Development of a workflow for SNP detection in grapevine species: MAPHiTS.

MAPHiTS: Mappping Analysis Pipeline for High-Throughput Sequences

ALIMENTATION
AGRICULTURE
ENVIRONNEMENT

Overview

I. Background and objectives of the pipeline
II. Existing Tools
III. MAPHiTS Development Tools
IV. Integration of tools in Galaxy
V. Preliminary Results
VI. Perspectives

I. Background and objectives of the pipeline

I. Background and objectives of the pipeline

A **SNP** (Single-Nucleotide Polymorphism) is a DNA sequence variation. SNPs are used to detect complex traits such as diseases restistance or agronomical performance.

URGI team develops a **pipeline** *(MAPHiTS)* for **SNPs detection** from short reads. It's fully integrated in **Galaxy**.

Users : 50% biologists / 50% bioinformaticians

I. Background and objectives of the pipeline

- **Objectives:**

  Detect a set of SNPs between various species of Grape after mapping short reads against a reference genome.

- **Data:**
  - Project 1 : 6 species
  - Project 2 : 16 species

  Short reads are in paired-ends with 76, 101 or 114 bp *(Illumina GAII)*.

➡ Other projects are also in progress with others species.

II. Existing Tools

II. Existing Tools

- **FASTX-Toolkit:** tools for FASTA / FASTQ files preprocessing.
- **BWA / Bowtie:** mapping softwares, particularly suitable for short reads alignement (in paired-ends or single-ends) against one reference genome (Burrows – Wheeler Alignement tool).
- **SAMtools:** toolkit for working on the output SAM file (BWA, Bowtie, …).
- **VarScan:** software used to filter SNPs and small indels by:
  - coverage
  - number of variant
  - base quality
  - variant allele frequency
  - pValue

III. MAPHiTS Development Tools

III. MAPHiTS Development Tools

- **Optimization tools:**
  - BWA in parallel
  - SAM-to-BAM in parallel

Time Saving: 10x average

Exemple:
- Before: 11 -12 hours
- Now: 1 - 2 hours

- **Preprocessing tools:**
  - Remove duplicated short-reads
  - Remove short reads not in paired-ends
  - Remove short reads > ‘N’%
  - Remove informations in each FASTA file header

III. MAPHiTS Development Tools

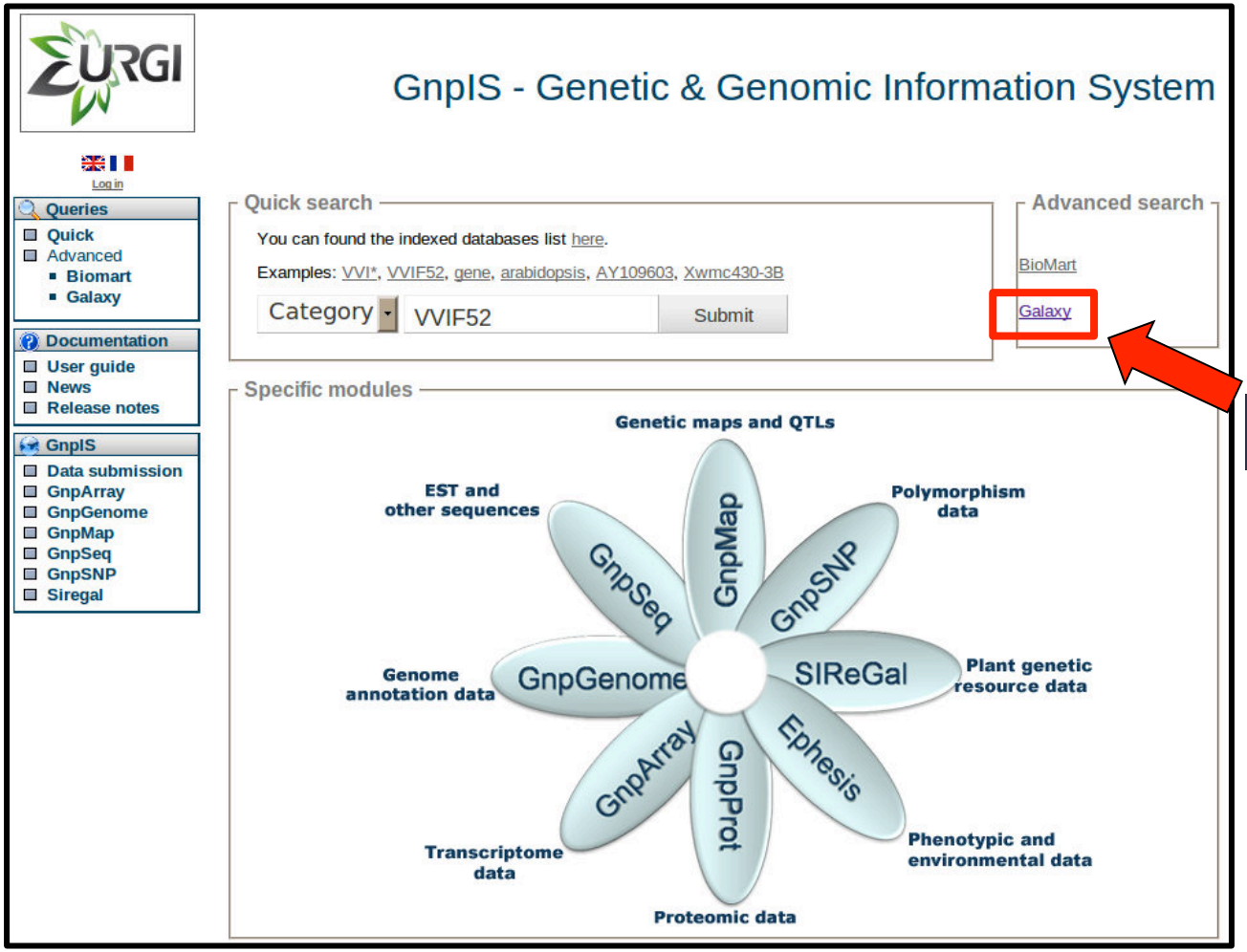
- **Postprocessing tools:**
  - Count multiple hits from the results of BWA
  - Extract short reads from SAM file
  - Extract SNPs with flanks 5' and 3'
  - Keep SNPs without other SNPs in an interval
  - Keep SNPs without 'N' in an interval
  - Remove sequences > 'N' % or 'GC' %
  - VarScan compare (intersection, merge or unique)
  - VarScan to Gff3

IV. Integration of tools in Galaxy

IV. Integration of tools in Galaxy

http://urgi.versailles.inra.fr/

IV. Integration of tools in Galaxy



IV. Integration of tools in Galaxy

http://urgi.versailles.inra.fr/galaxy/

IV. Integration of tools in Galaxy

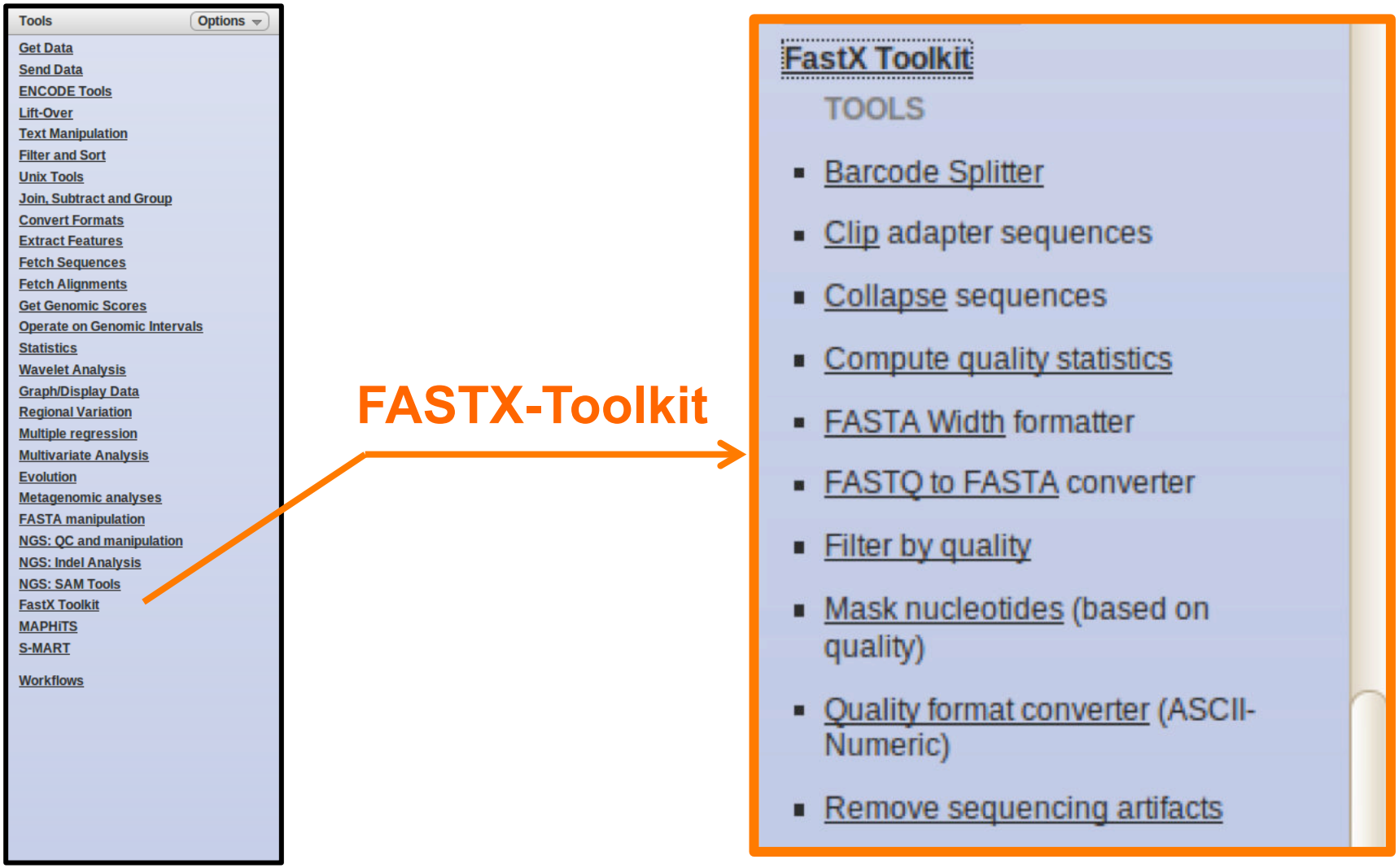
IV.1. Installation of URGI Galaxy

Galaxy is installed on URGI cluster with:

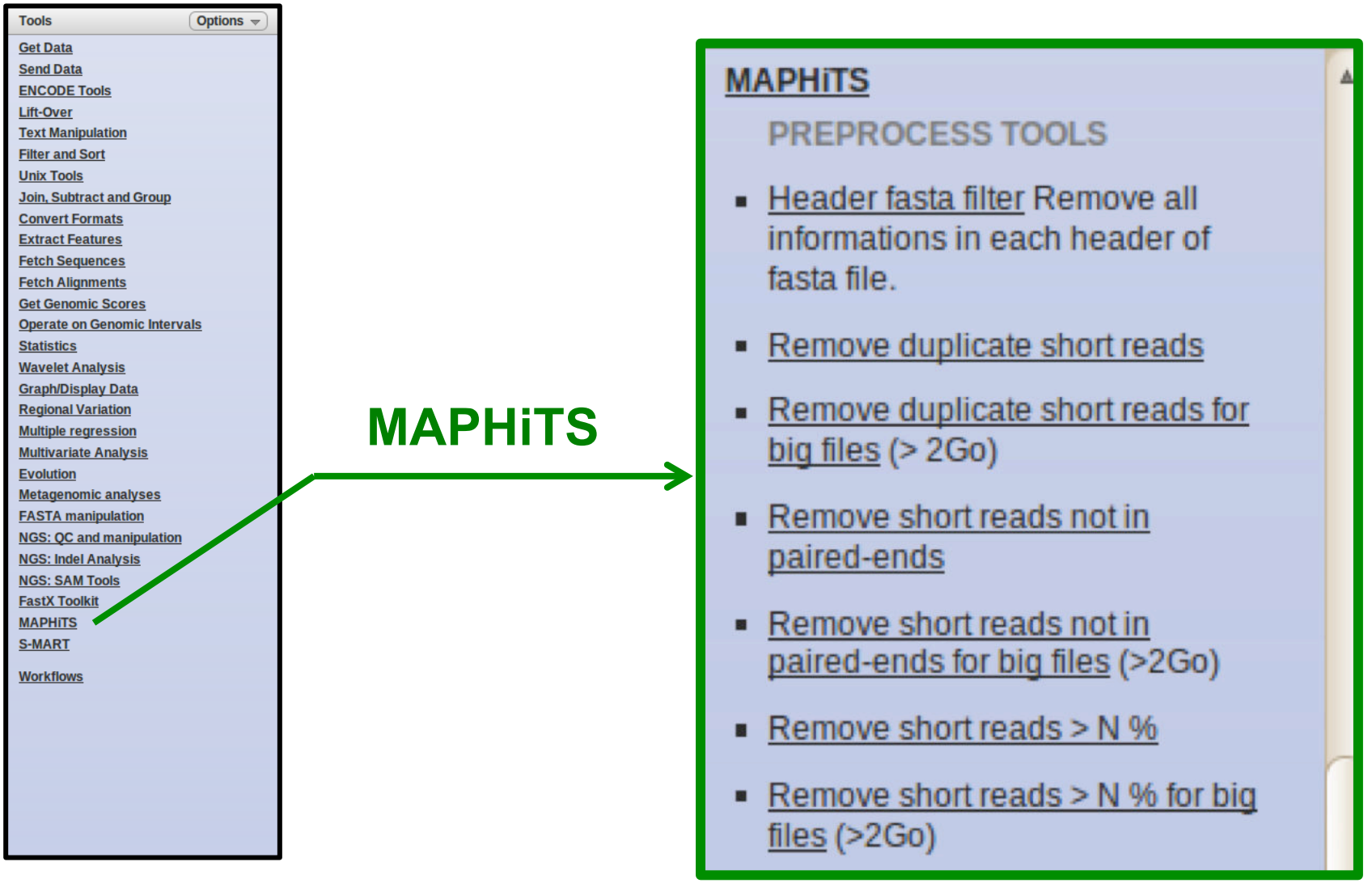
- CPU: **704** (Intel Xeon)
- RAM max: **96 Gb** per job
- Storage: **60 Tb**

Using Sun Grid Engine (for job managment) and a PostgreSQL Database (for Galaxy).

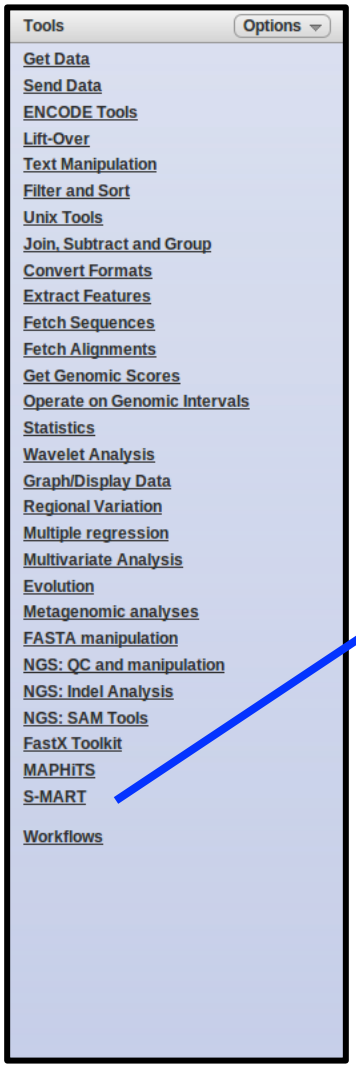
IV.2. New Integrated tools



IV.2. New URGI Integrated tools



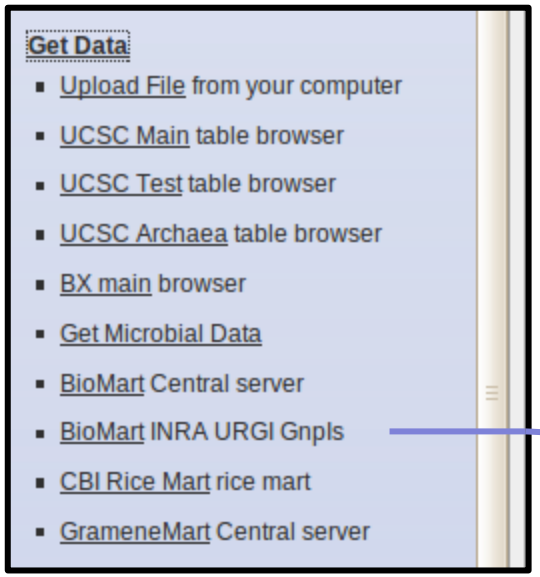
IV.2. New Others URGI Integrated tools



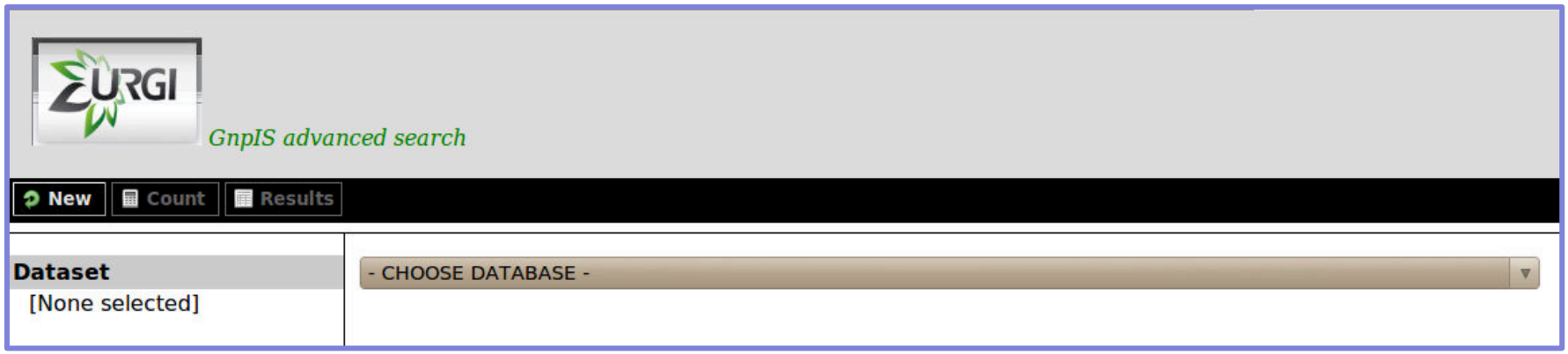
S-MART

IV.2. New Others URGI Integrated tools

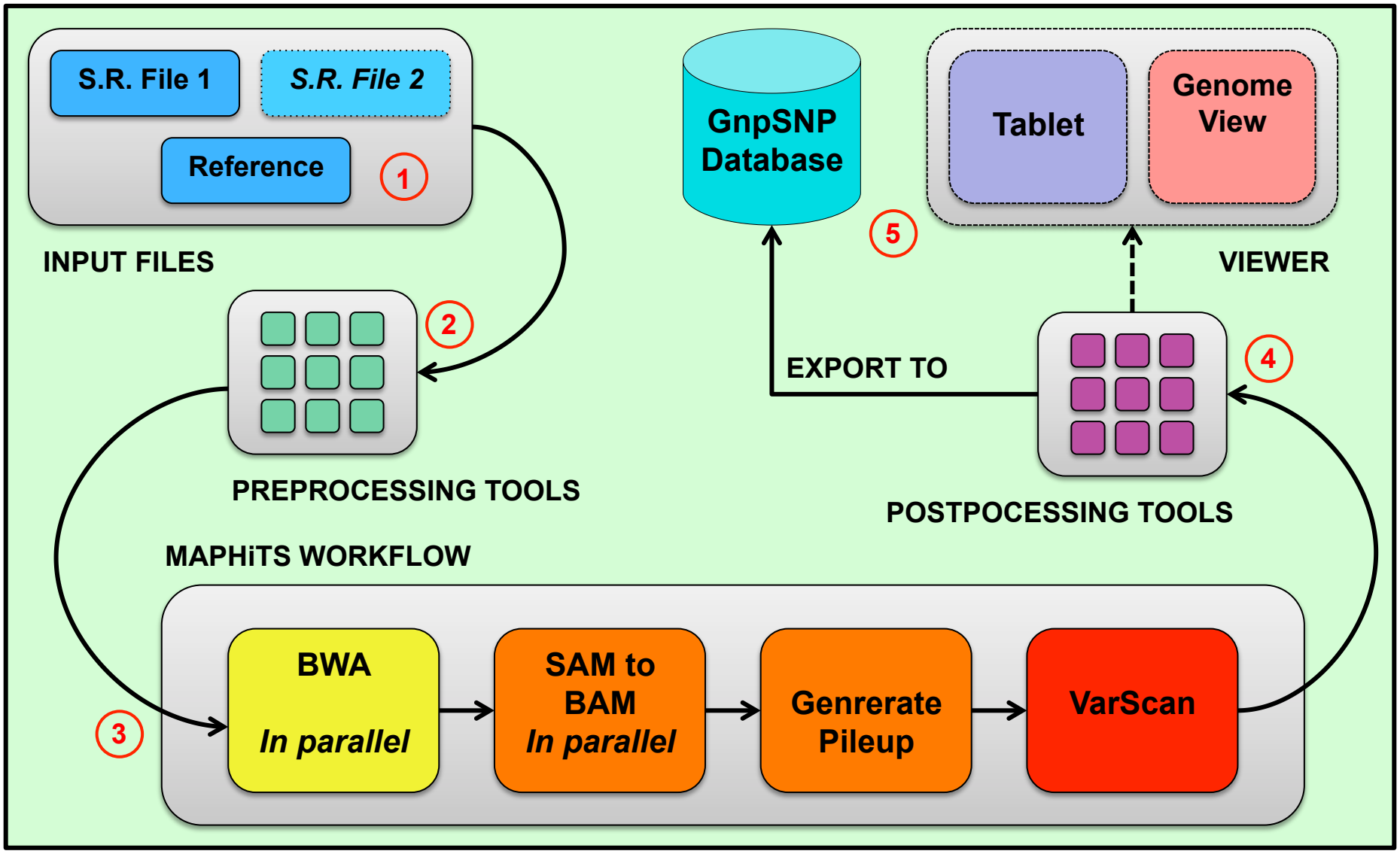
Access to URGI Information System via **BioMart** software



BioMart URGI Gnpls



IV.3. MAPHiTS: Resume



IV.3. MAPHiTS: Resume

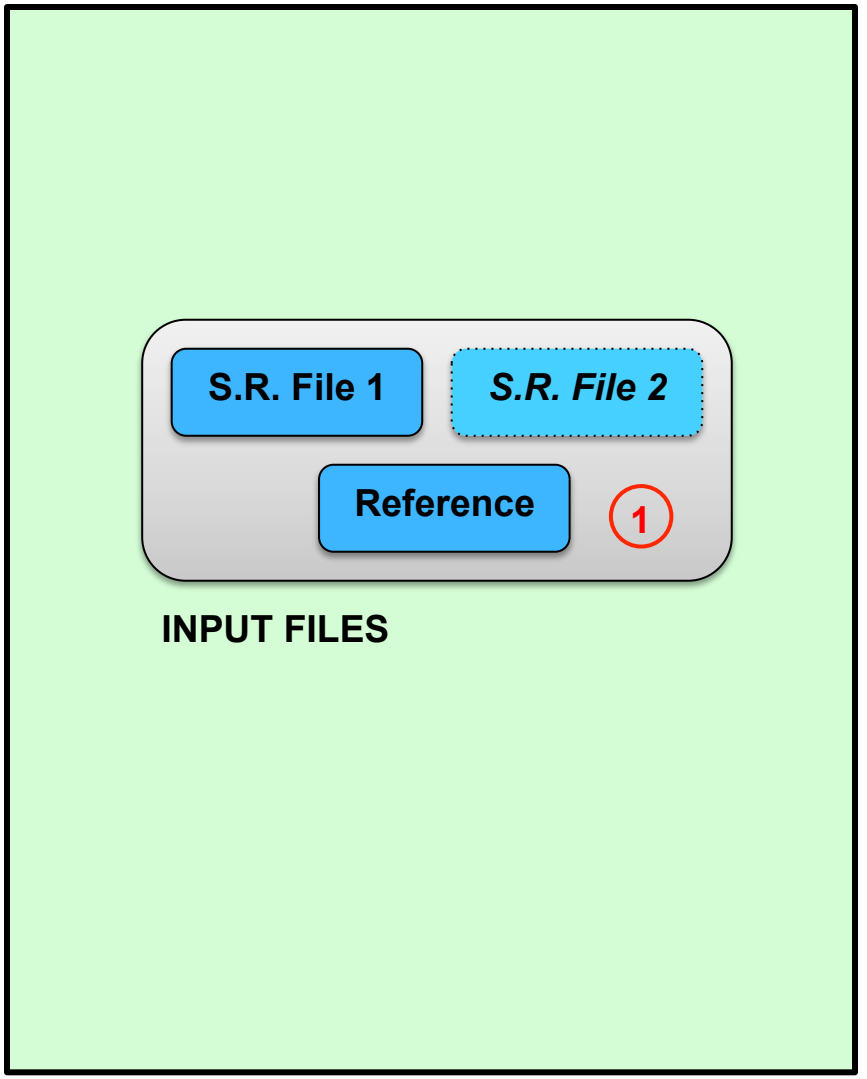
Step 1

Upload your input files:

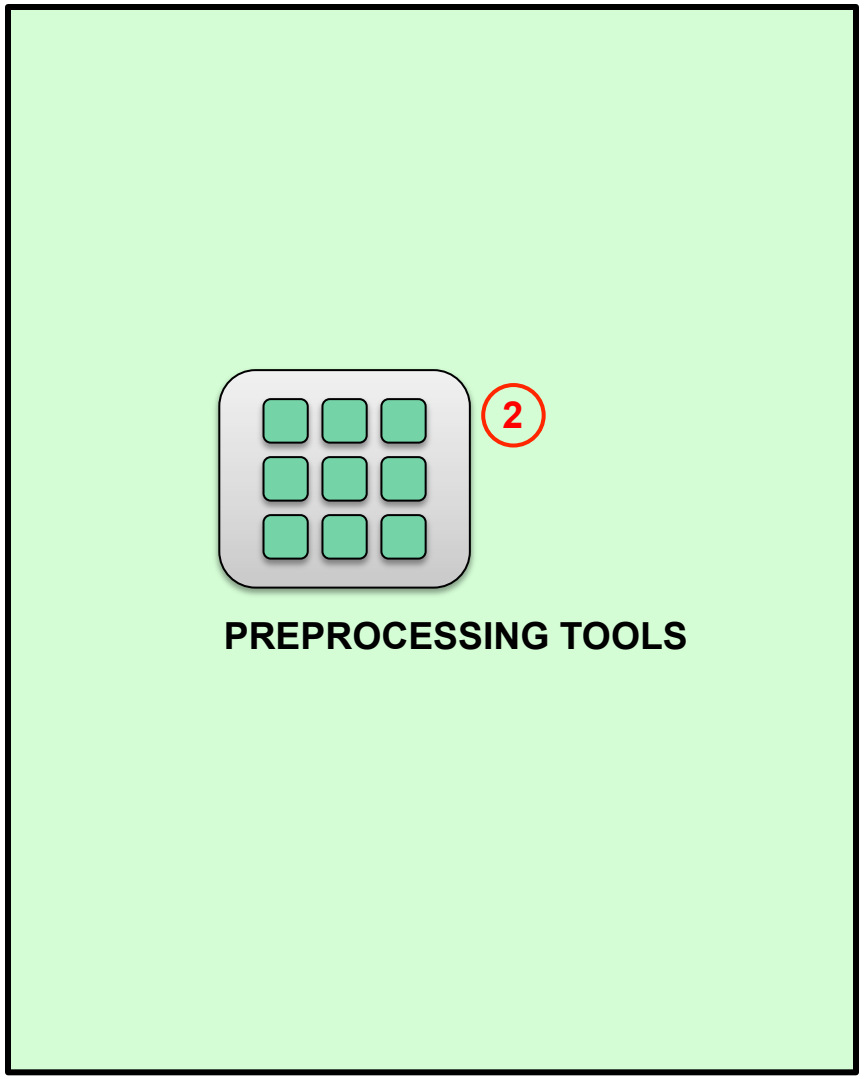
- 1 reference genome (FASTA)
- 1 short reads file if you are in single-ends (FASTA / FASTQ)

OR

- 2 short reads files if you are in paired-ends (FASTA / FASTQ)



IV.3. MAPHiTS: Resume



Step 2

You can filter your input files with one or some preprocessing tools.

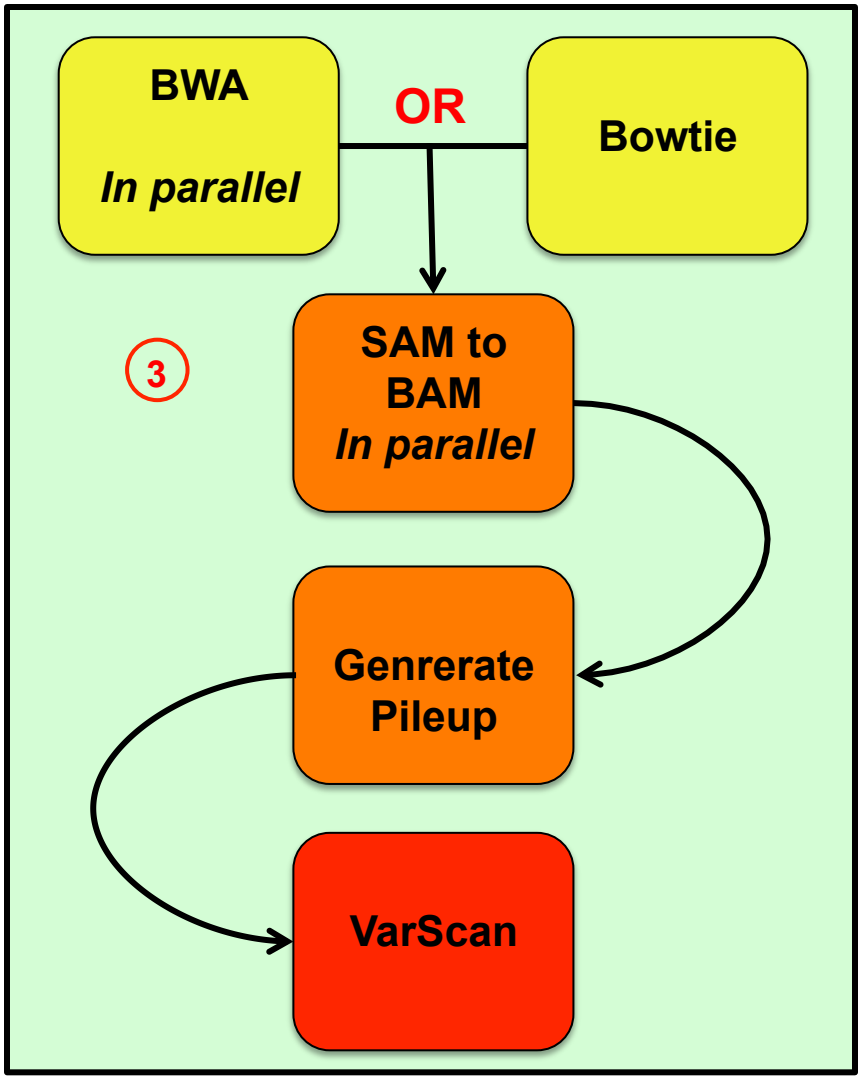
Examples:

- Remove all duplicated short reads
- Trim short reads by quality
- Remove short reads not in paired-ends

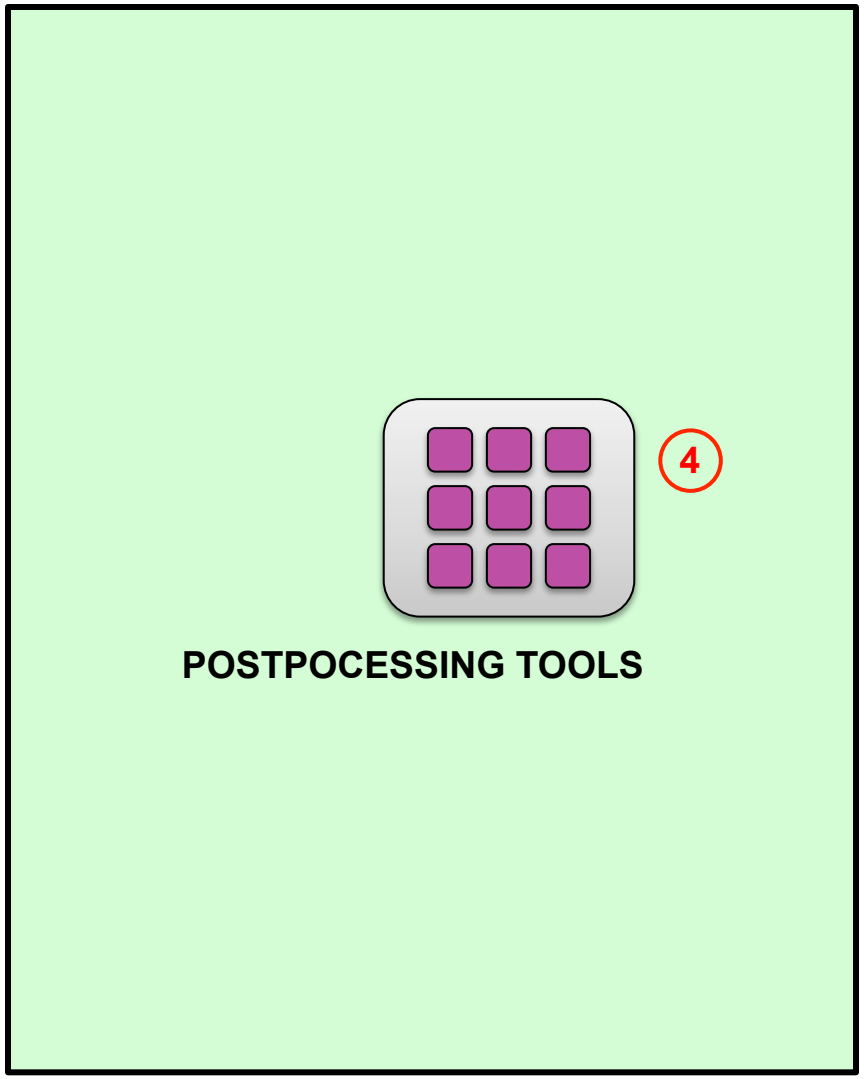
IV.3. MAPHiTS: Resume

Step 3

You have to launch MAPHiTS workflow.



IV.3. MAPHiTS: Resume



Step 4

You can filter your output files with one or some postprocessing tools.

Examples:

- Count multiple hits from the results of BWA
- Extract short reads from SAM file
- VarScan compare

IV.3. MAPHiTS: Resume

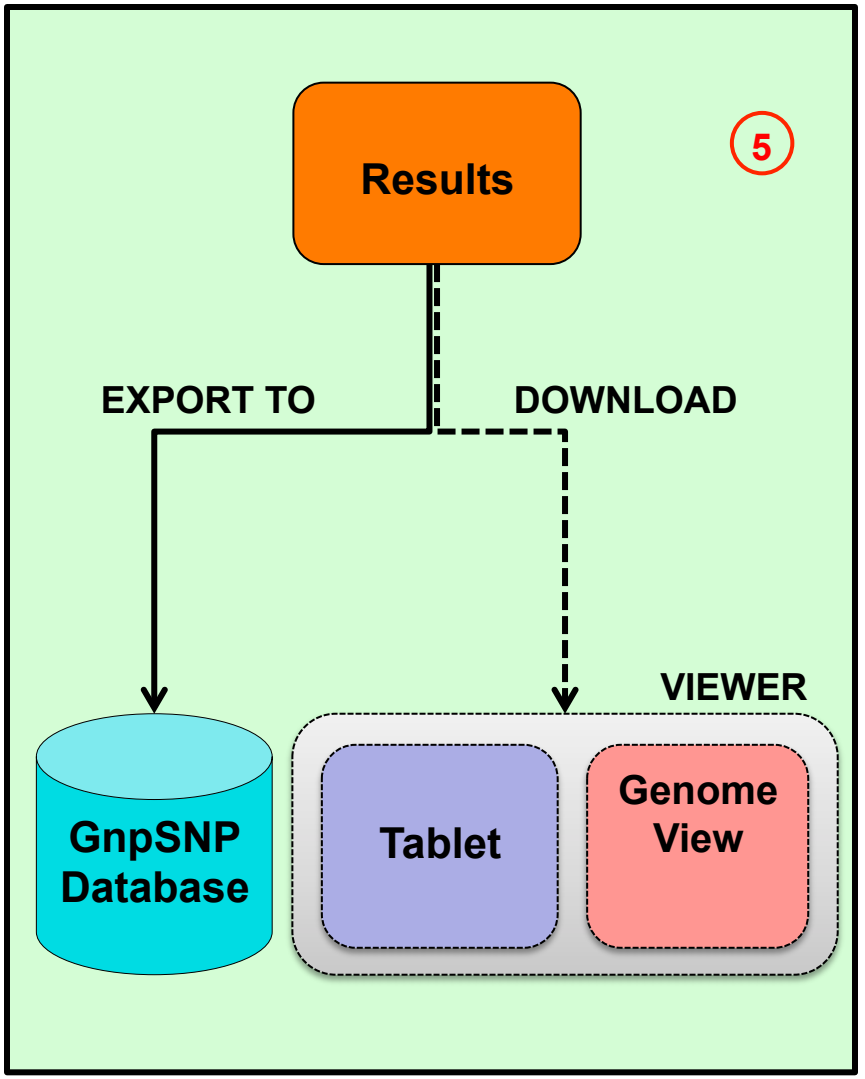
Step 5

Finally, you can :

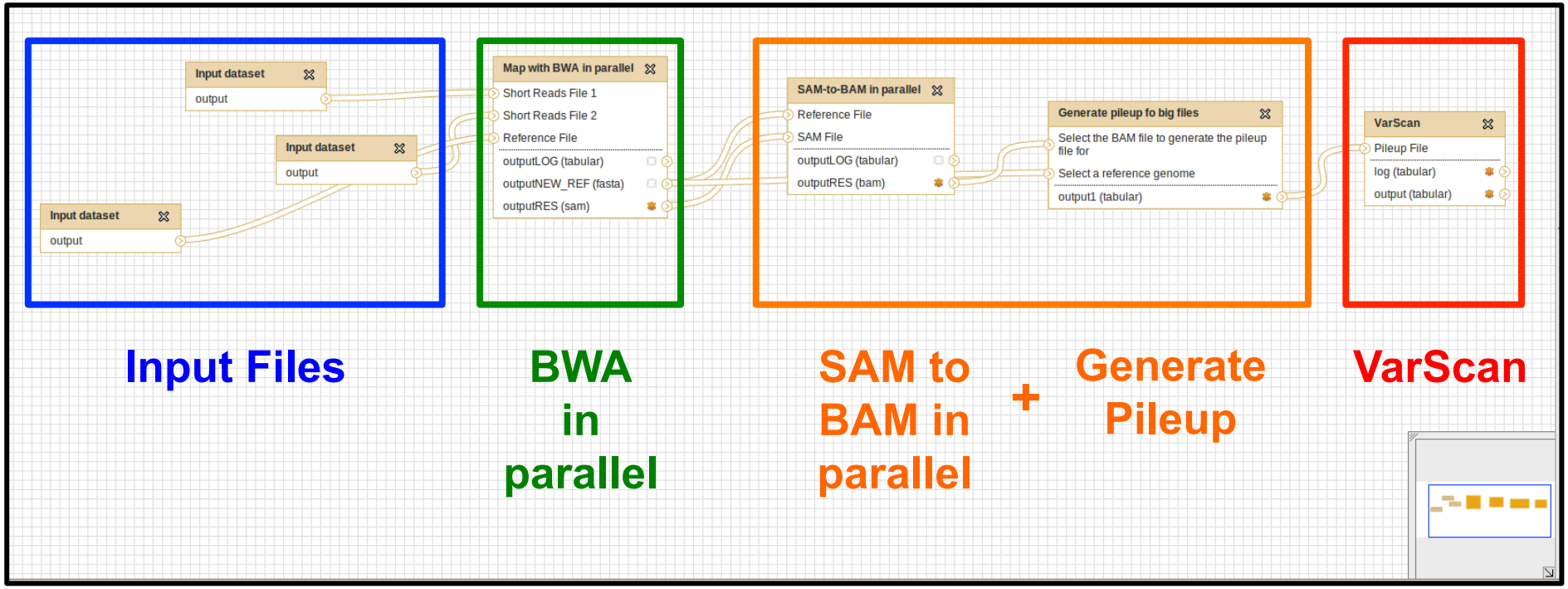
- Export your results to our GnpSNP database

OR / AND

- Download your results and visualize them with your favorite viewer software (Tablet, GenomeView, Gbrowse 2, IGV, …)

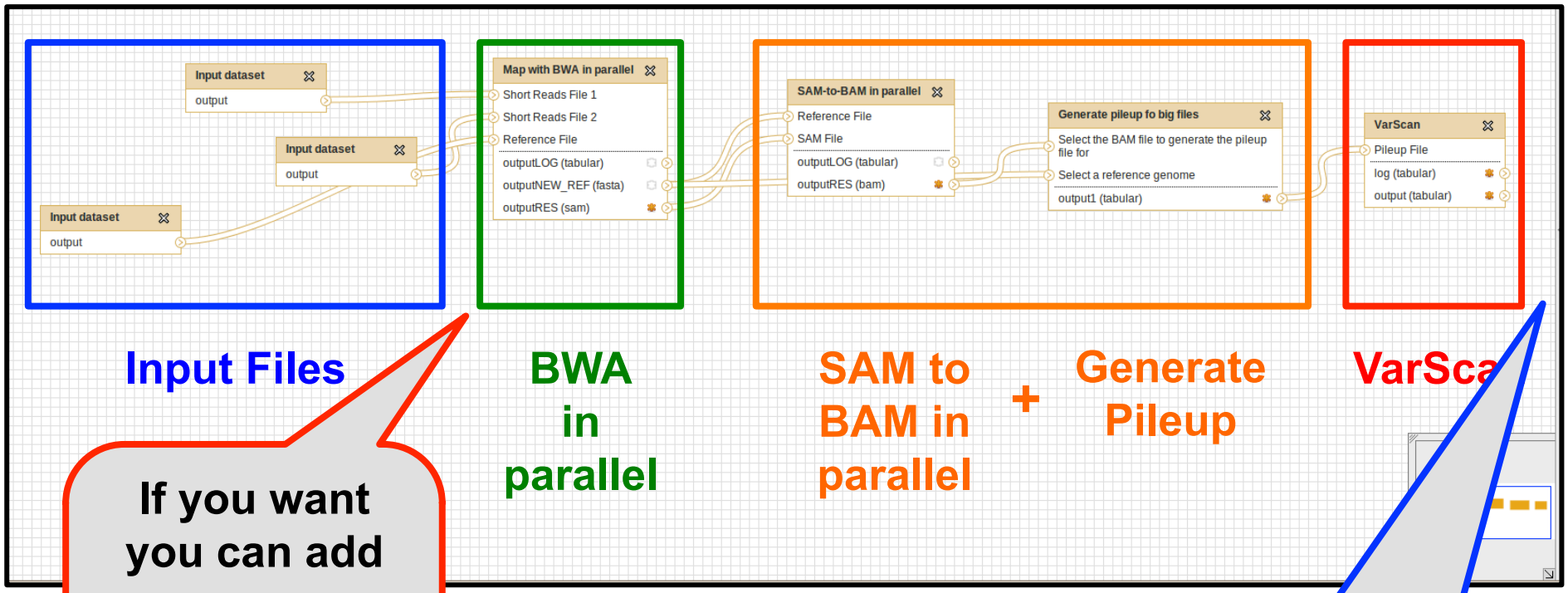


IV.3. MAPHiTS: Build

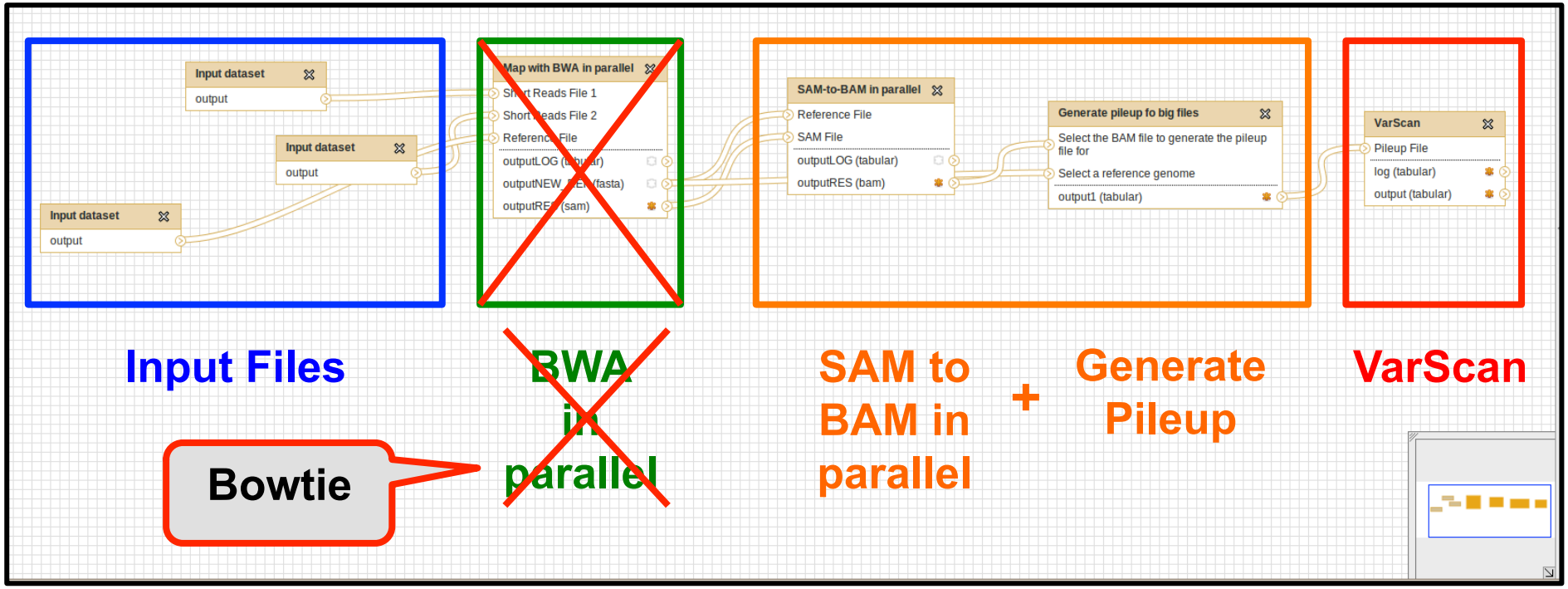


MAPHiTS is build using the graphical interface of Galaxy.

IV.3. MAPHiTS: Build

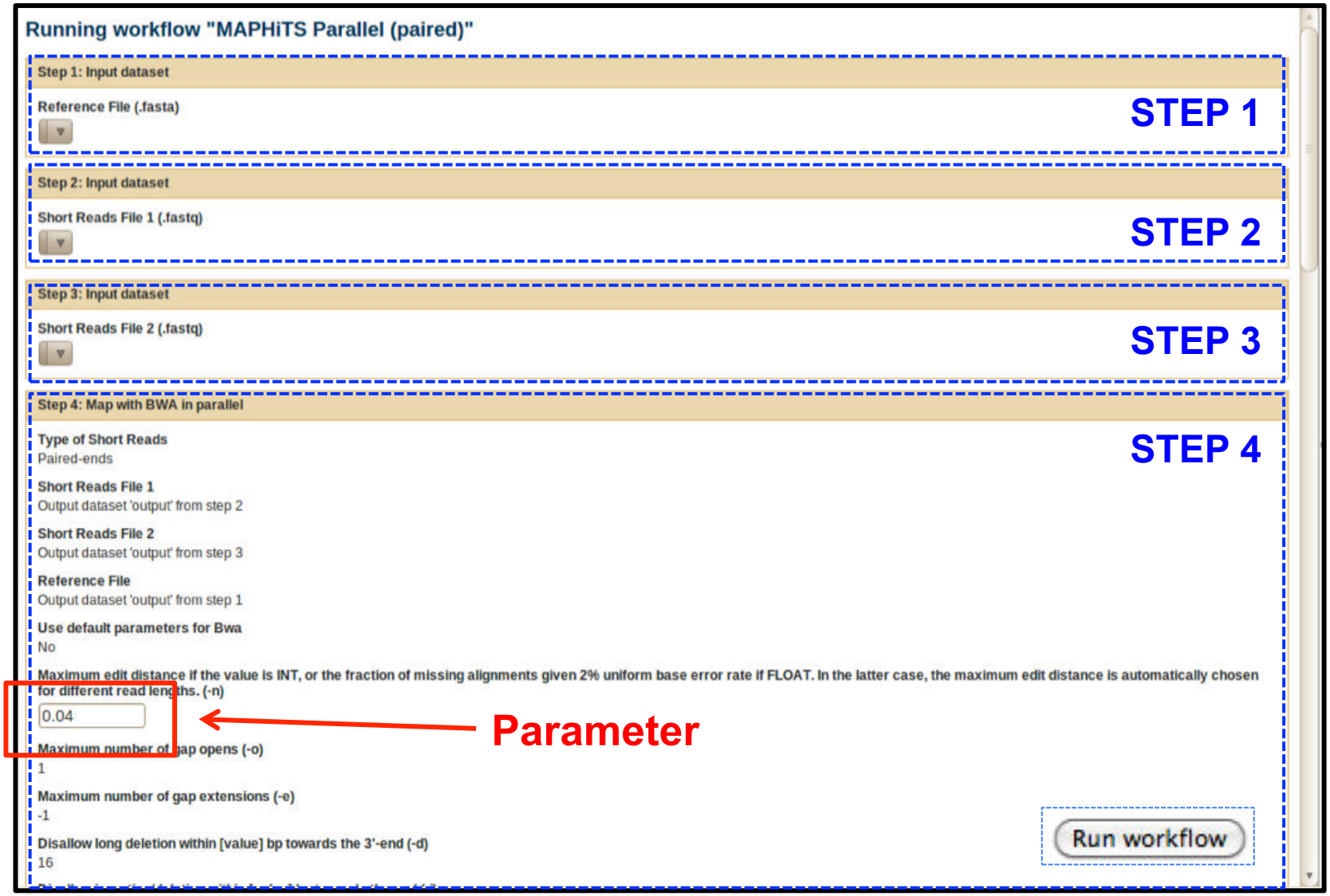


IV.3. MAPHiTS: Build



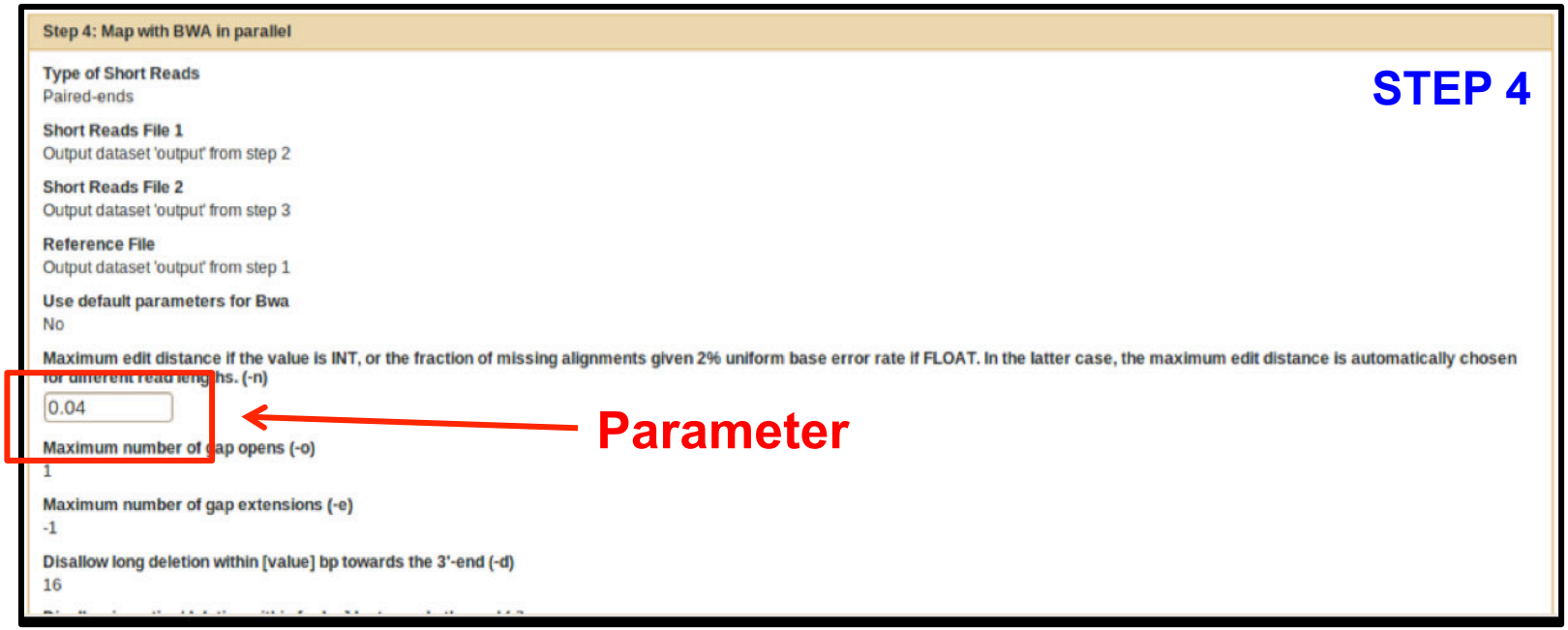
You can **remove** one tool **and replace** it by an other tool very **quickly**.

IV.3. MAPHiTS: Launch

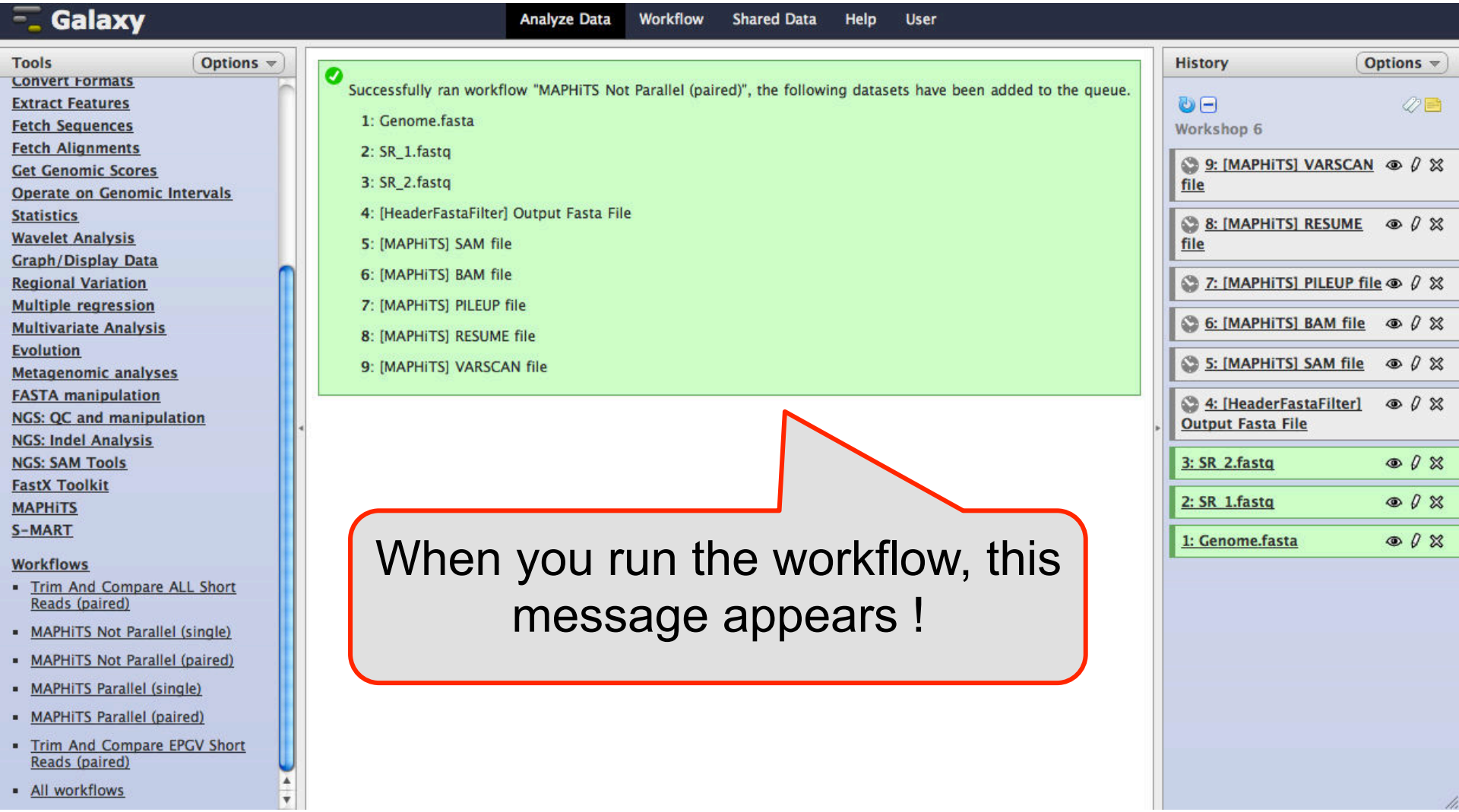


IV.3. MAPHiTS: Launch

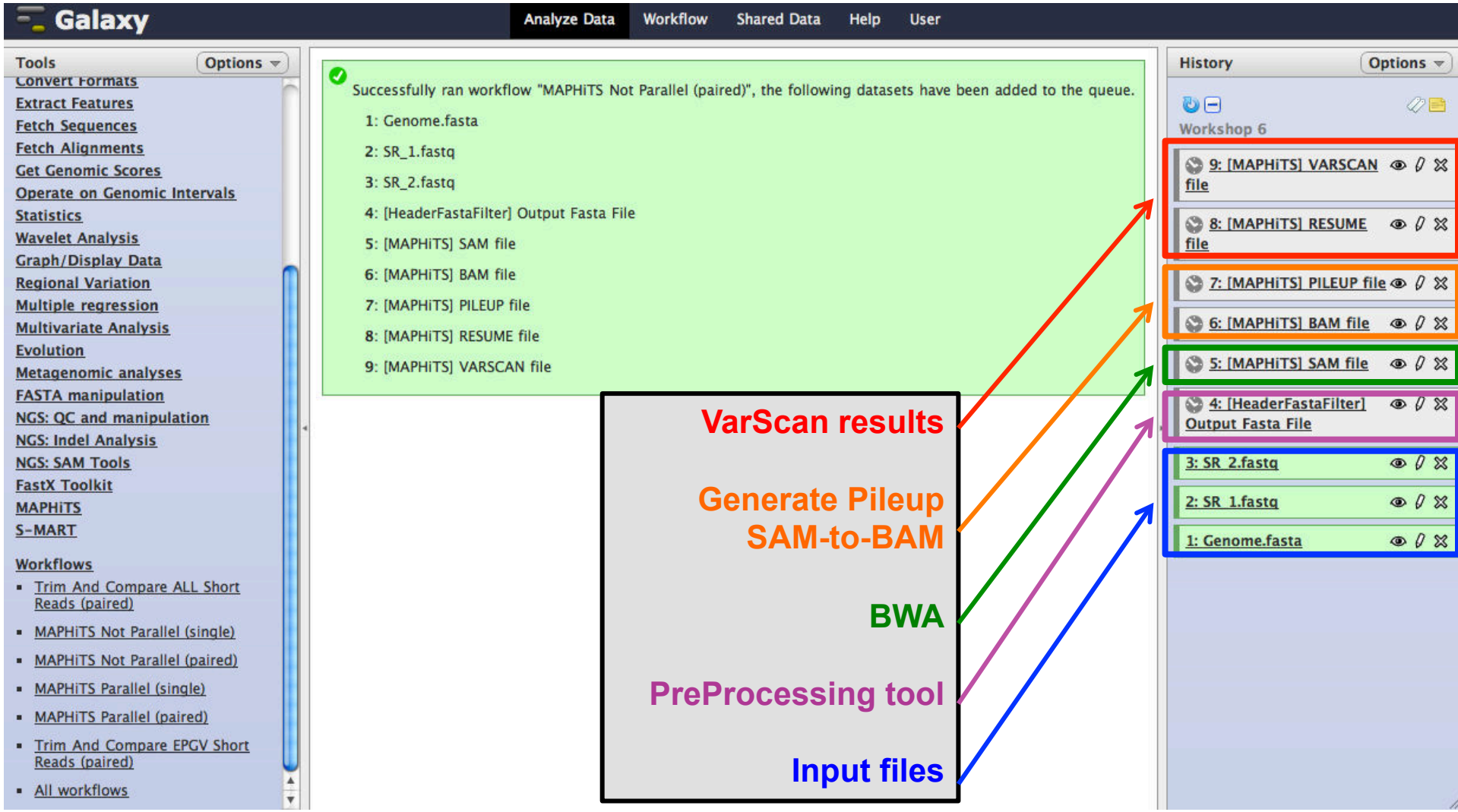
When I build the workflow, I **can choose** what are the parameters that users **can modify or not**.



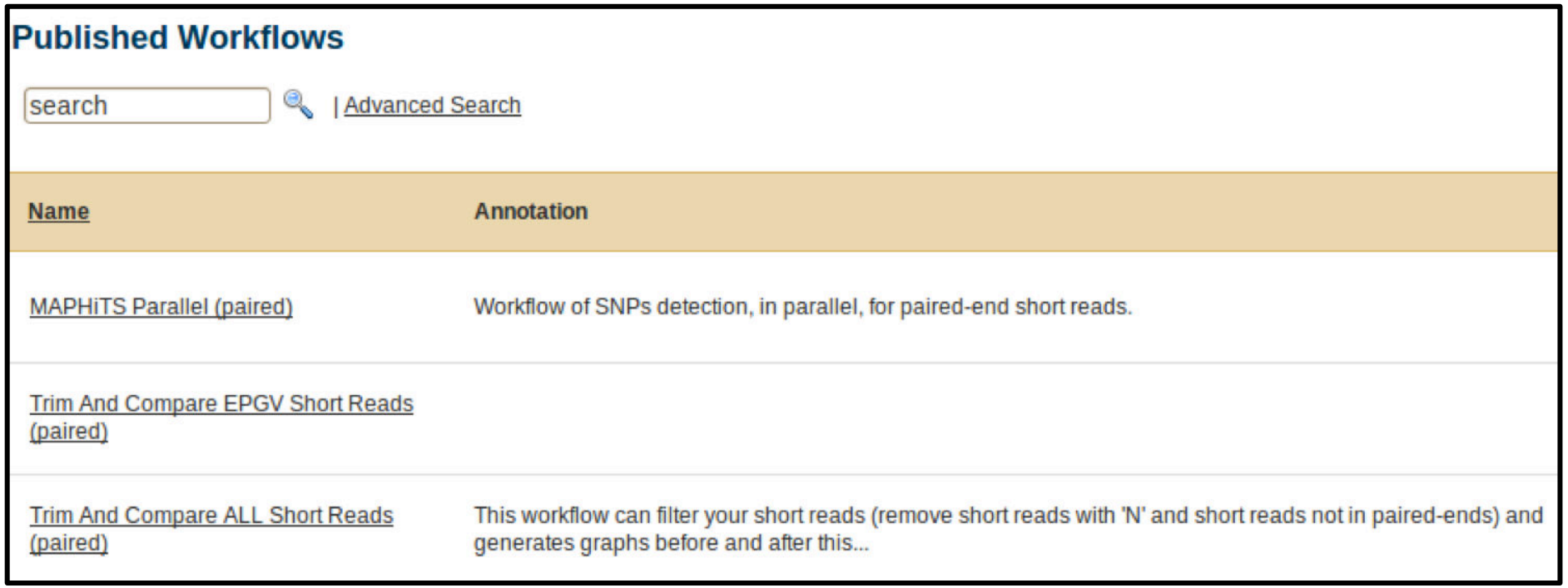
IV.3. MAPHiTS: Launch



IV.3. MAPHiTS: Launch



IV.4. Shared Workflows / Data



Published Workflows

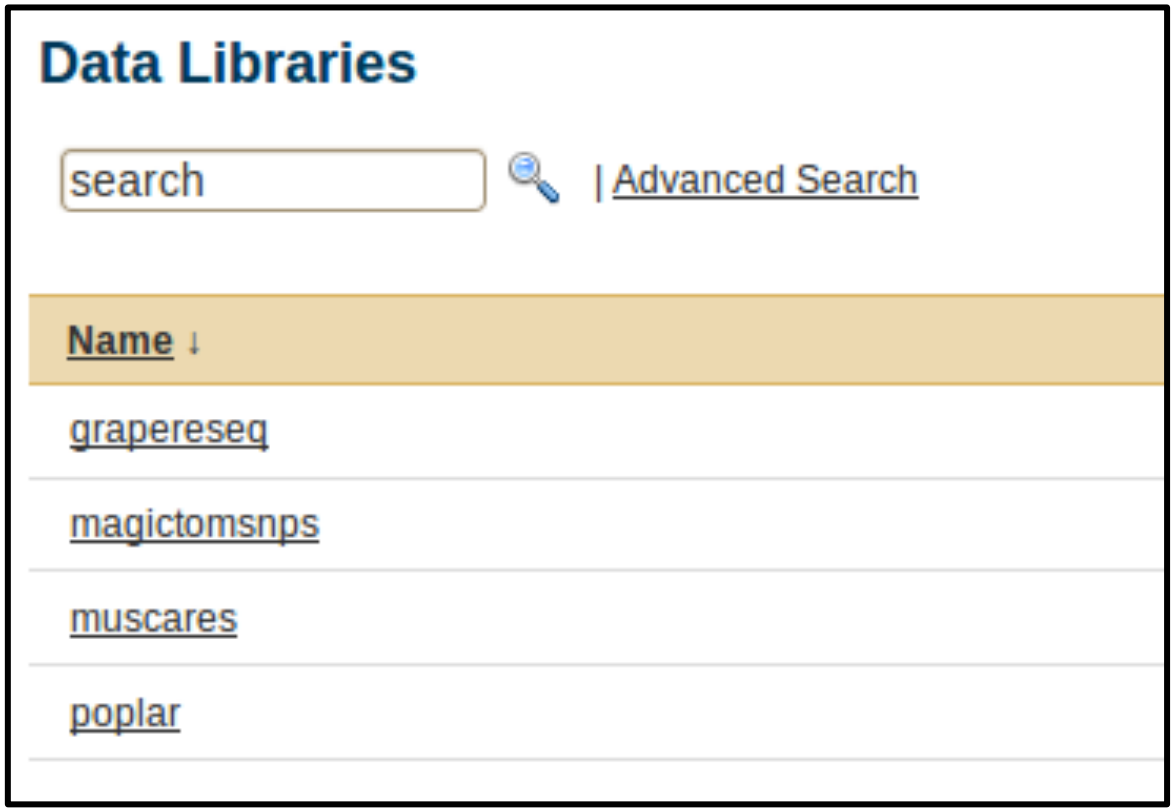
search | Advanced Search

Name	Annotation
MAPHiTS Parallel (paired)	Workflow of SNPs detection, in parallel, for paired-end short reads.
Trim And Compare EPGV Short Reads (paired)	
Trim And Compare ALL Short Reads (paired)	This workflow can filter your short reads (remove short reads with 'N' and short reads not in paired-ends) and generates graphs before and after this...

Some workflows are available for logged users in *'Shared Data'* and *'Published Workflows'* section.

IV.4. Shared Workflows / Data

- In ‘Shared Data’ and ‘Data Libraries’ section, logged users can see 1 directory per Project.
- Users can only see their projects.



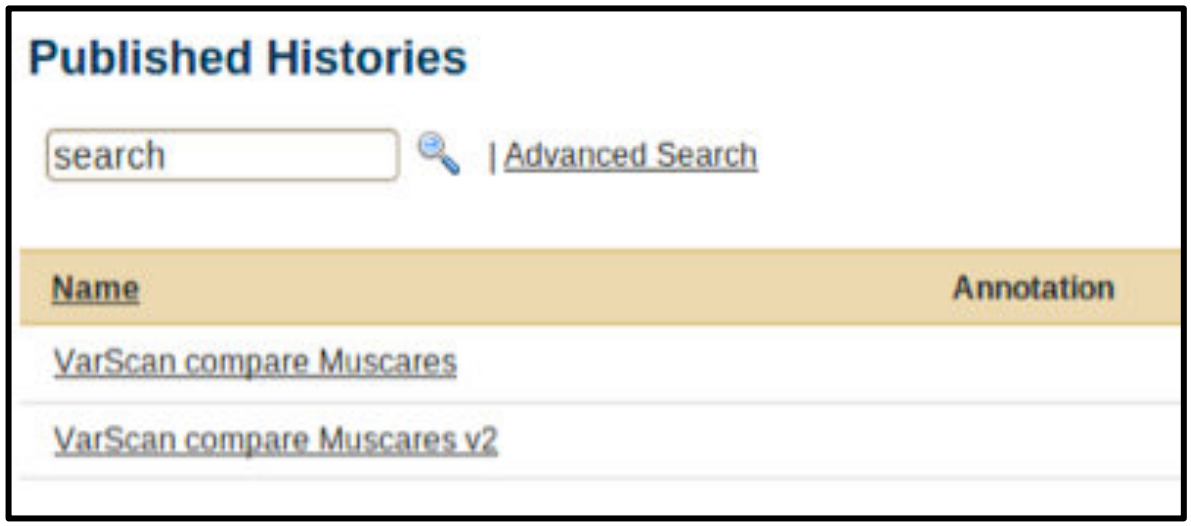
IV.4. Shared Workflows / Data

They can import their data into the history quickly.

➡ Usefull for **NGS** !

IV.5. Shared your History

If a user wants to share its results with other users or a specific user, it's possible !



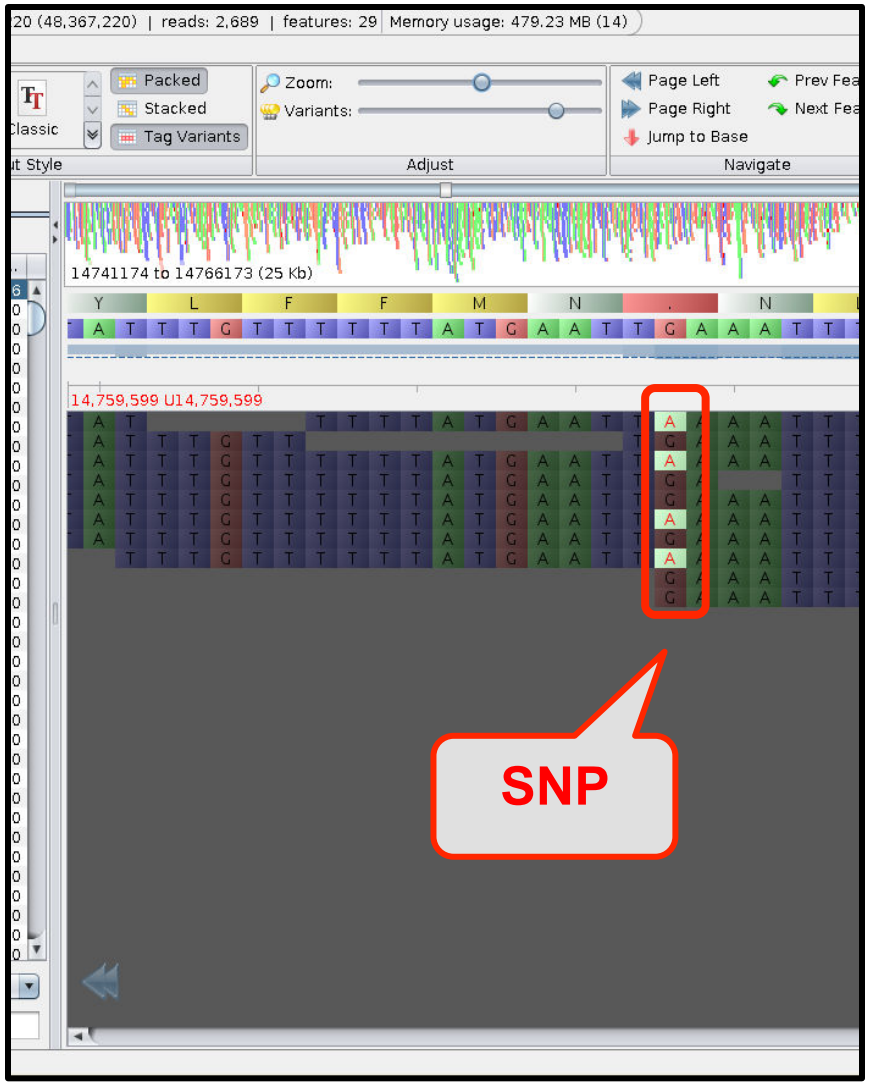
All this histories are in *'Shared Data'* and *'Published Histories'*.

V. Preliminary Results

V. Preliminary Results

We consider that a variant is a SNP if you have:

- **10** short reads in minimum at this position
- **4** variants in minimum
- **30** of mapping quality in minimum
- **30%** of variant allele frequency in minimum
- Pvalue threshold <= **1e10-3**



V. Preliminary Results

	A	B	C
S.R.	**71 Millions**	**70 Millions**	**45 Millions**
STEP 1	62 Millions short reads (**88%**)	60 Millions short reads (**85%**)	38 Millions short reads (**85%**)
STEP 2	**84,94 %** mapped in PE	**48,20 %** mapped in PE	**83,11 %** mapped in PE
SNPs	**847.130**	**3.245.011**	**532.756**

0/ I start with short reads in paired-ends (101 nucleotides).

1/ I run one workflow to filter and trim all my short reads in input files.

⇨ 15 % of short reads are removed for **ALL** species.

2/ I run MAPHiTS.

⇨ 85 % of short reads are mapped in paired-ends for **A** and **C** but only 48% for **B**.

I've got 500.000 SNPs for **A** and **C** and 3 millions for **B** !

⇨ **A** and **C** are closest to reference genome than **C**.

VI. Perspectives

VI. Perspectives

- **Add new tools** (all tools used in all our pipelines)
- **Link Galaxy to a visualization software** (Gbrowse 2, Tablet, GenomeView, …)
- **Application Note in progress (2011)**

Acknowledgements

- Dave Clements
- Galaxy developers
- Galaxy community

Acknowledgements

Thank you for your attention !!!

ALIMENTATION
AGRICULTURE
ENVIRONNEMENT