Réalisation d'un Pipeline d'analyse d'exome





Les données

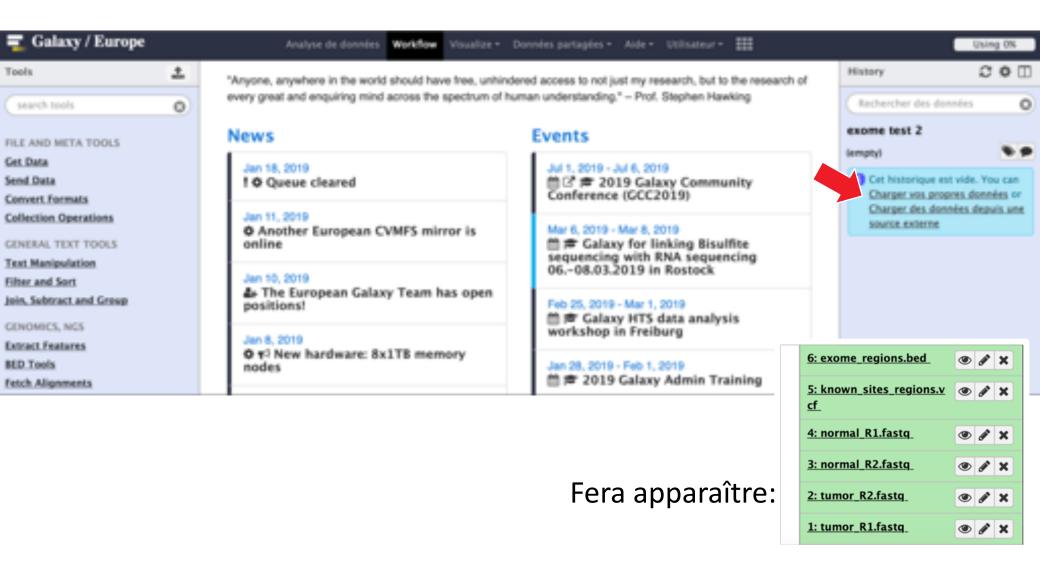


Normal tissue (blood)
Tumor tissue (non small cell lung cancer)

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

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

Chargez vos données



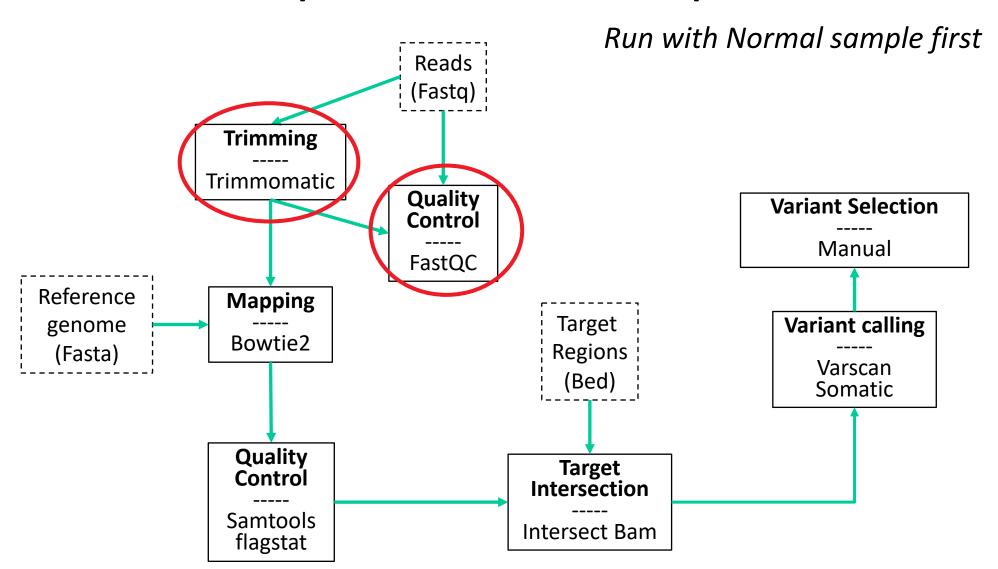
(alternativement: à partir de données partagées)

- Menu « Données partagées »
- Histories
- Choisir History « ... IFSBM ...»
- Import History

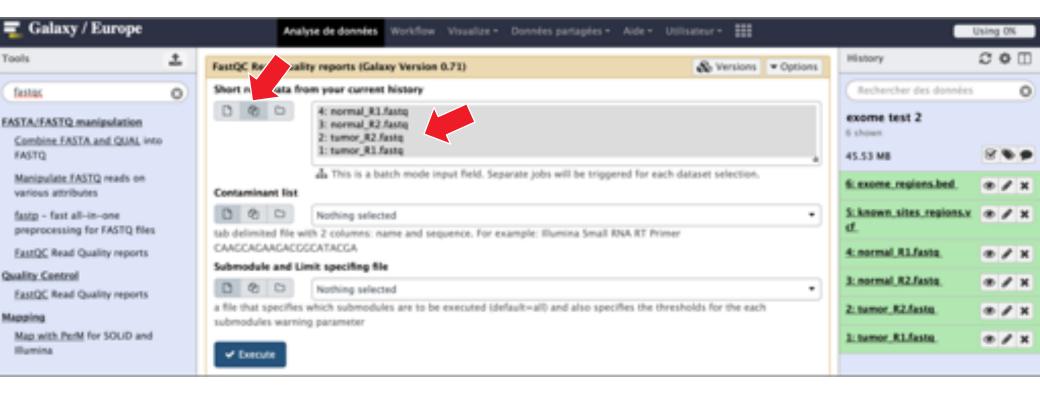


Fera apparaître:

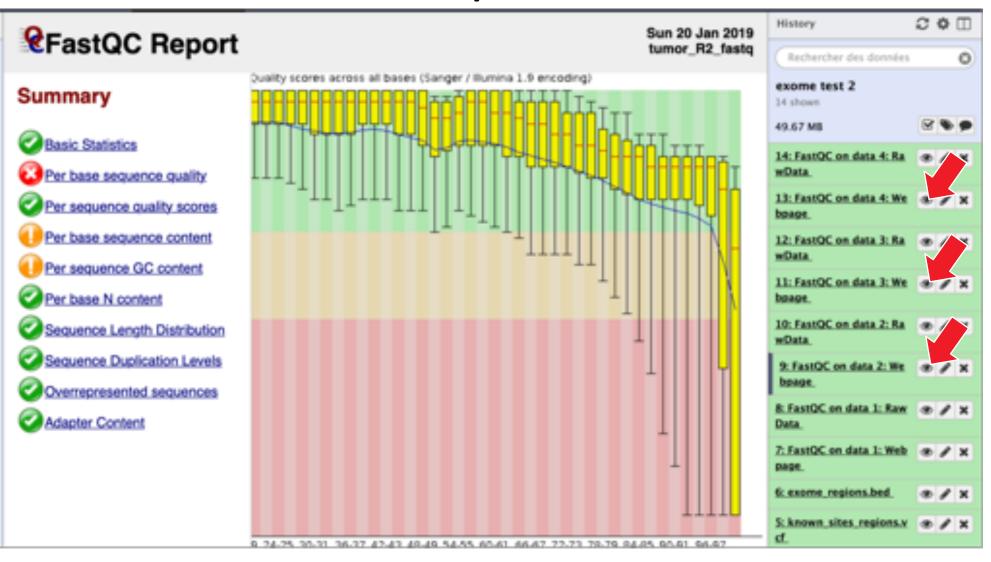
A simplified Variant Pipeline



fastqc

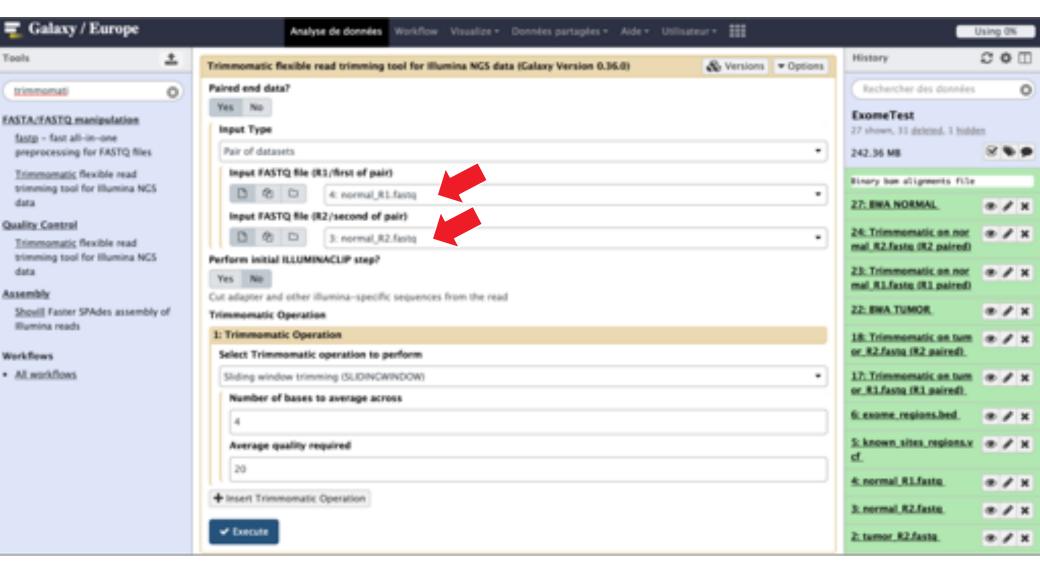


Fastqc results



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

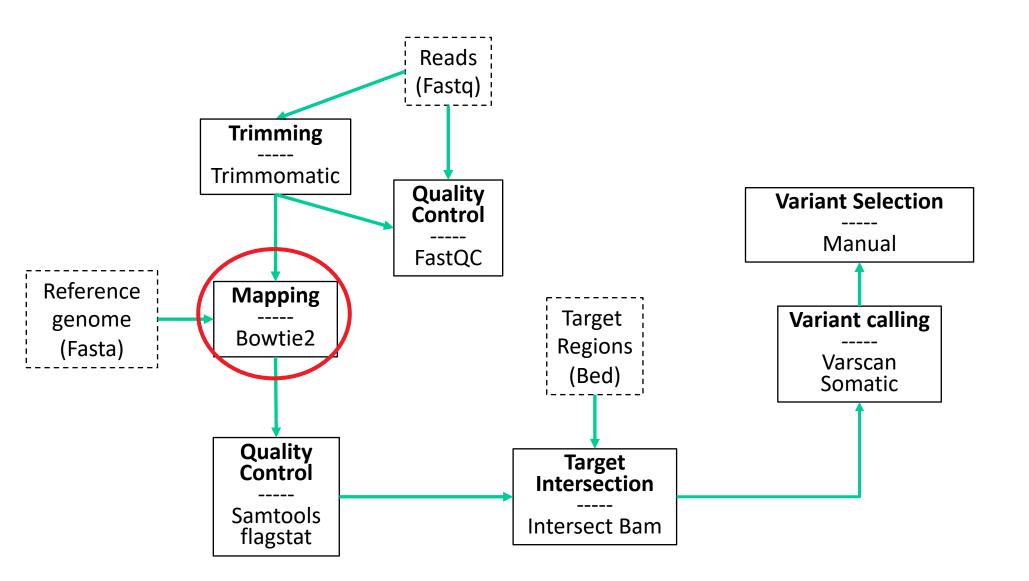
Trimmomatic



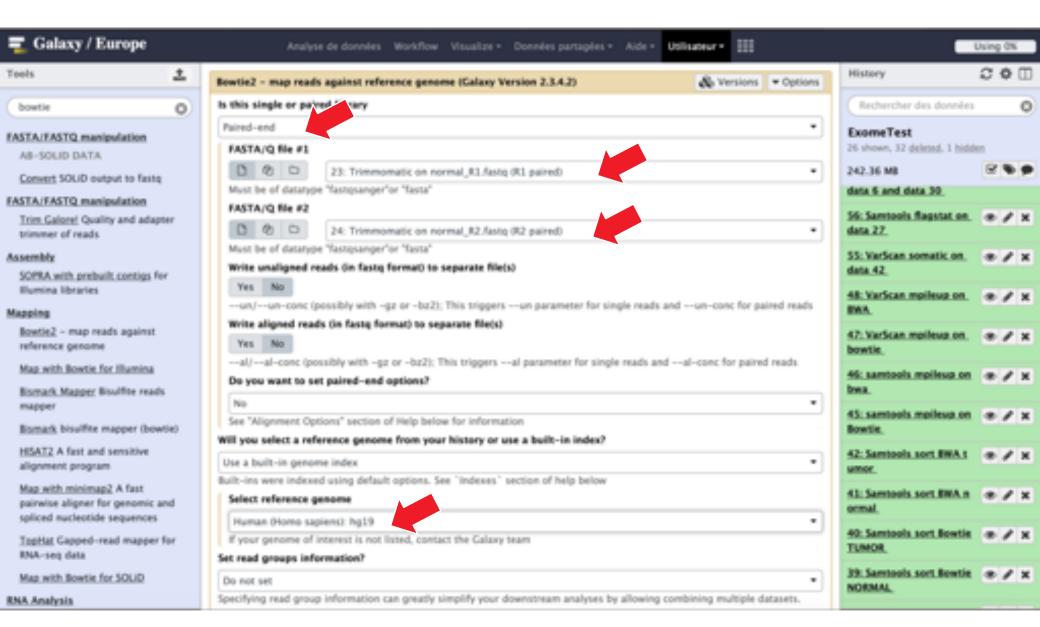
Vérifiez à nouveau les fichiers corrigés avec fastqc

Trimmomatic (fin)

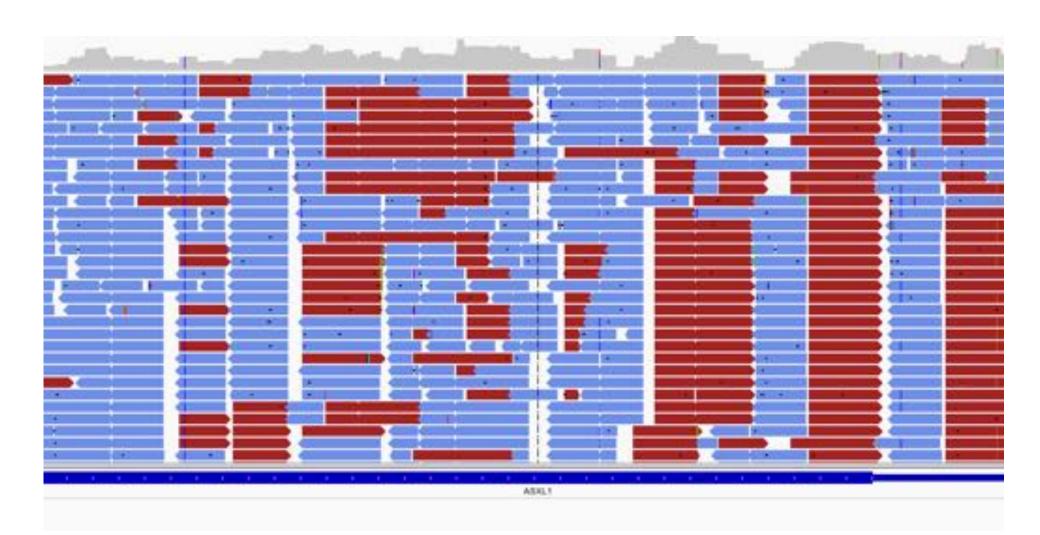
- Vérifiez le gain de qualité (fastqc d'un fastq)
- Eliminez les données « unpaired »



Bowtie



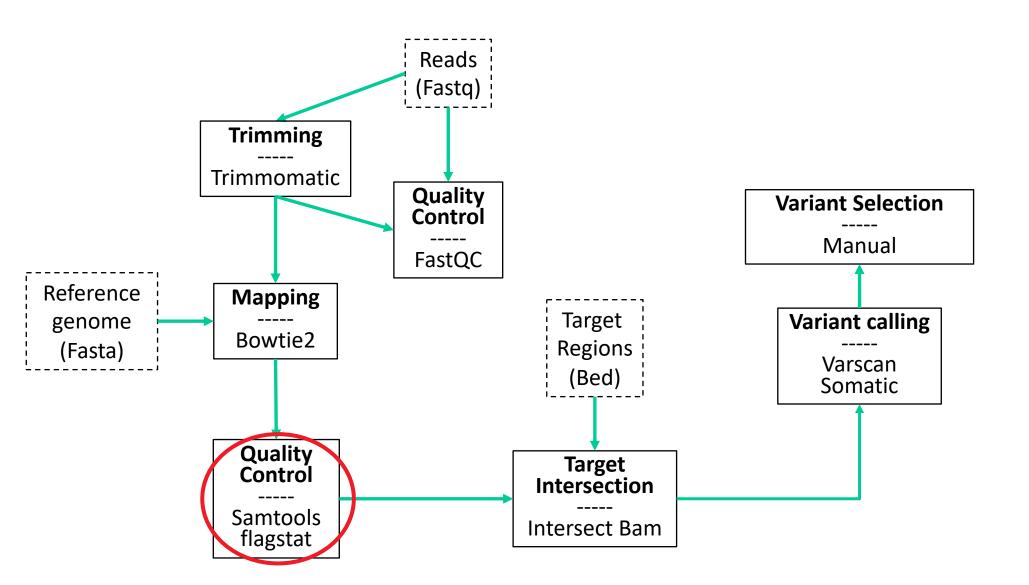
Reads alignés: le format BAM/SAM



BAM format

Rappel BAM:

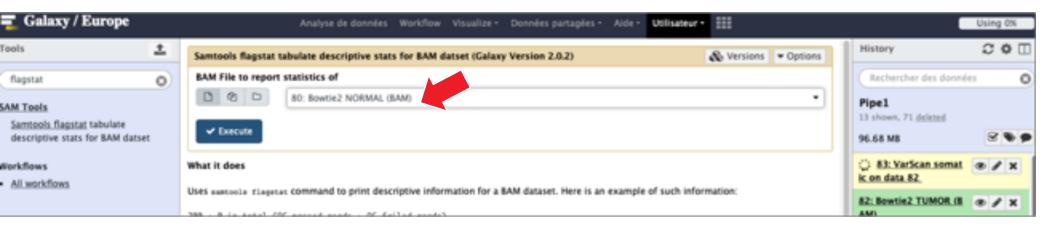
OBC	TD 1	CH . 1 43F	6034	DI - TI I I MITAIA		Billion				
@RG	ID:group1	SM: 1425_	_CD34	PL:ILLUMINA	LB:lib1	PU:uni	111			
@PG	ID:bwa PN:bwa	VN:0.7.1	12-r1039	CL:bwa mem -M -	-t 2 -A 2	-E 1 -	-R @RG\tID:	:group1\tSM:1425	_CD34\tPl	L:ILLUMINA\tLB:lib1\tPU:unit1 /root/myd
ERR16633	8.13782800	83	chr13	32890449	60	101M	=	32890343	-207	GGGACTGAATTAGAATTCAAACAAATTTTCCAGCGCTT
ERR16633	8.13782800	163	chr13	32890343	60	75M	=	32890449	207	CACTAGCCACGTTTCGAGTGCTTAATGTGGCTAGTGGC
ERR16633	8.26716588	99	chr13	32890406	60	101M	=	32890553	222	AATGTTCCCATCCTCACAGTAAGCTGTTACCGTTCCAG
ERR16633	8.26716588	147	chr13	32890553	60	75M	=	32890406	-222	TTGCAGACTTATTTACCAAGCATTGGAGGAATATCGTA
ERR16633	8.27259961	99	chr13	32890496	60	101M	=	32890558	137	ACCTCAGTCACATAATAAGGAATGCATCCCTGTGTAAG
ERR16633	8.27259961	147	chr13	32890558	60	75M	=	32890496	-137	GACTTATTTACCAAGCATTGGAGGAATATCGTAGGTAA
ERR16633	8.63037998	99	chr13	32890496	60	101M	=	32890558	137	ACCTCAGTCACATAATAAGGAATGCATCCCTGTGTAAG
ERR16633	8.63037998	147	chr13	32890558	60	75M	=	32890496	-137	GACTTATTTACCAAGCATTGGAGGAATATCGTAGGTAA



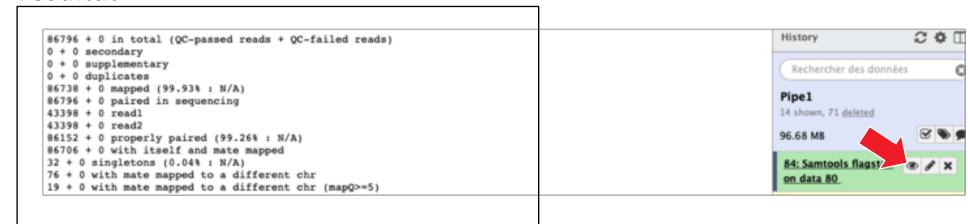
Samtools

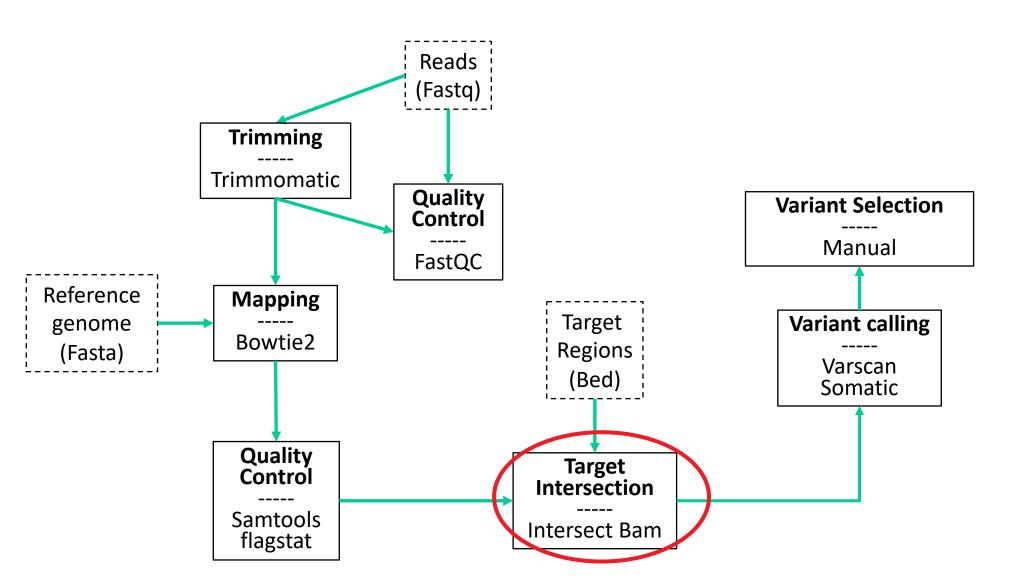
- La boîte à outils pour traiter les BAMs/SAMs
 - BAM <-> SAM
 - BAM <-> FASTQ
 - Tri de BAM
 - Indexation du BAM (création fichier .bai)
 - Obtenir un rapport sur le BAM (flagstat)

Samtools flatgstats



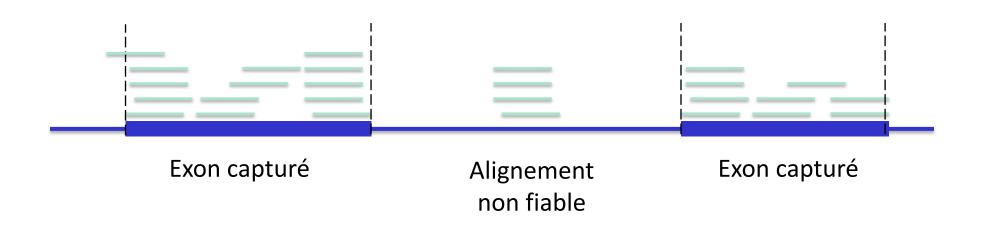
résultat



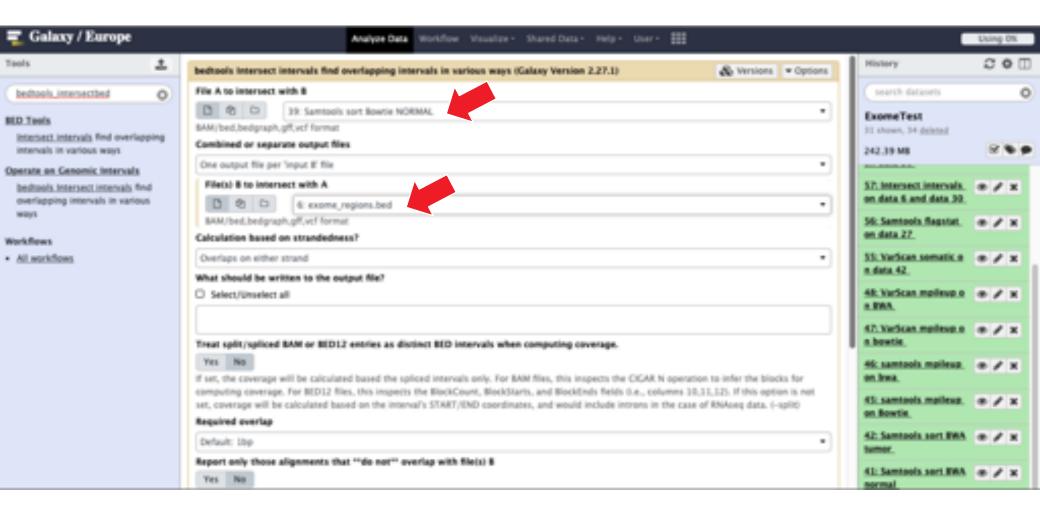


Target intersection

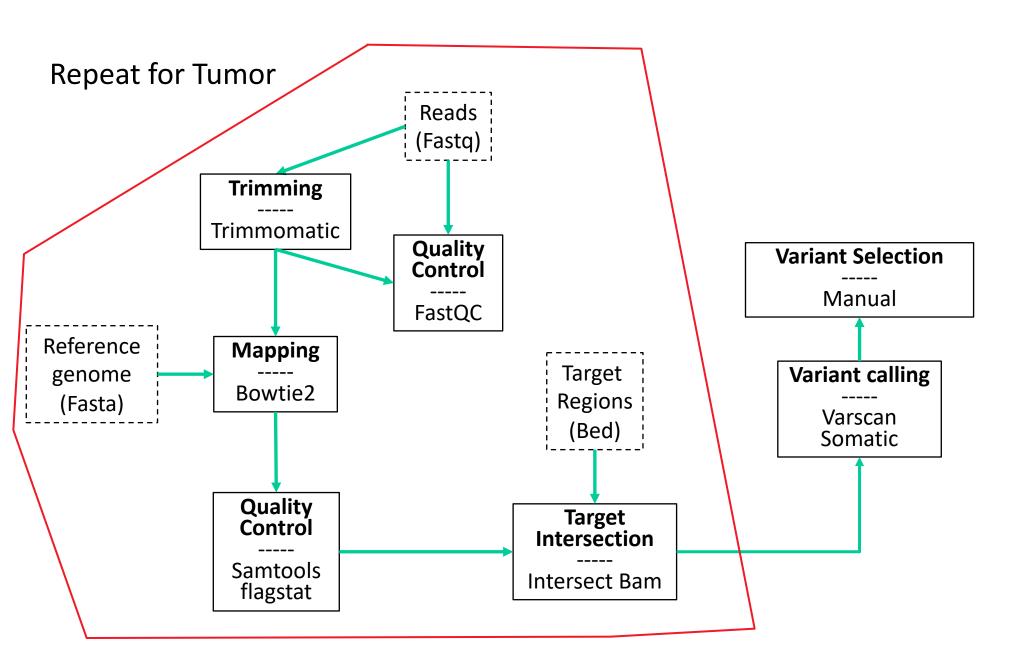
 Comparer l'alignement obtenu à la liste des positions visées par le protocole de capture



Bedtools intersect intervals



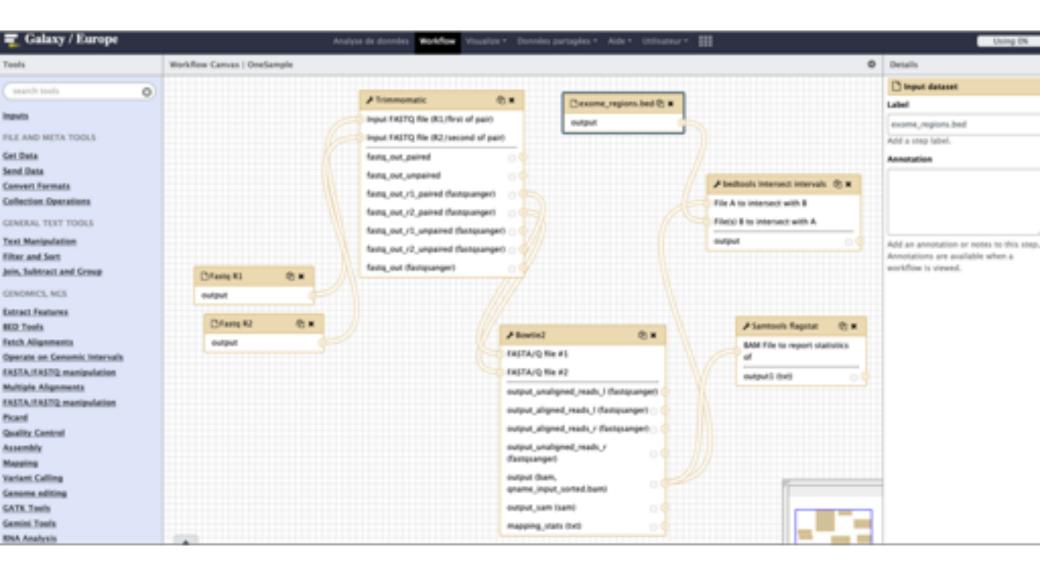
Vérifiez la réduction de taille du fichier BAM!



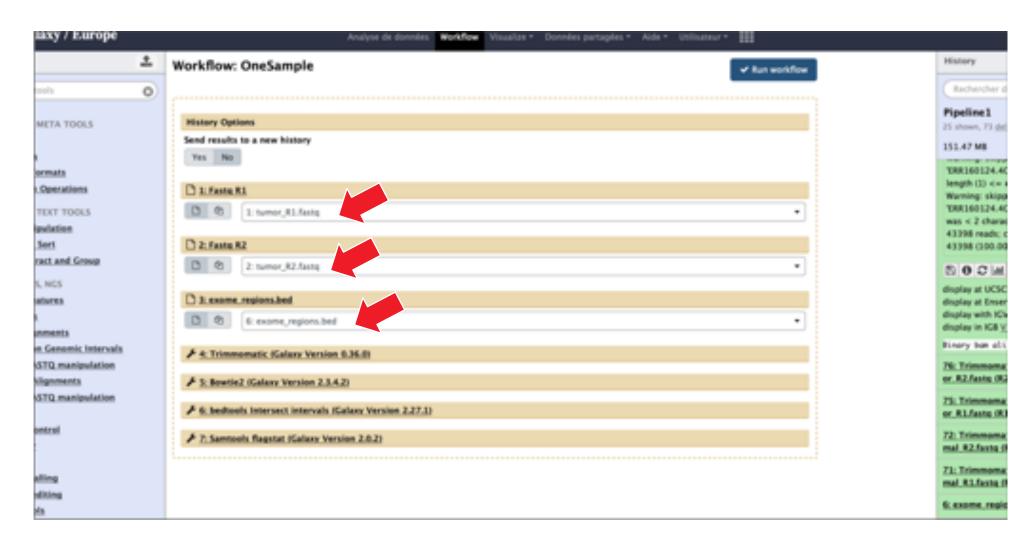
Extraire un workflow

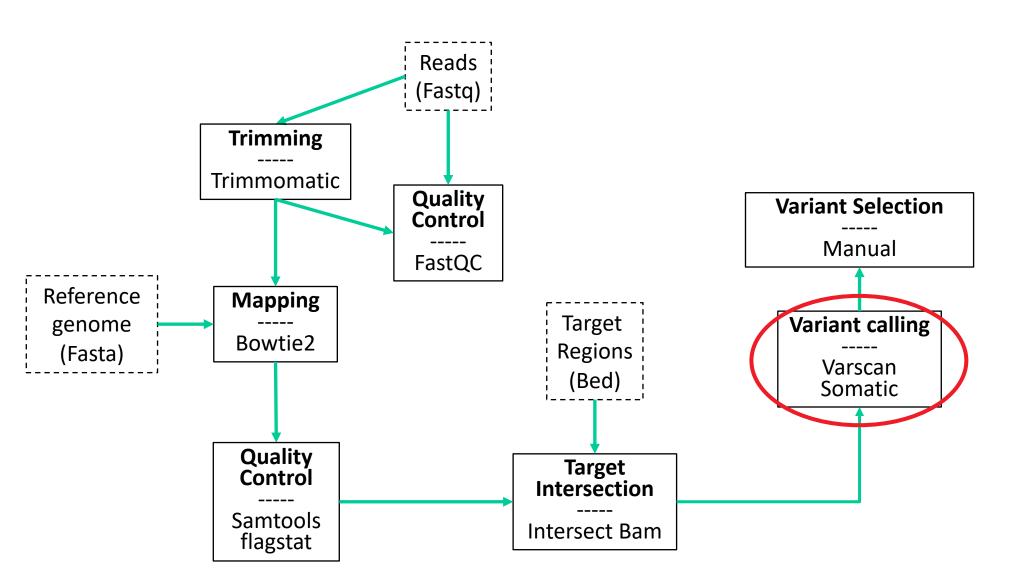


- Extraire workflow
- Le nommer
- Editer le workflow
- Choisir les données pertinentes (juste 2 fastq et regions.bed)
- Choisir les étapes de Trimmomatic à Intersect bed
- Enlever les data inutilisées (fastq tumor)
- Renommer les objets de façon générique (« sample » plutot que « normal »)
- Puis save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



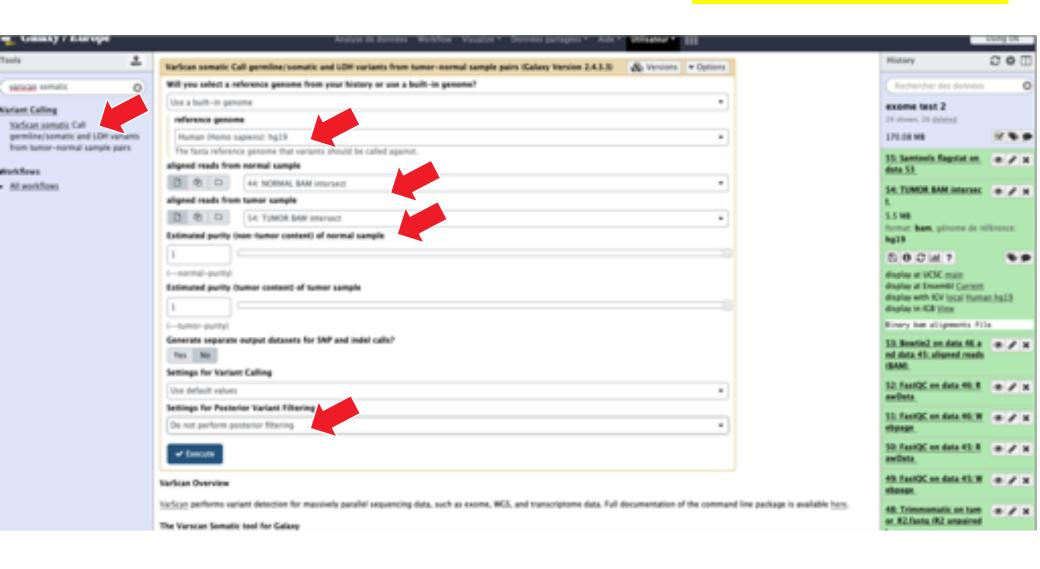
Maintenant lancez le workflow sur les données Tumor (run workflow)





Somatic variant calling: Varscan

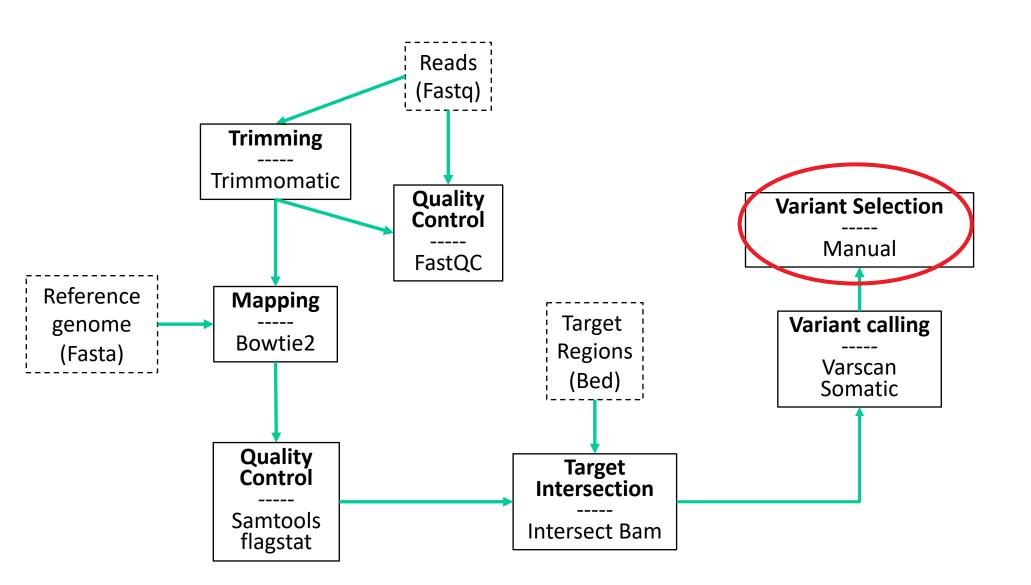
Attention: étape de 30min!



Check Varscan output

	chr17	18874685	c	CGGT	,	PASS	DP=32:55=3:55C=16:GPV=1:5PV=0.022989;INDEL	Pipeline1
FILE AND META TOOLS	chr17	18874720	c	G		PASS	DP=33;55=1;55C=0;CPV=1:3852e-19;5PV=1	24 shown, 72 deleted
Get Duta	chr17	18882991	T	A		PASS	DP=60;55=1;55C=0;CPV=1:035e-35;5PV=1	151.47 MB
Send Data	chr17	41256074	c	CA		PASS	DP=81;55=1;55C=1;CPV=0.0015196;5PV=0.63343;INDEL	
Convert Formats	chr17	73759304	G	T		PASS	DP=36;55=1;55C=0;CPV=2:2598e-21;5PV=1	88: VarScan somatic on. 🐵 🖋 🗙
Collection Operations	chr29	6374813	T	c	,	PASS	DP=33;55=1;55C=0;CPV=2:8029e-05;5PV=0:8425	data 82 and data 80.
GENERAL TEXT TOOLS	chr29	7550844	G	A		PASS	DP=44;55=1;55C=4;CPV=2:3358e-10;5PV=0:35332	153 lines, 113 comments format: vcf, génome de néférence:
Text Manipulation	chr29	36504365	c	T		PASS	OP=34;55=1;55C=1;CPV=5.1914e-07;5PV=0.63966	hg19
Filter and Sort	chr1	10596341	c	T		PIASS	DP=44;35=1;55C=2;CPV=7;4746e-10;5PV=0:53262	7
	chr1	160251792	A	G		PASS	DP=37;55=1;55C=0;CPV=5.1339e-06;5PV=0.87856	Starting variant calling _
Join, Subtract and Group	chri	167082869	G	A		PASS	DP=71;55=1;55C=8;CPV=2:0173e-19;5PV=0:13252	Calling variants for contig: chr10
GENOMICS, NGS	chr1	167095163	G	c		PASS	DP=52;55=1;55C=5;CPV=6.8522e-13;5PV=0.28624	Contig chr10 finished.
Extract Features	chrli	167097739	c	A		PASS.	9P=64;35=1;55C=3;CPV=4.3049e-14;5PV=0.44587	
BED Tools	chr1	214788427	c	T		PASS	15:35=1:35C=1:CPV=8:5784e-10:SPV=0:66234	Calling variants for contig: chr11
Fetch Alignments	ehr1	214802553	CT	C		PASS	ur=83;50MATIC;55=2;55C=18;CPV=1;5PV=0.015148;NOEL	Contig chr11 finished.
Operate on Cenomic Intervals	chrl	214803969	G	c		PASS	DP=111;S0MATIC;S5=2;S5C=35;GPV=1;SPV=0.00029013	Coming Cor 11 minoring.
FASTA/FASTQ manipulation	chr1	214804041	c	A		PASS	DP=65;35=1;35C=0;CPV=2:7963e-08;5PV=0:9934	Calling variants for contig:
Multiple Alignments	chrli	214811174	G	A		PASS	OP=76;55=1;55C=0;CPV=3:6183e-12;5PV=0:99124	chr11_gl000202_random
FASTA/FASTQ manipulation	chr1	214811244	c	C		PASS	DP=120,55=1,55C=0,CPV=1.7875e-19,5PV=0.92629	Calling variants for contig: chr12
Ficand	chr1.	214813487	A	G		PASS	DP=291;55=1;55C=3;GPV=1.3526e-38;5PV=0.47444	Contig chr12 finish
Quality Control	chrl	214813782	Α.	G	,	PASS	DP=108;55=1;55C=0;GPV=1:7692e-19;5PV=0:98472	Constitution of the Consti
Assembly	chr1	214813941	c	G		PASS	DP=86;35=1;55C=4;CPV=8.038e-16;5PV=0.34707	5 0 CM 7 % 9
	chrli	214814125	G	A.		PASS	DP+80;55=1;55C+0;CPV+1;2414e-11;5PV+0;85982	display at UCSC main
Mapping	chr1	214814582	G	A		PASS	OP=226;55=1;55C=5;CPV=3:0361e-32;5PV=0:28302	display with ICV local Human.hg12
Variant Calling	chr1	214814733	T	G		PASS	DP=244;55=1;55C=0;GPV=2:27499e-40;5PV=0:97323	display at RViewer main
France ofities								

How many variants? Somatic variants found?



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 <u>local</u> ». (will automatically connect with our local IGV session)
- Visualize somatic events.

IGV view





Variant annotation with VEP

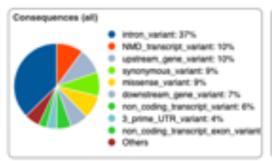
- Download the Varscan VCF file
- Go to https://www.ensembl.org/Tools/VEP
- Select GRCh37.p13 (=hg19)
- Launch VEP
- Sort results by impact

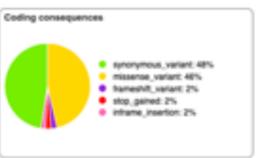
Variant Effect Predictor results @

Job details III

Summary statistics III

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



Į.	Showhide	Showfride columns (2 hidden)											
	Spicaded rariant	Location	Afele	Consequence	Impact	Symbol	Gene	Feature type	Feature	Biotype	Exon	Intron	eDN/ posit
		1:248058779- 248058779	A	frameshilt_variant	HIGH	ORZWS	ENSG00000238243	Transcript	ENST00000060358	protein_coding	1/1		891-6
		1:248059779- 248059779	A	frameshift_variant	HIGH	OR2W3	ENSG00000238243	Transcript	ENST00000537741	protein_coding	3/3		1148-
		3:121416308- 121416308	T	stop_gained	HIGH	GOLG81	ENSG00000173230	Transcript	ENST00000340645	protein_coding	13/22		3173
		3:121416308- 121416308	T	stop_pained	HIGH	GOL081	ENSG00000173230	Transcript	ENST000000999667	protein_coding	13/22		3173
		3:121416306-	T	stop_gained	HIGH	GOLGB1	ENSG00000173230	Transcript	ENST00000489400	profein_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 »

Galaxy: récupérer des données partagées

- Menu « Données partagées »
- Histories
- Choisir History « ... IFSBM ... dgauth»
- Import History