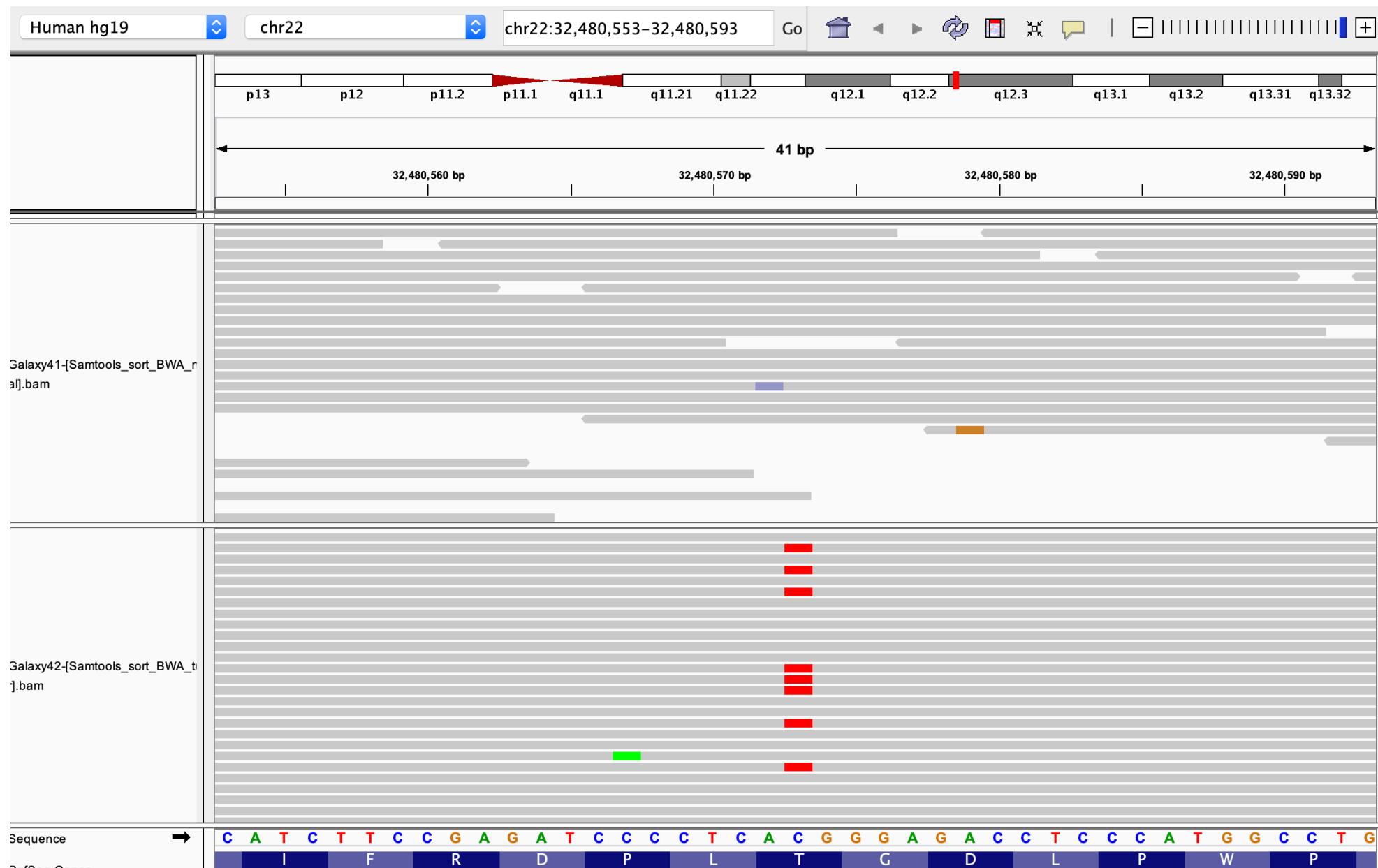
- Vous pouvez utiliser la fonction « grep » pour filter les lignes avec somatic ou LOH
 - Vérifiez la somatic P-value (SPV), les comptages
 - Regardez les sous IGV



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

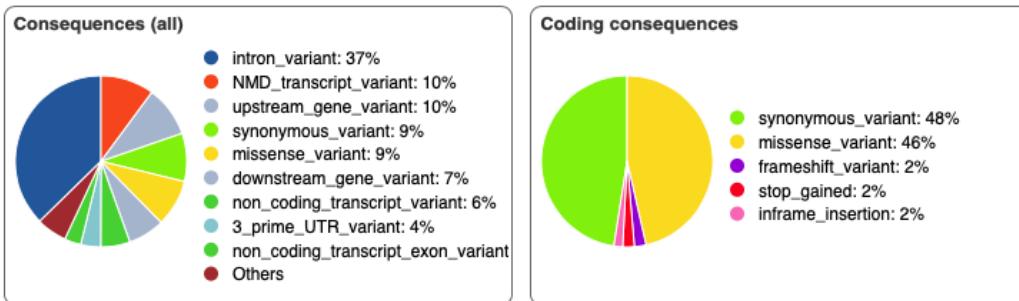
Note: the highest impact variants are not necessarily somatic!

Variant Effect Predictor results

Job details 

Summary statistics 

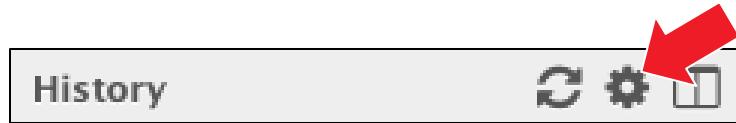
Category	Count
Variants processed	153
Variants filtered out	0
Novel / existing variants	6 (3.9) / 147 (96.1)
Overlapped genes	55
Overlapped transcripts	318
Overlapped regulatory features	23



Results preview

Show/hide columns (2 hidden)												
Uploaded variant	Location	Allele	Consequence	Impact	Symbol	Gene	Feature type	Feature	Biotype	Exon	Intron	cDNA position
. 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	

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 large red arrow pointing to a "+" button. Below this, there's an "Author" section. To the right of the modal is a list of files, each with edit and delete icons:

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:

Samtools mpileup sur fichier intersect bed

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