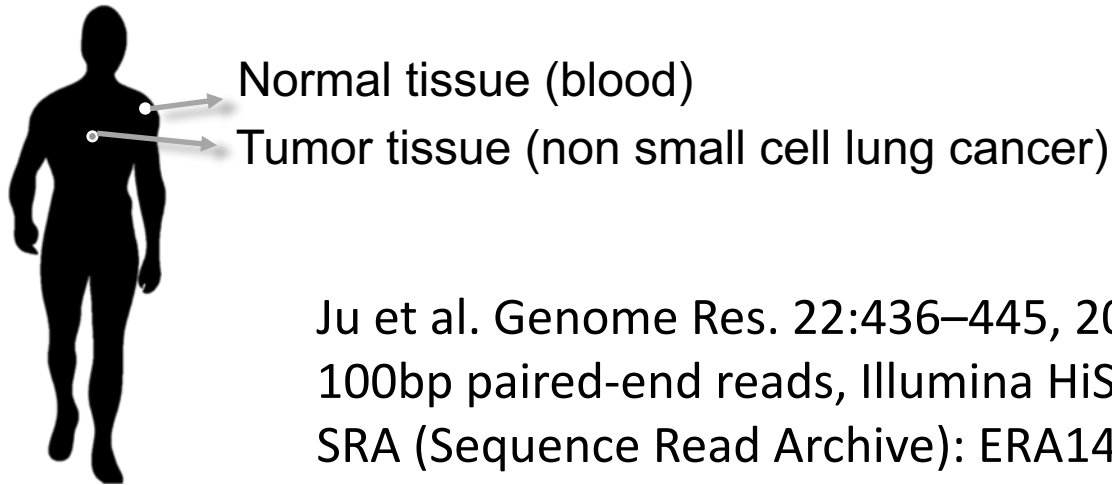


Réalisation d'un Pipeline d'analyse d'exome

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: ~ 15 Go tumor ; ~ 7 Go blood
Complete analysis processing Hme: $\sim 20h$
- **Fastq files restricted to a few regions ($\sim 112kb$)**
- **to limit processing time**

Chargez vos données

Galaxy / Europe

Analyse de données Workflow Visualize Données partagées Aide Utilisateur Using OR

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

"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." – Prof. Stephen Hawking

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:

(alternativement: à partir de données partagées)

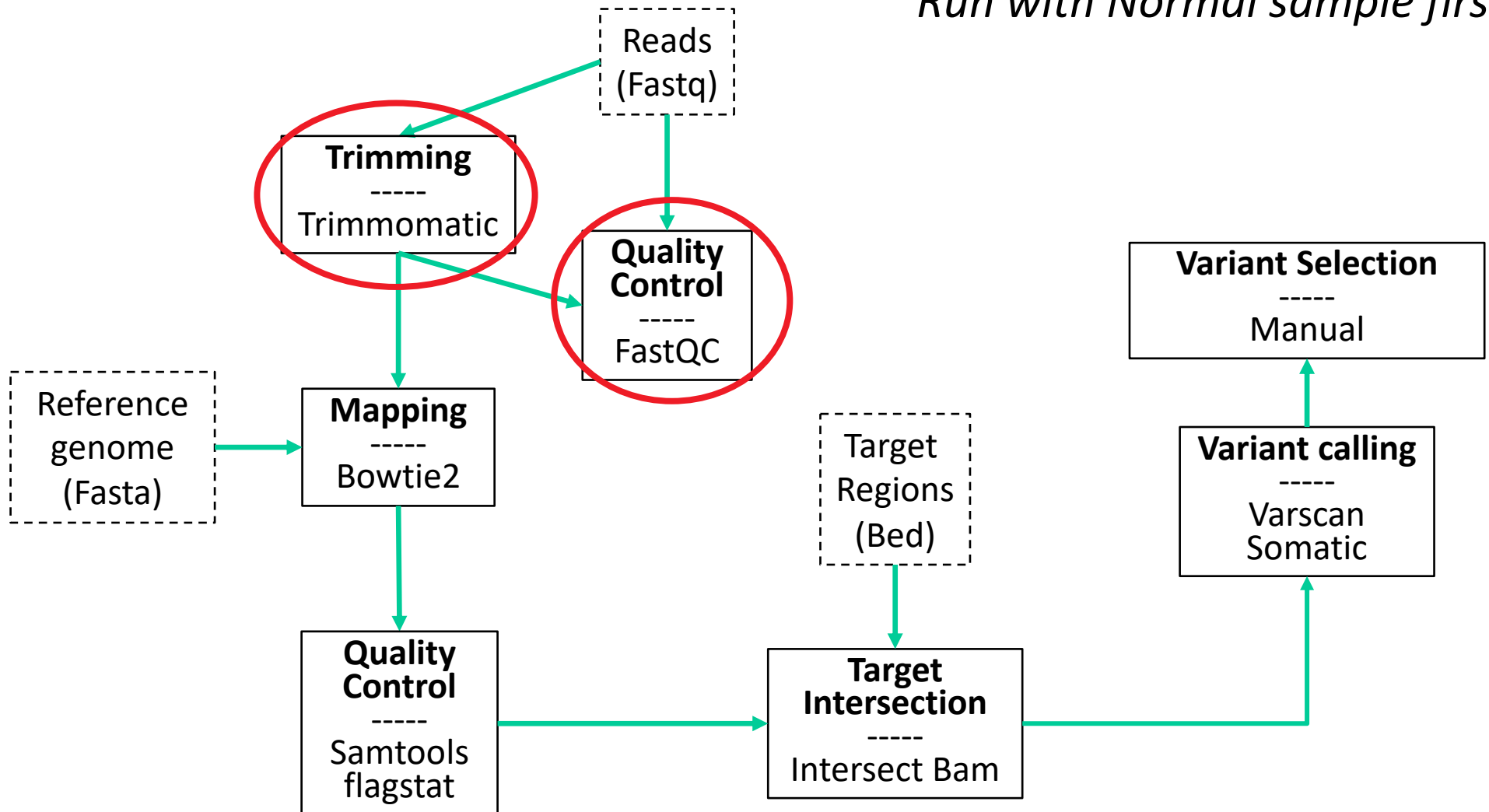
- Menu « Données partagées »
- Histories
- Choisir History « ... IFSBM ... »
- Import History

Fera apparaître:

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>	  

A simplified Variant Pipeline

Run with Normal sample first



fastqc

The screenshot shows the Galaxy web interface with the FastQC tool selected. The left sidebar contains navigation links for Tools, FASTA/FASTQ manipulation, Quality Control, and Mapping. The main panel displays the FastQC configuration page, which includes a search bar for 'fastqc', a list of input files, and options for contaminant list and submodules. Two red arrows highlight the input field and the file list.

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) | Versions | Options

Short name | **Input** | **Output**

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
CAACCAGAACGCGCATACGA

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

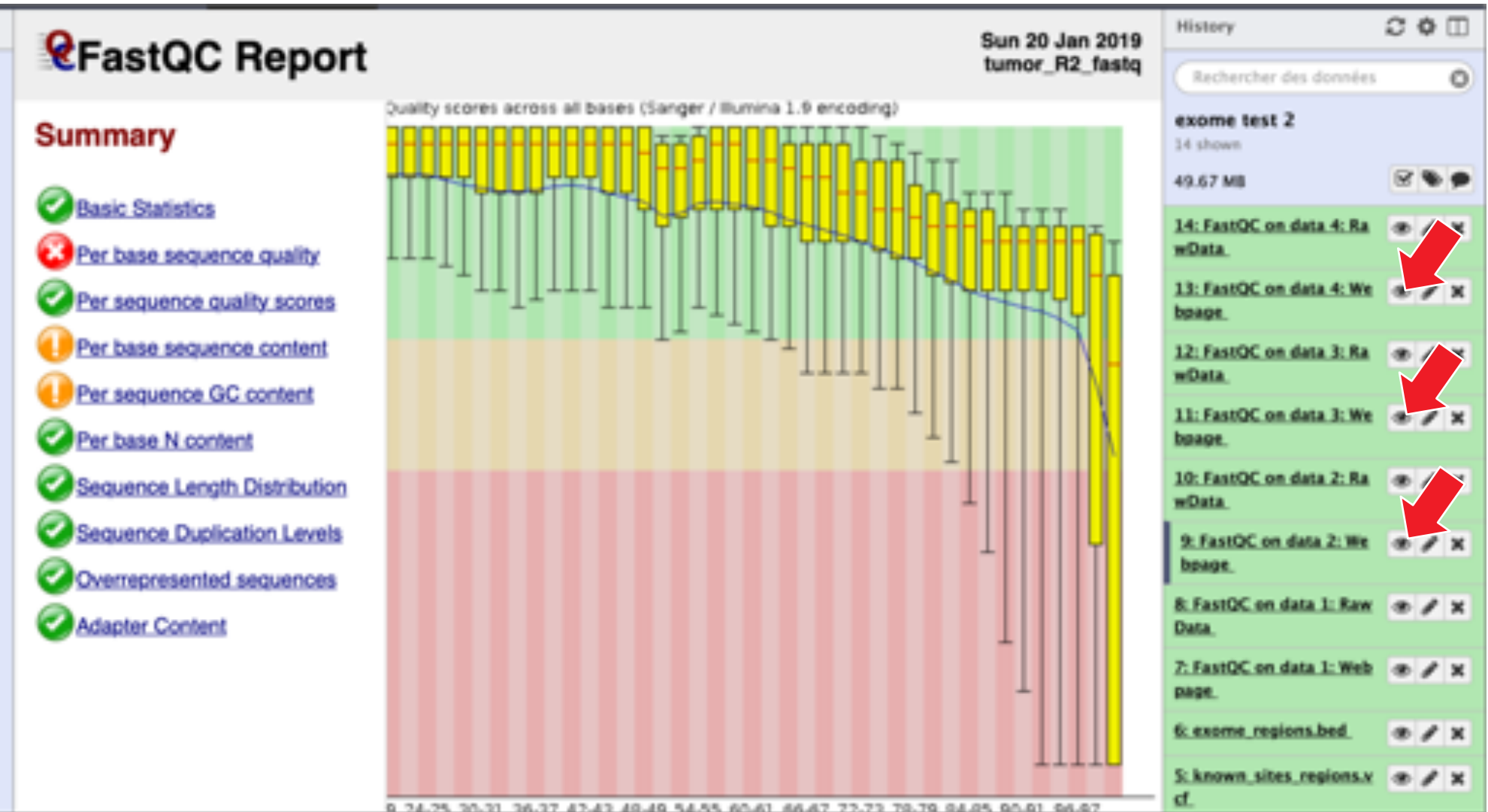
Execute

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



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

Trimmomatic

Galaxy / Europe

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

Tools

trimmomatic

FASTA/FASTQ manipulation

fastq - 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

Show 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, 11 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

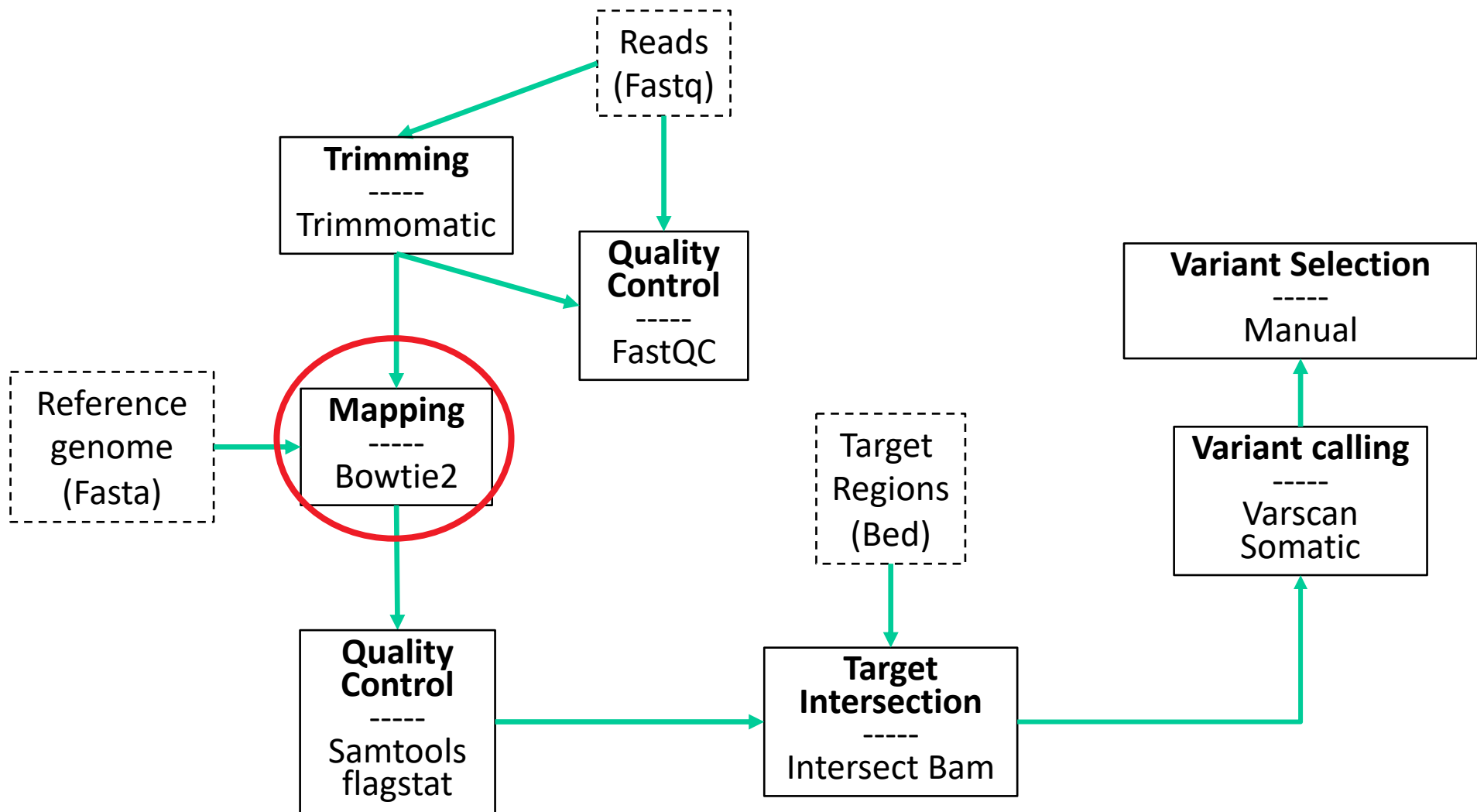
3: normal_R2.fastq

2: tumor_R2.fastq

Vérifiez à nouveau les fichiers corrigés avec fastqc

Trimmomatic (fin)

- Vérifiez le gain de qualité (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 Capped-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 39.

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.

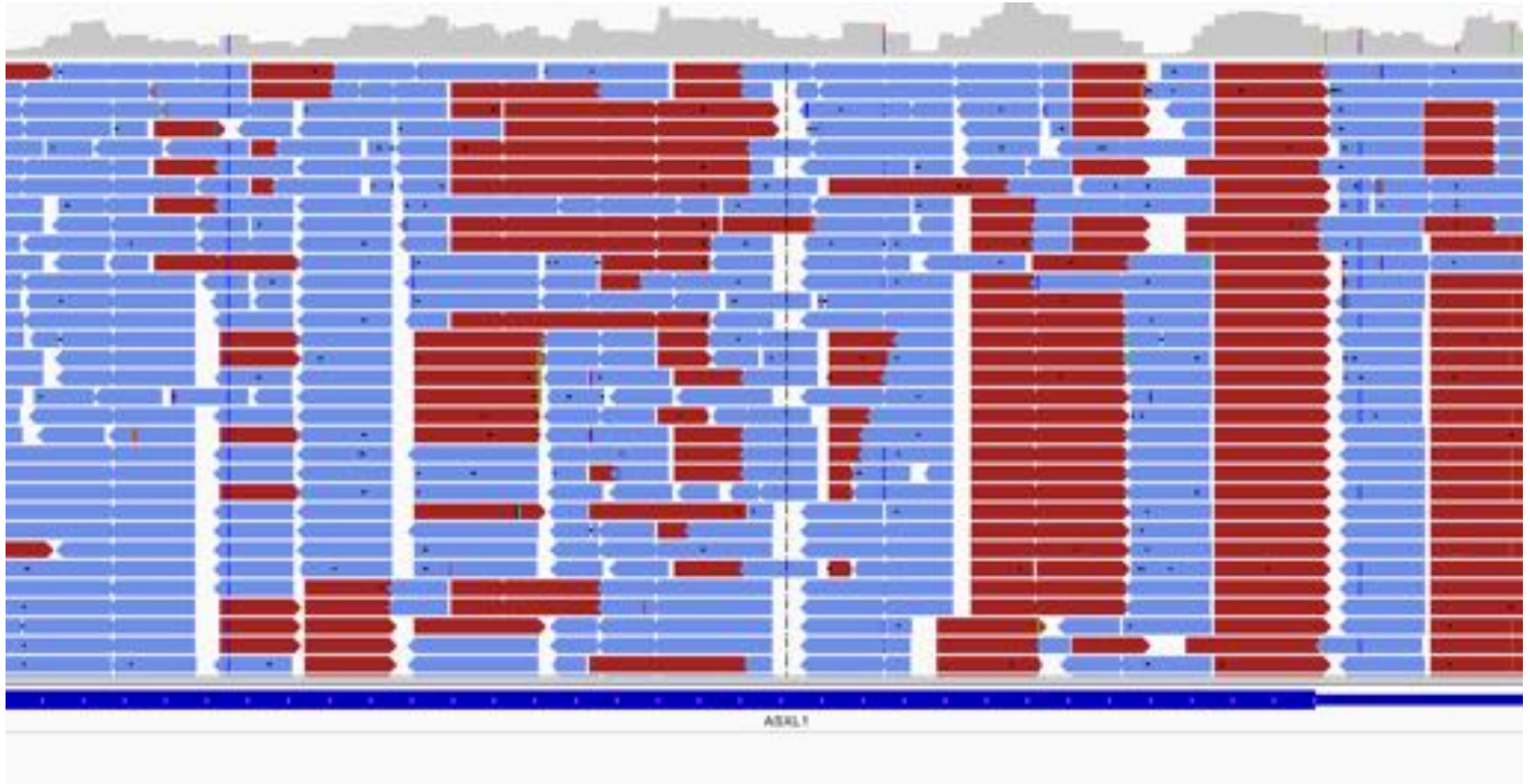
42: Samtools sort BWA.tumor.

41: Samtools sort BWA.normal.

40: Samtools sort Bowtie TUMOR.

39: Samtools sort Bowtie NORMAL.

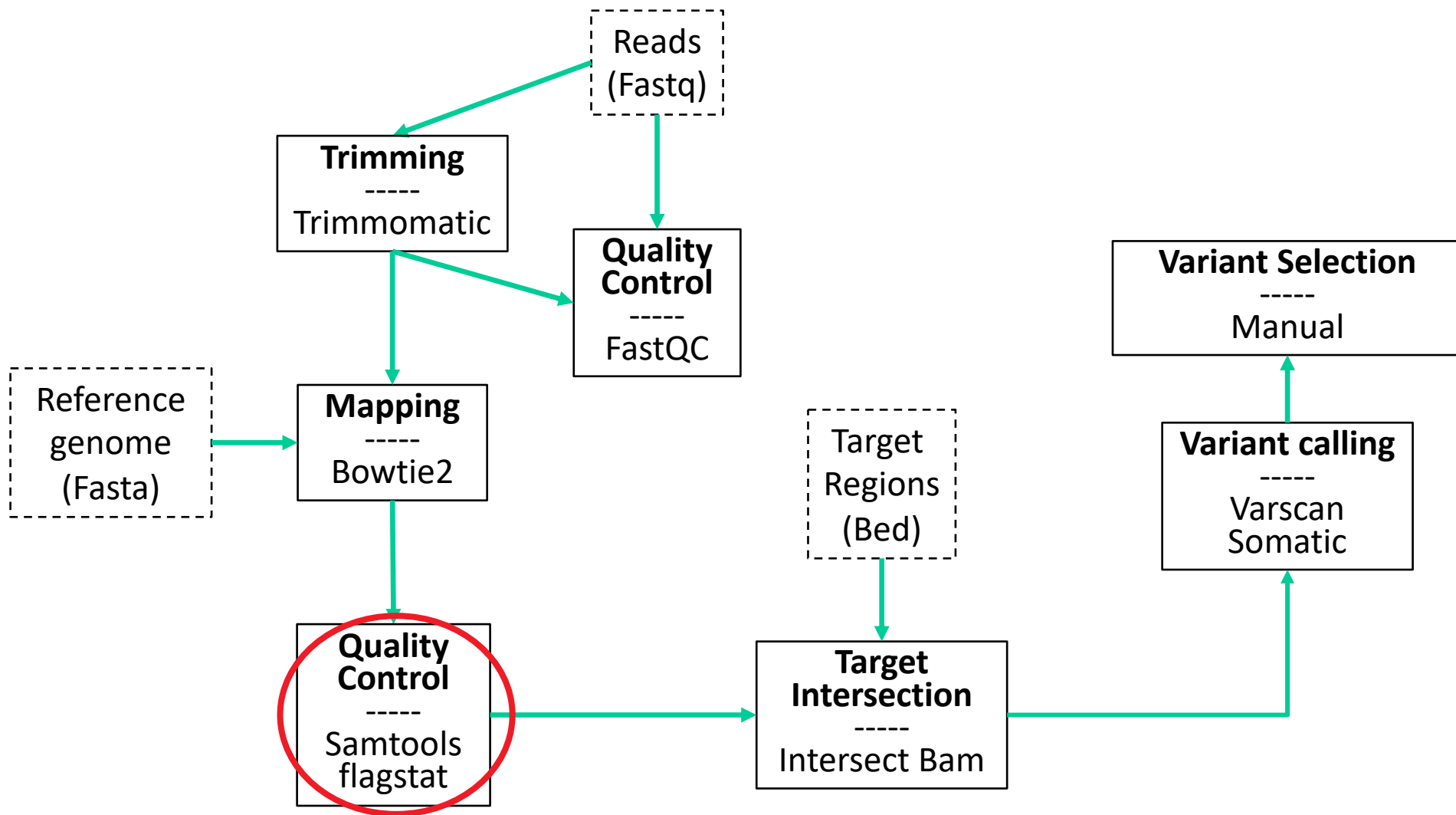
Reads alignés: le format BAM/SAM



BAM format

Rappel BAM:

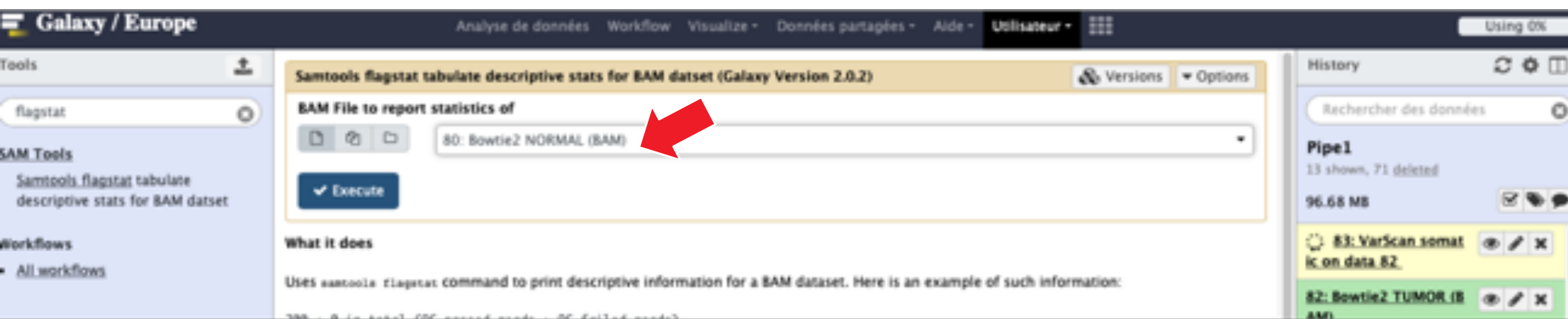
```
@RG      ID:group1      SM:1425_CD34      PL:ILLUMINA      LB:lib1 PU:unit1
@PG      ID:bwa      PN:bwa      VN:0.7.12-r1039 CL:bwa mem -M -t 2 -A 2 -E 1 -R @RG\tID:group1\tSM:1425_CD34\tPL:ILLUMINA\tLB:lib1\tPU:unit1 /root/myd
ERR166338.13782800      83      chr13      32890449      60      101M      =      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)

Samtools flatgstats



Galaxy / Europe

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

Tools

flagstat

SAM Tools

Samtools flagstat tabulate descriptive stats for BAM dataset

Workflows

All workflows

Samtools flagstat tabulate descriptive stats for BAM dataset (Galaxy Version 2.0.2)

BAM File to report statistics of

80: Bowtie2 NORMAL (BAM)

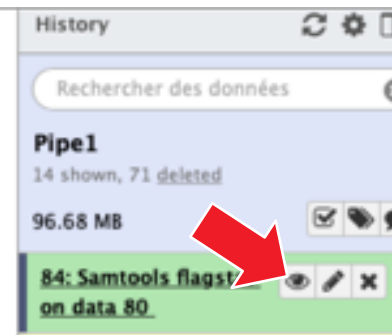
Execute

What it does

Uses samtools flagstat command to print descriptive information for a BAM dataset. Here is an example of such information:

résultat

```
86796 + 0 in total (QC-passed reads + QC-failed reads)
0 + 0 secondary
0 + 0 supplementary
0 + 0 duplicates
86738 + 0 mapped (99.93% : N/A)
86796 + 0 paired in sequencing
43398 + 0 read1
43398 + 0 read2
86152 + 0 properly paired (99.26% : N/A)
86706 + 0 with itself and mate mapped
32 + 0 singletons (0.04% : N/A)
76 + 0 with mate mapped to a different chr
19 + 0 with mate mapped to a different chr (mapQ>=5)
```



History

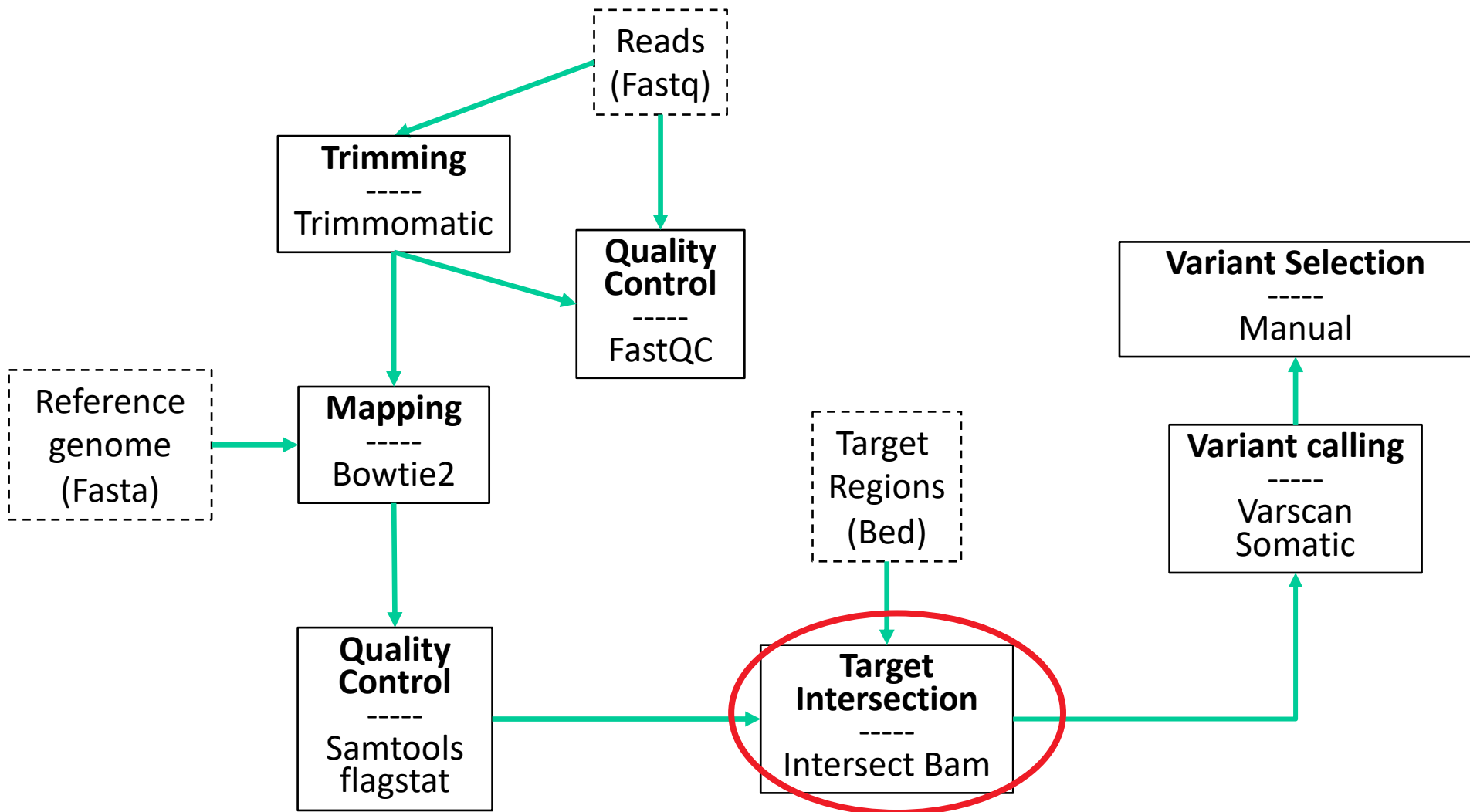
Rechercher des données

Pipe1

14 shown, 71 deleted

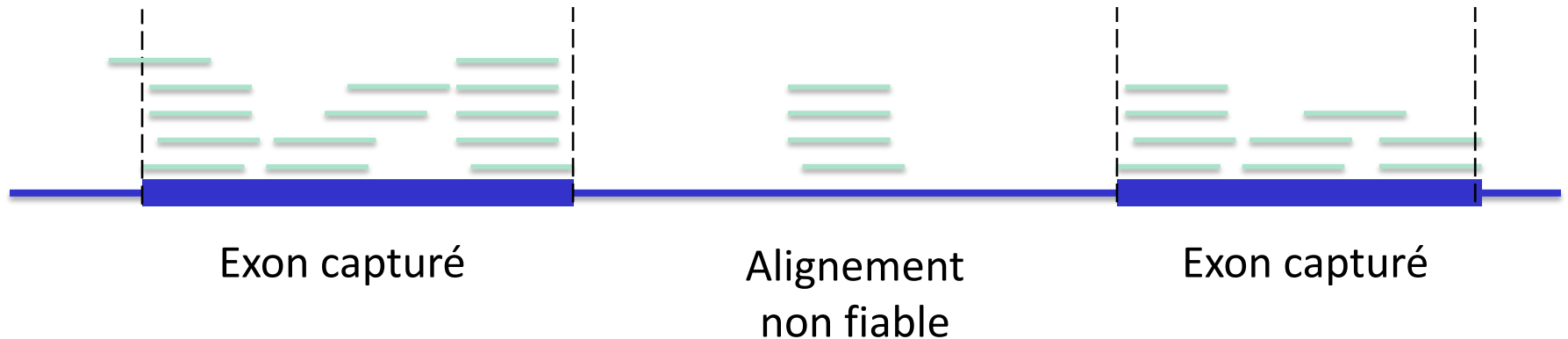
96.68 MB

84: Samtools flagstat on data 80



Target intersection

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



Bedtools intersect

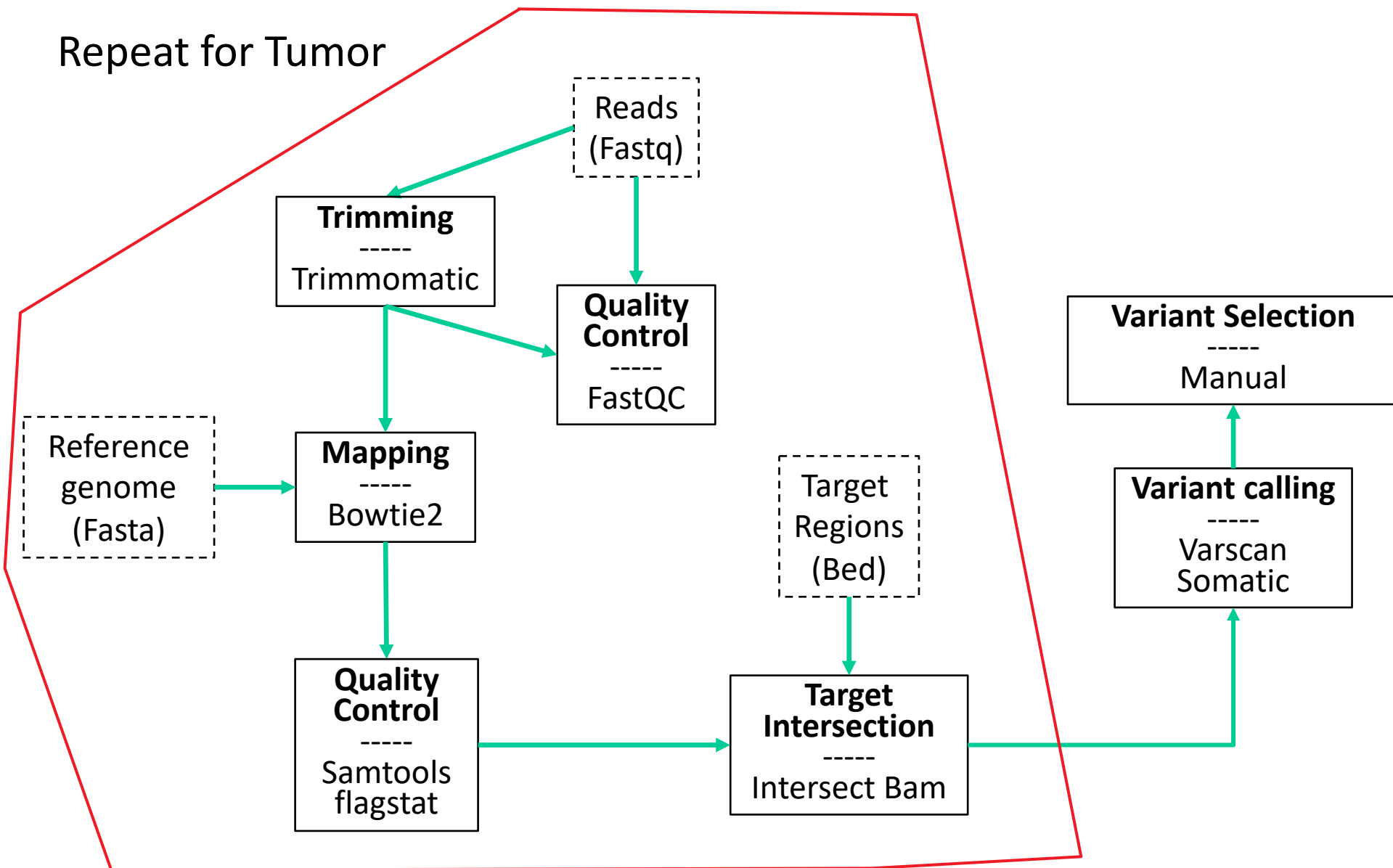
The screenshot displays the Galaxy web interface for the 'bedtools intersect' tool. The tool's title is 'bedtools intersect intervals find overlapping intervals in various ways (Galaxy Version 2.27.1)'. The configuration is as follows:

- File A to intersect with B:** 39: Samtools sort Bowtie NORMAL (indicated by a red arrow).
- File(s) B to intersect with A:** 6: exome_regions.bed (indicated by a red arrow).
- 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 sidebar shows a history of jobs, including '57: Intersect intervals on data 6 and data 39'.


Vérifiez la réduction de taille du fichier BAM!

Repeat for Tumor



Extraire un workflow



- Extraire workflow
- Le nommer
- 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

Tools Workflow Canvas | OneSample

Search tools

Inputs

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

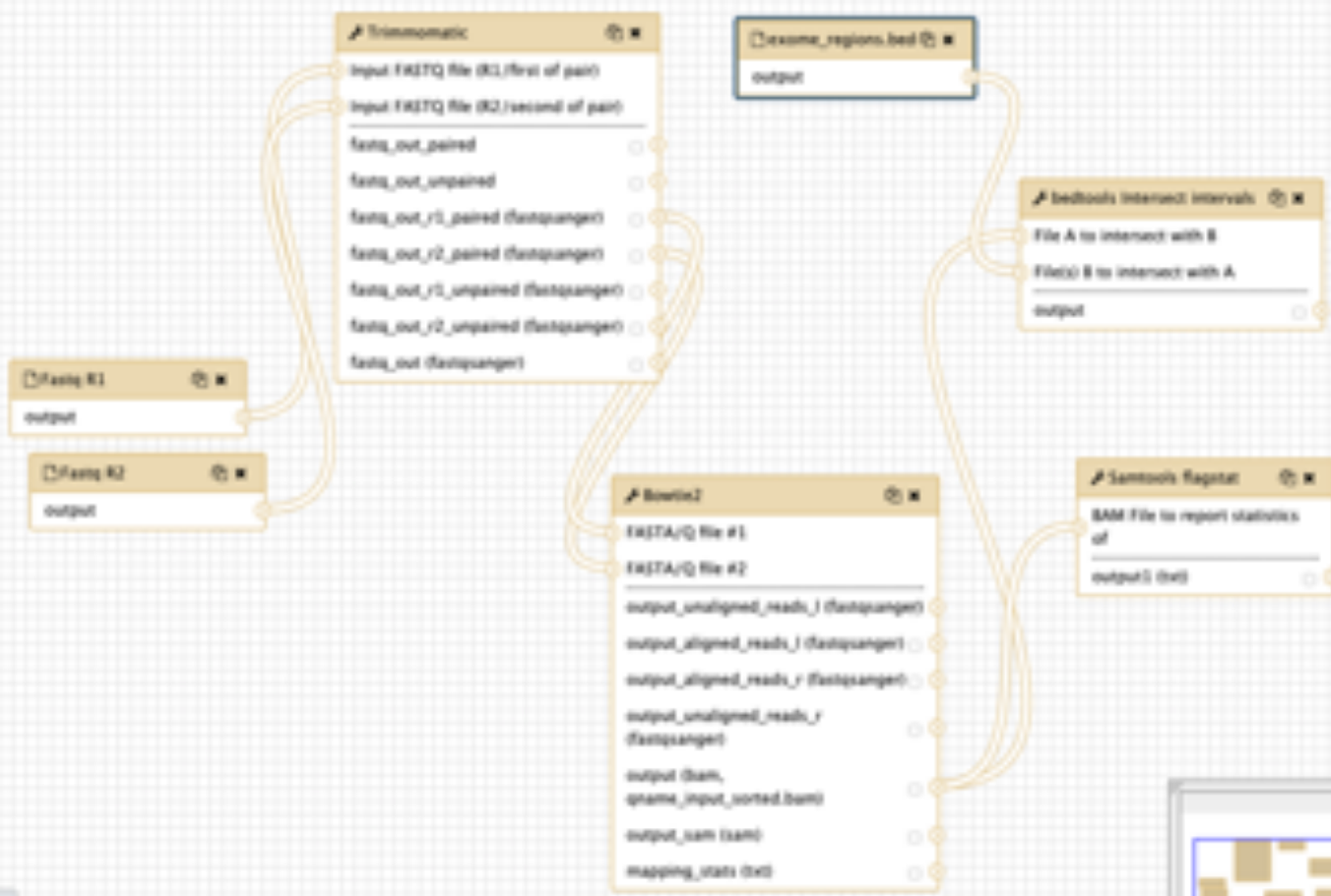
Variant Calling

Genome editing

GATK Tools

Genomic Tools

RNA Analysis



Details

Input dataset

Label

exome_regions.bed

Add a step label.

Annotation

Add an annotation or notes to this step.
Annotations are available when a workflow is viewed.

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

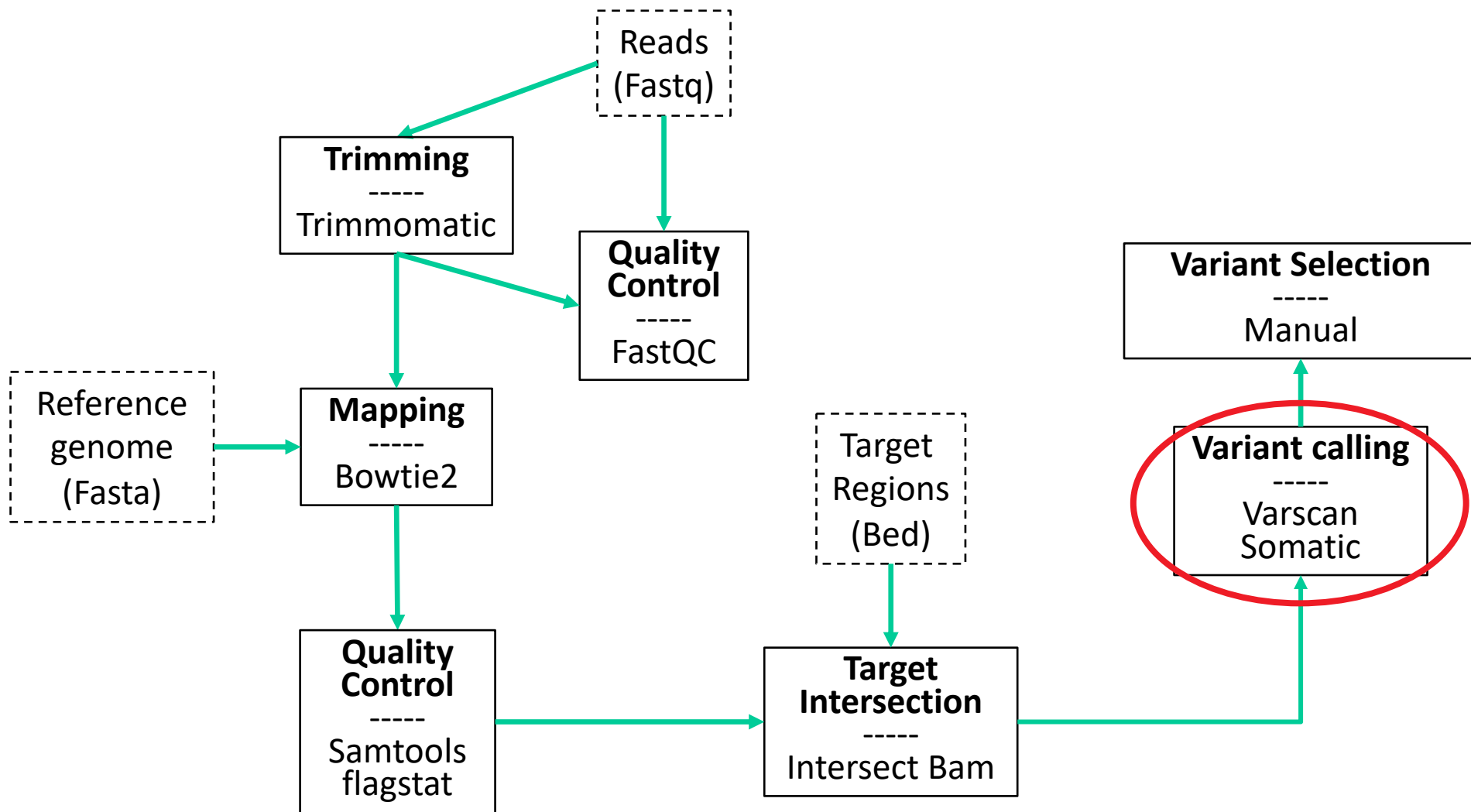
The screenshot displays the Galaxy web interface for a workflow named "Workflow: OneSample". The interface includes a top navigation bar with links for "Analyse de données", "Workflow", "Visualize", "Données partagées", "Aide", and "Utilisateur". A left sidebar lists various tool categories like "META TOOLS", "Formats", "Operations", "TEXT TOOLS", "Manipulation", "Sort", "Fact and Group", "NCS", "Sequences", "Annotations", "Genomic Intervals", "STQ manipulation", "Alignments", "STQ manipulation", "Control", "Calling", "Editing", and "Jobs".

The main content area shows the workflow steps:

- History Options**
 - Send results to a new history:
- 1: Fastq R1**
 - Input field:
- 2: Fastq R2**
 - Input field:
- 3: exome_regions.bed**
 - Input field:
- 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)**

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

The right sidebar shows a "History" panel with a search bar and a list of datasets. The first dataset is "Pipeline 1" (25 shown, 73 deleted), with a size of 151.47 MB. Below it, a list of datasets is shown, including "TERR160124.4C", "length (1)", "Warning: skip", "TERR160124.4C", "was < 2 charac", "43398 readic", "43398 (100.00)", "display at UCSC", "display at Enser", "display with IGV", "display in IGV V", "Binary bam alt", "76: Trimmoma", "or_R2.fastq (R2)", "75: Trimmoma", "or_R1.fastq (R1)", "72: Trimmoma", "maf_R2.fastq (R2)", "71: Trimmoma", "maf_R1.fastq (R1)", and "6: exome_regio".



Somatic variant calling: Varscan

Attention: étape de 30min!

The screenshot displays the Galaxy web interface for the Varscan tool. The left sidebar shows the 'Tools' section with 'Varscan' selected under 'Variant Calling'. The main panel shows the 'Varscan somatic Call germline/somatic and LOH variants from tumor-normal sample pairs (Galaxy Version 2.4.3.3)' configuration page. The page includes several input fields and dropdown menus, with red arrows highlighting specific elements: the 'Varscan' tool in the sidebar, the 'reference genome' dropdown (set to 'Human (Homo sapiens): hg19'), the 'aligned reads from normal sample' and 'aligned reads from tumor sample' dropdowns (both set to '82: Bowtie2 TUMOR (BAM)'), and the 'Settings for Posterior Variant Filtering' dropdown (set to 'Do not perform posterior filtering'). The right sidebar shows the 'History' section with a list of datasets, including '83: Varscan.somatic on data 82'.

Galaxy / Europe

Analyse de données Workflow Visualize Données partagées Aide Utilisateur Using DR

Tools

Varscan

Variant Calling

Varscan.somatic Call germline/somatic and LOH variants from tumor-normal sample pairs

Varscan.mpileup for variant detection

Varscan.copynumber Determine relative tumor copy number from tumor-normal pileups

Varscan for variant detection

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

82: Bowtie2 TUMOR (BAM)

aligned reads from tumor sample

82: Bowtie2 TUMOR (BAM)

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

History

Rechercher des données

Pipe1

13 shown, 71 deleted

96.68 MB

83: Varscan.somatic on data 82

82: Bowtie2 TUMOR (BAM)

80: Bowtie2 NORMAL (BAM)

77: Trimmomatic on tumor_R1.fastq (R1 unpaired)

76: Trimmomatic on tumor_R2.fastq (R2 paired)

75: Trimmomatic on tumor_R1.fastq (R1 paired)

72: Trimmomatic on normal_R2.fastq (R2 paired)

71: Trimmomatic on normal_R1.fastq (R1 paired)

6: exome_regions.bed

Check Varscan output

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

Genomic edition

chr17	18874685	.	C	CCGT	.	PASS	DP=32;SS=1;SSC=16;CPV=1;SPV=0.021989;INDEL
chr17	18874720	.	C	G	.	PASS	DP=33;SS=1;SSC=0;CPV=1.3852e-19;SPV=1
chr17	18882991	.	T	A	.	PASS	DP=60;SS=1;SSC=0;CPV=1.035e-35;SPV=1
chr17	41256074	.	C	CA	.	PASS	DP=81;SS=1;SSC=1;CPV=0.0015196;SPV=0.63343;INDEL
chr17	73759304	.	G	T	.	PASS	DP=36;SS=1;SSC=0;CPV=2.2598e-21;SPV=1
chr19	6374813	.	T	C	.	PASS	DP=33;SS=1;SSC=0;CPV=2.8029e-05;SPV=0.8425
chr19	7550844	.	G	A	.	PASS	DP=44;SS=1;SSC=4;CPV=2.3358e-10;SPV=0.35332
chr19	36504365	.	C	T	.	PASS	DP=34;SS=1;SSC=1;CPV=5.1914e-07;SPV=0.63966
chr1	10596341	.	C	T	.	PASS	DP=44;SS=1;SSC=2;CPV=7.4746e-10;SPV=0.53262
chr1	160251792	.	A	G	.	PASS	DP=37;SS=1;SSC=0;CPV=5.1339e-06;SPV=0.87856
chr1	167082869	.	G	A	.	PASS	DP=71;SS=1;SSC=8;CPV=2.0173e-19;SPV=0.13252
chr1	167095363	.	G	C	.	PASS	DP=52;SS=1;SSC=5;CPV=6.8522e-13;SPV=0.28624
chr1	167097739	.	C	A	.	PASS	DP=64;SS=1;SSC=3;CPV=4.3049e-14;SPV=0.44587
chr1	214788427	.	C	T	.	PASS	DP=35;SS=1;SSC=1;CPV=8.5784e-10;SPV=0.66234
chr1	214802553	.	CT	C	.	PASS	DP=83;SOMATIC;SS=2;SSC=18;CPV=1;SPV=0.015148;INDEL
chr1	214803969	.	G	C	.	PASS	DP=111;SOMATIC;SS=2;SSC=35;CPV=1;SPV=0.00029013
chr1	214804041	.	C	A	.	PASS	DP=65;SS=1;SSC=0;CPV=2.7963e-08;SPV=0.9934
chr1	214811174	.	G	A	.	PASS	DP=76;SS=1;SSC=0;CPV=3.6183e-12;SPV=0.99124
chr1	214811244	.	C	G	.	PASS	DP=120;SS=1;SSC=0;CPV=1.7875e-19;SPV=0.92629
chr1	214813487	.	A	G	.	PASS	DP=291;SS=1;SSC=3;CPV=1.3526e-38;SPV=0.47444
chr1	214813782	.	A	G	.	PASS	DP=108;SS=1;SSC=0;CPV=1.7692e-19;SPV=0.98472
chr1	214813941	.	C	G	.	PASS	DP=86;SS=1;SSC=4;CPV=8.058e-16;SPV=0.34707
chr1	214814125	.	G	A	.	PASS	DP=80;SS=1;SSC=0;CPV=1.2414e-11;SPV=0.85982
chr1	214814582	.	G	A	.	PASS	DP=226;SS=1;SSC=5;CPV=3.0361e-32;SPV=0.28302
chr1	214814733	.	T	G	.	PASS	DP=244;SS=1;SSC=0;CPV=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, genome 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

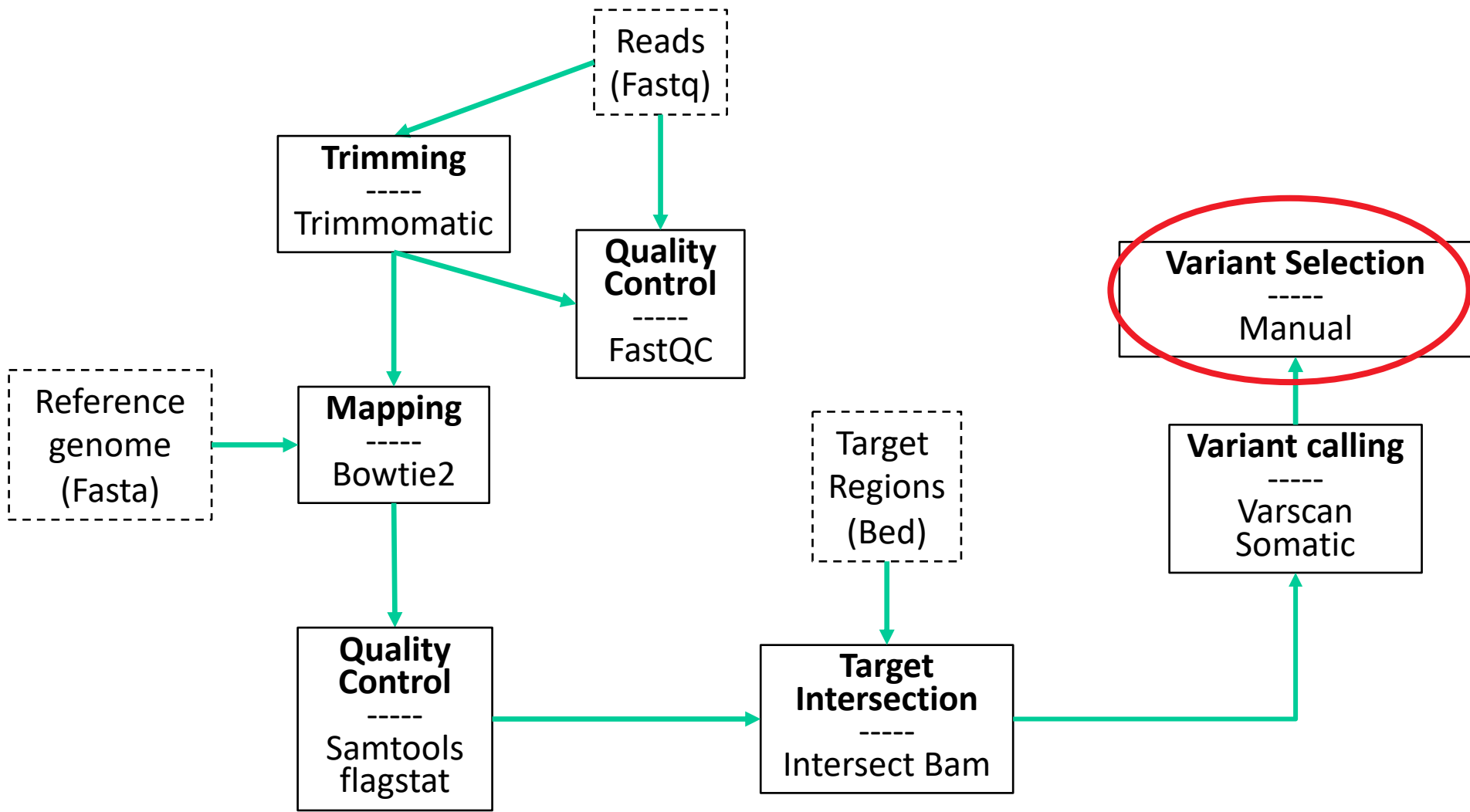
Contig chr12 finish

display at UCSC main

display with IGV local Human.hg19

display at RViewer main

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
- Download Normal and Tumor BAM files on your local computer
 - (select option « download bam_index »)
- Launch IGV with hg19 reference
- In IGV, load normal and tumor BAM files
- Visualize somatic events.

IGV view



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 ... »
- Import History