Development of a workflow for SNPs detection in grapevine species: *MAPHiTS*

-> Integration of MAPHiTS in Galaxy

BRAS Marc
Colloque EPGV 2011 – Monastir, Tunisie

ALIMENTATION
AGRICULTURE
ENVIRONNEMENT

A. Galaxy Presentation

A.1. Galaxy Homepage

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

A.2. 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).

A.3. Homepage presentation

Return to homepage

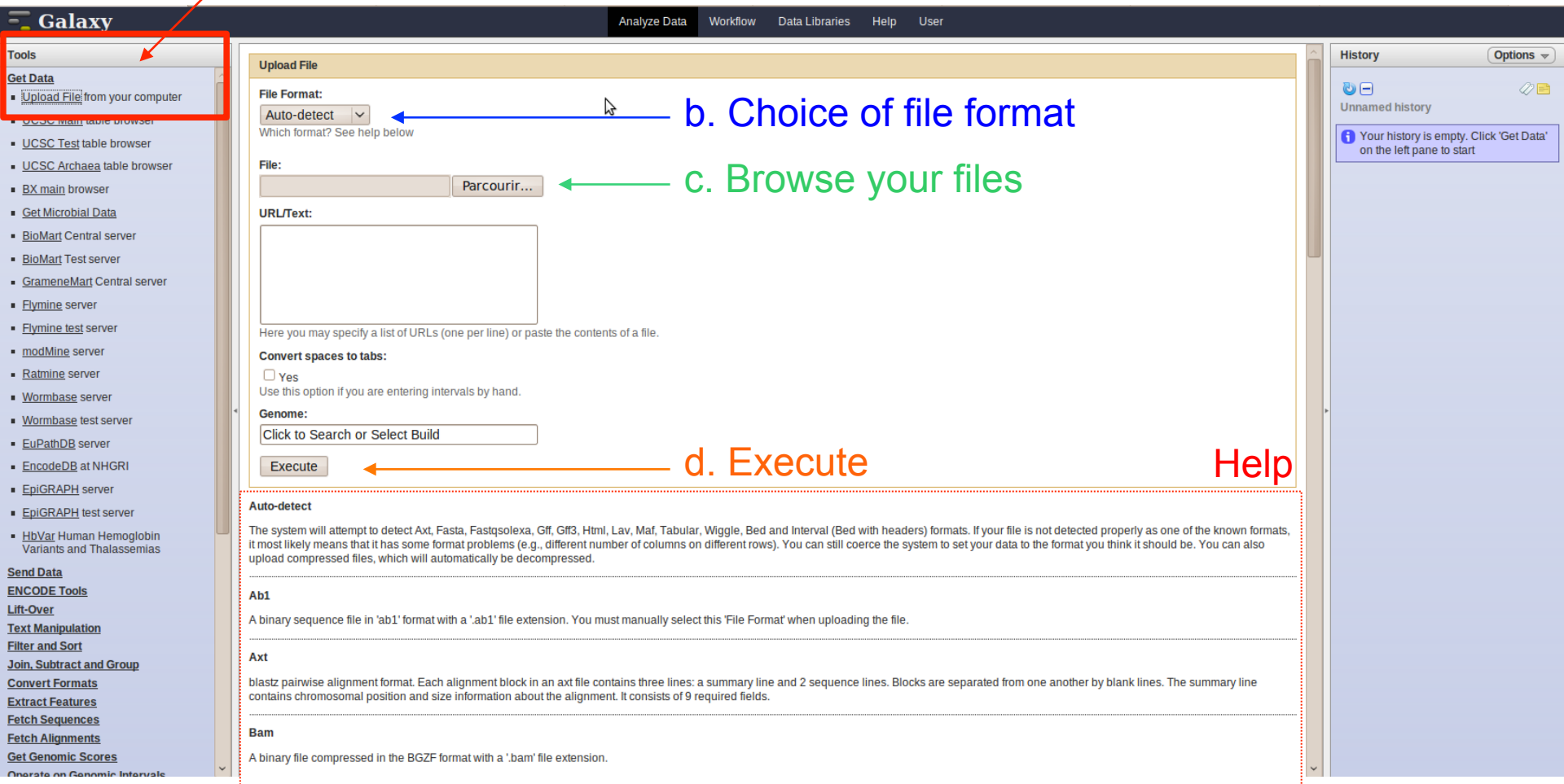
Tools

Managment of Galaxy

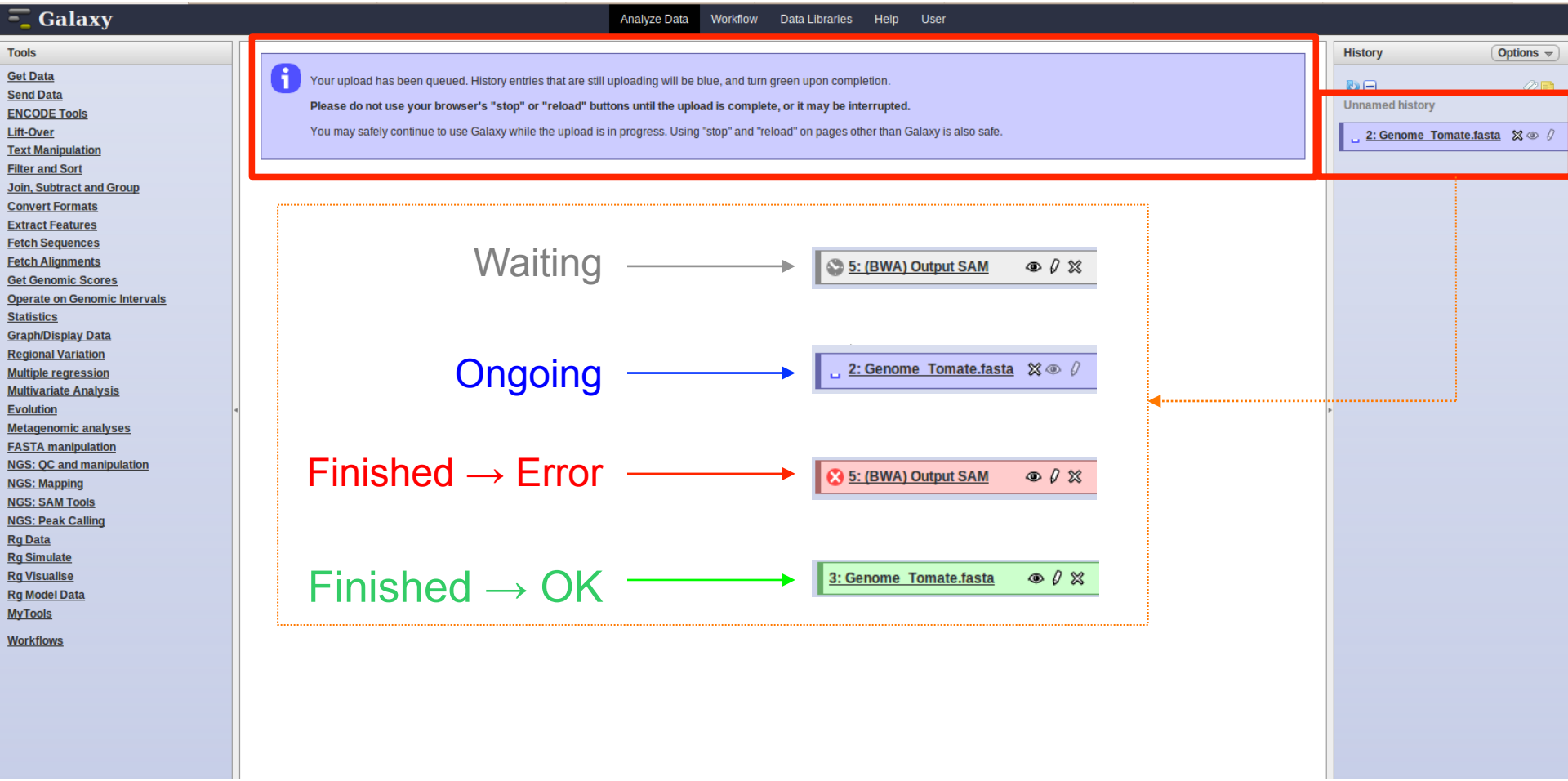
History

A.4. How to upload your data ?

a. Tools → Get Data → Upload File

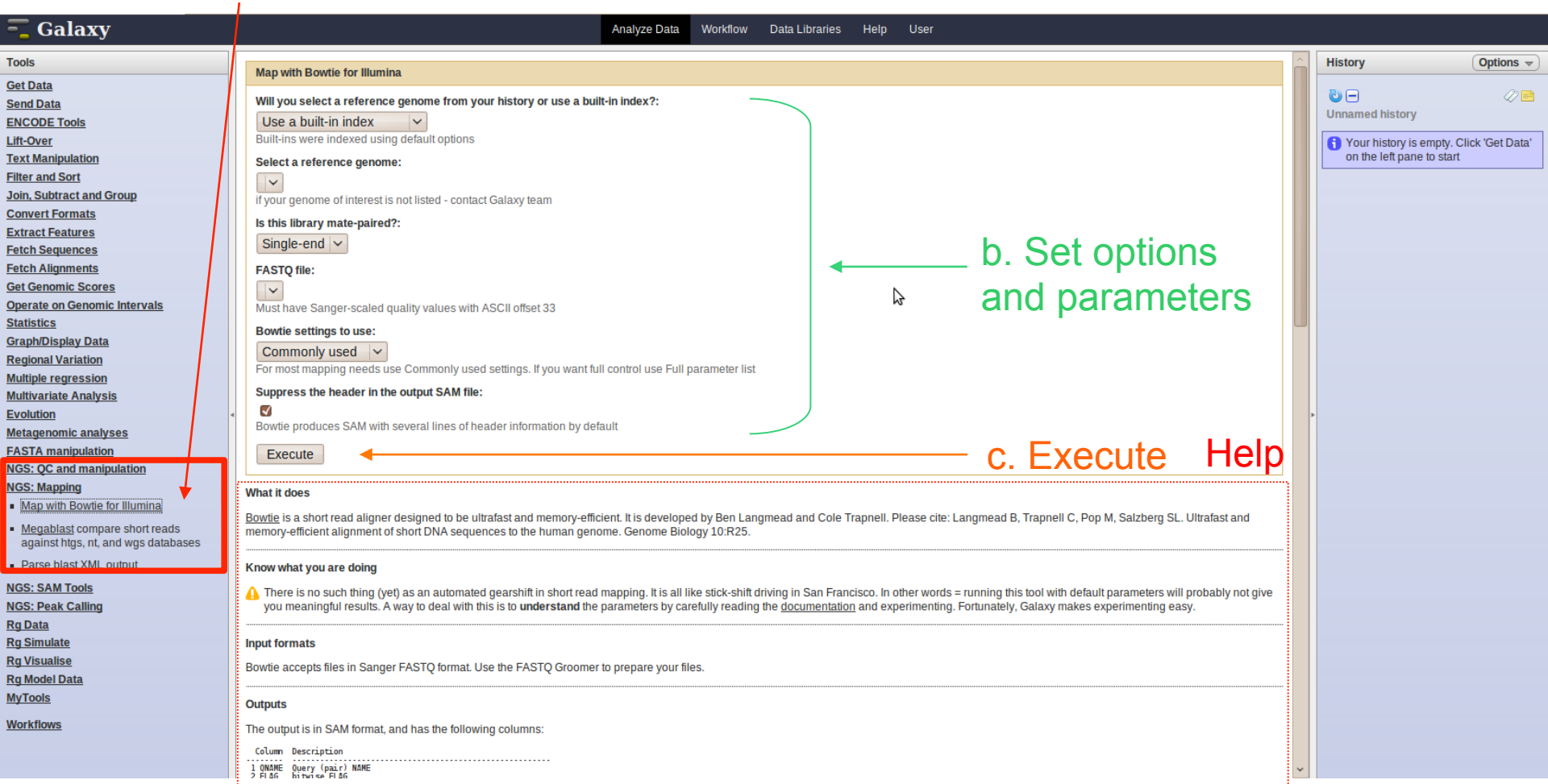


A.4. How to upload your data ?

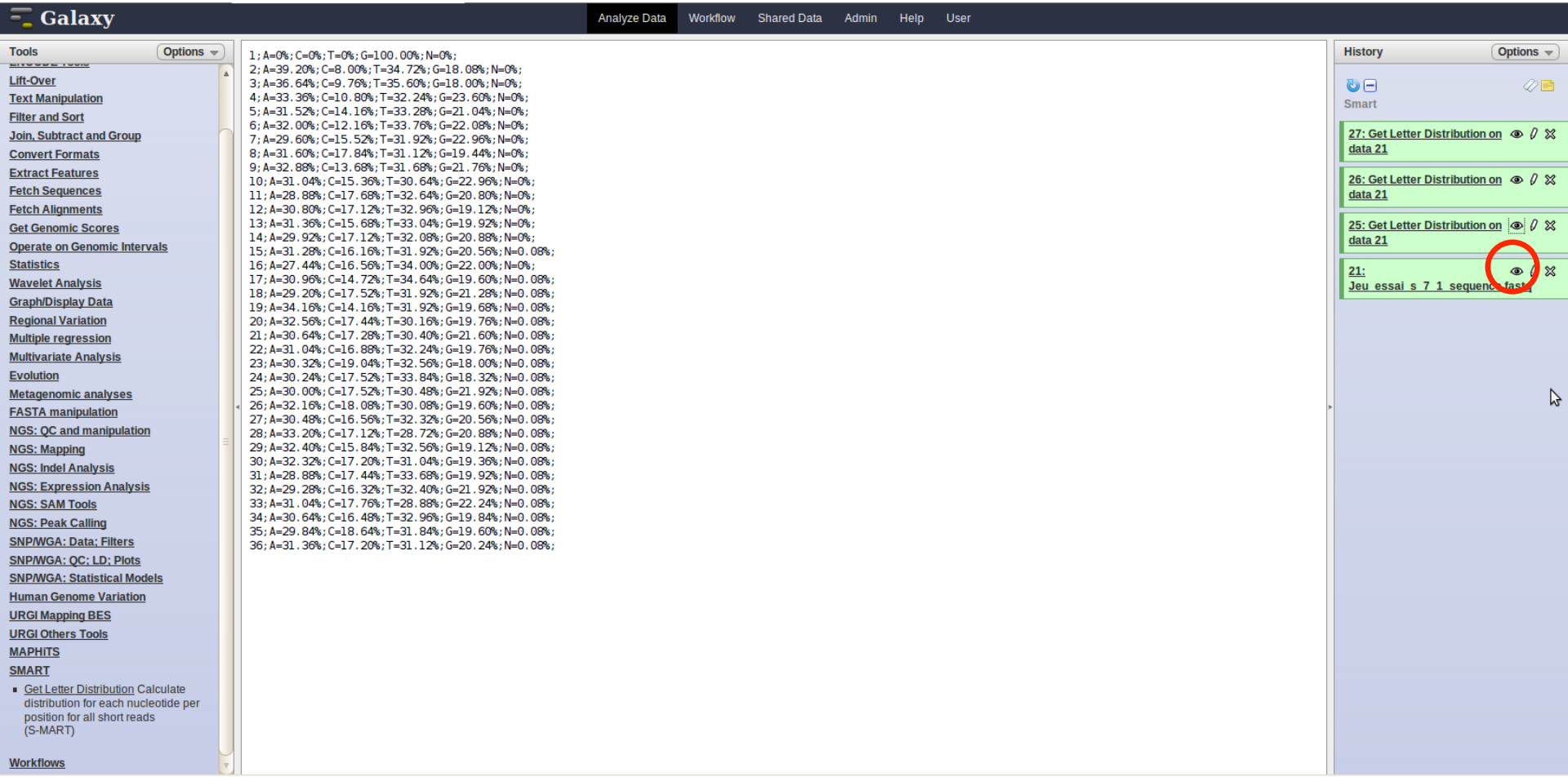


A.5. How to use a tool ?

a. Choose a tool from the list

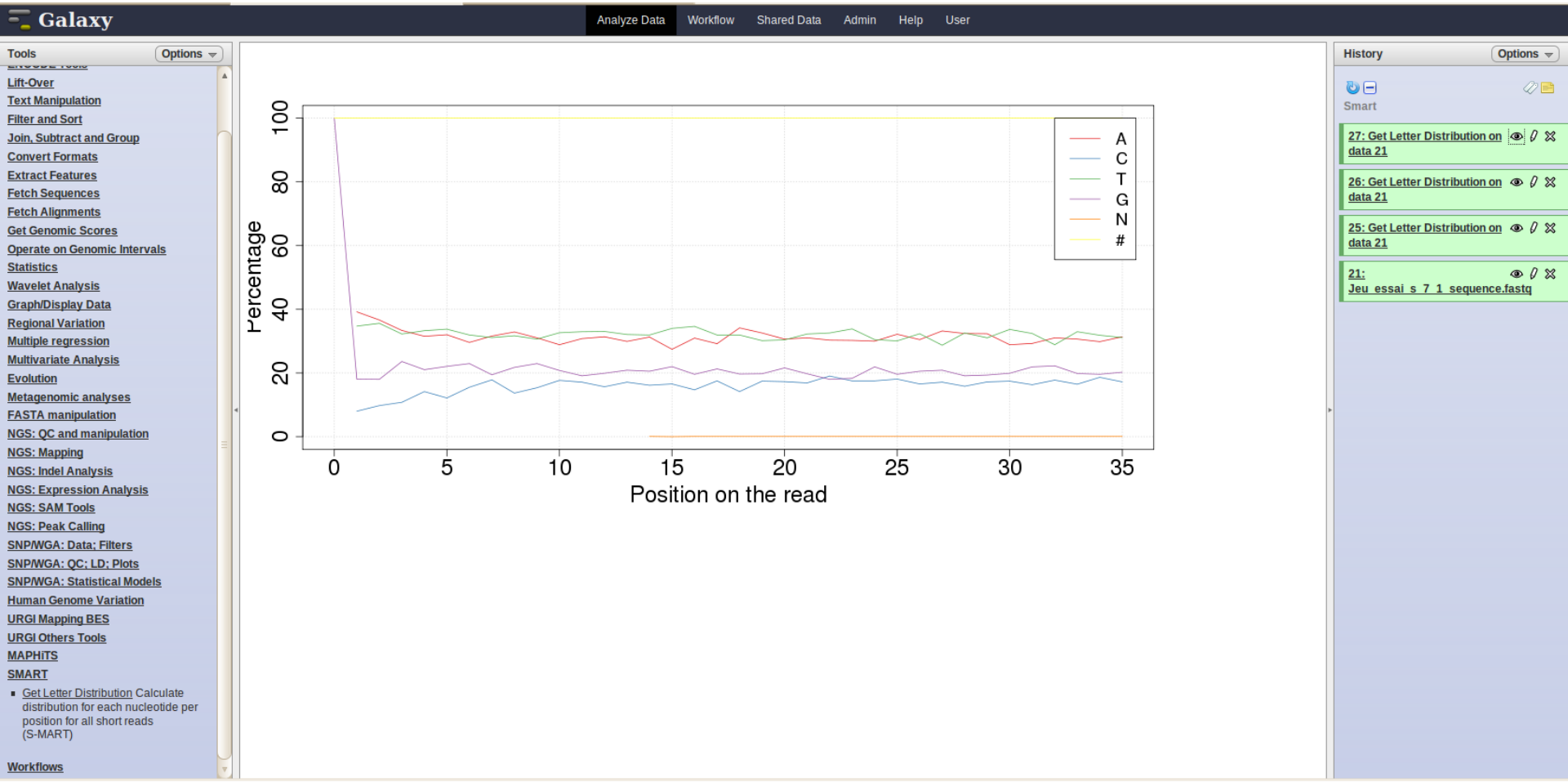


A.6. Example



SMART - Get Letter Distribution

A.6. Example



SMART - Get Letter Distribution

B. MAPHiTS Presentation

B.1. Background and objectives of the pipeline

- **Objectives:**

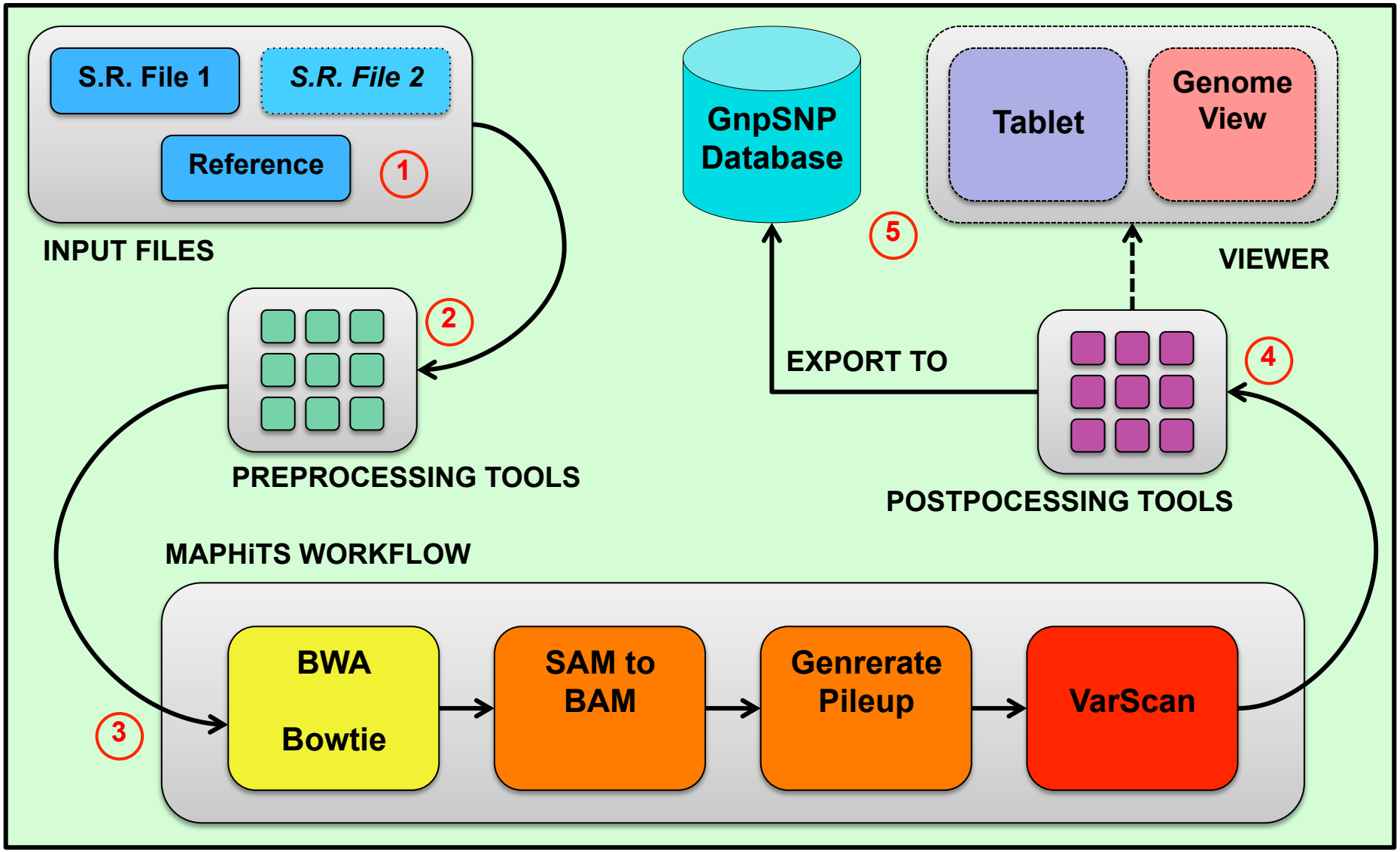
  Detect a set of SNPs between various species of Grape after mapping short reads against a reference genome.

- **Data:**
  - Muscares : 6 genotypes
  - GrapeReSeq : 16 genotypes

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

➡ Other projects are also in progress with others species.

B.2. MAPHiTS: Resume



B.3. 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

Fewer Ressources Required !

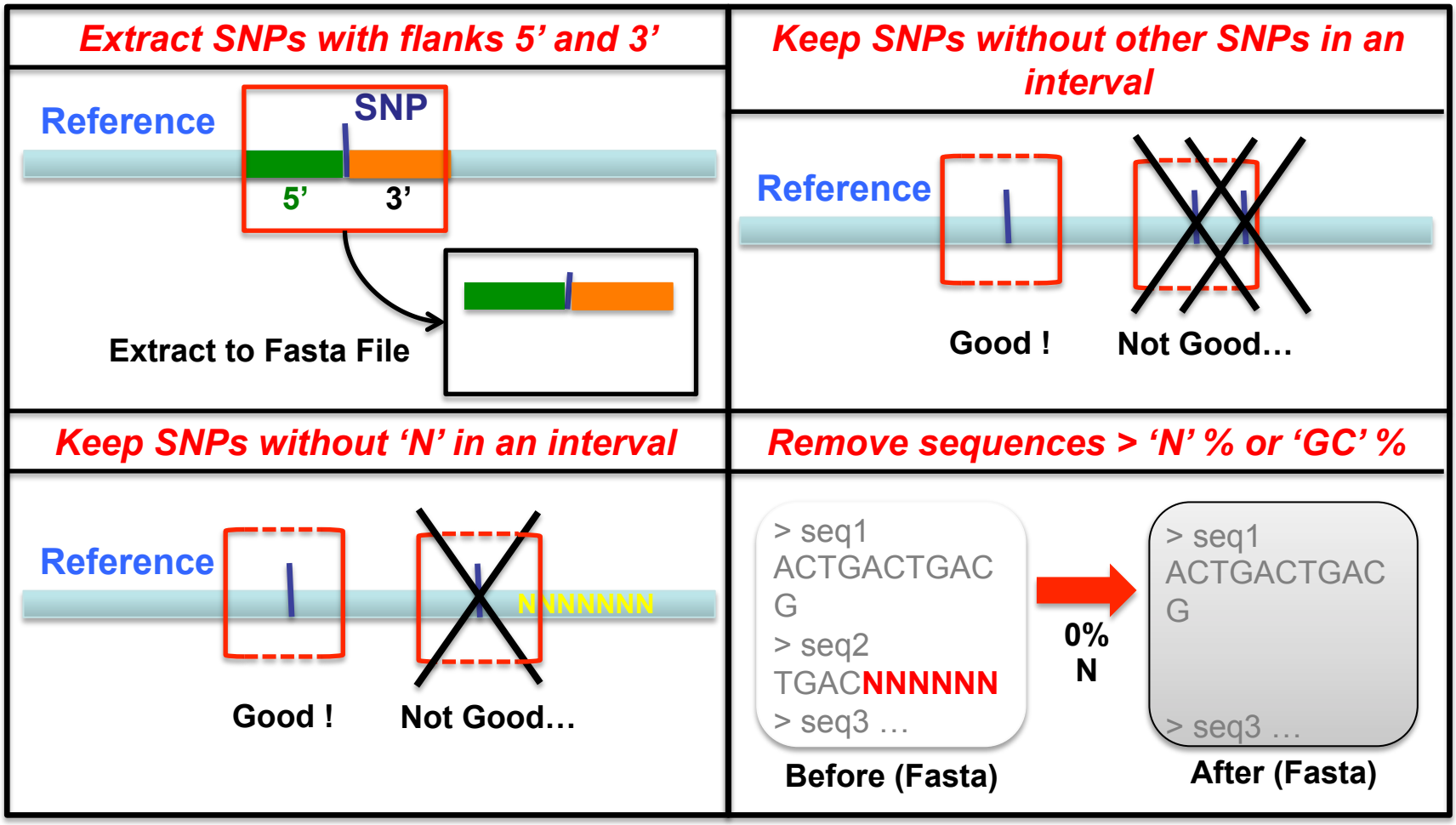
Exemple:
- Before: 1 big job in 1 cluster node
- Now: N little jobs in N cluster nodes

B.3. MAPHiTS Development Tools

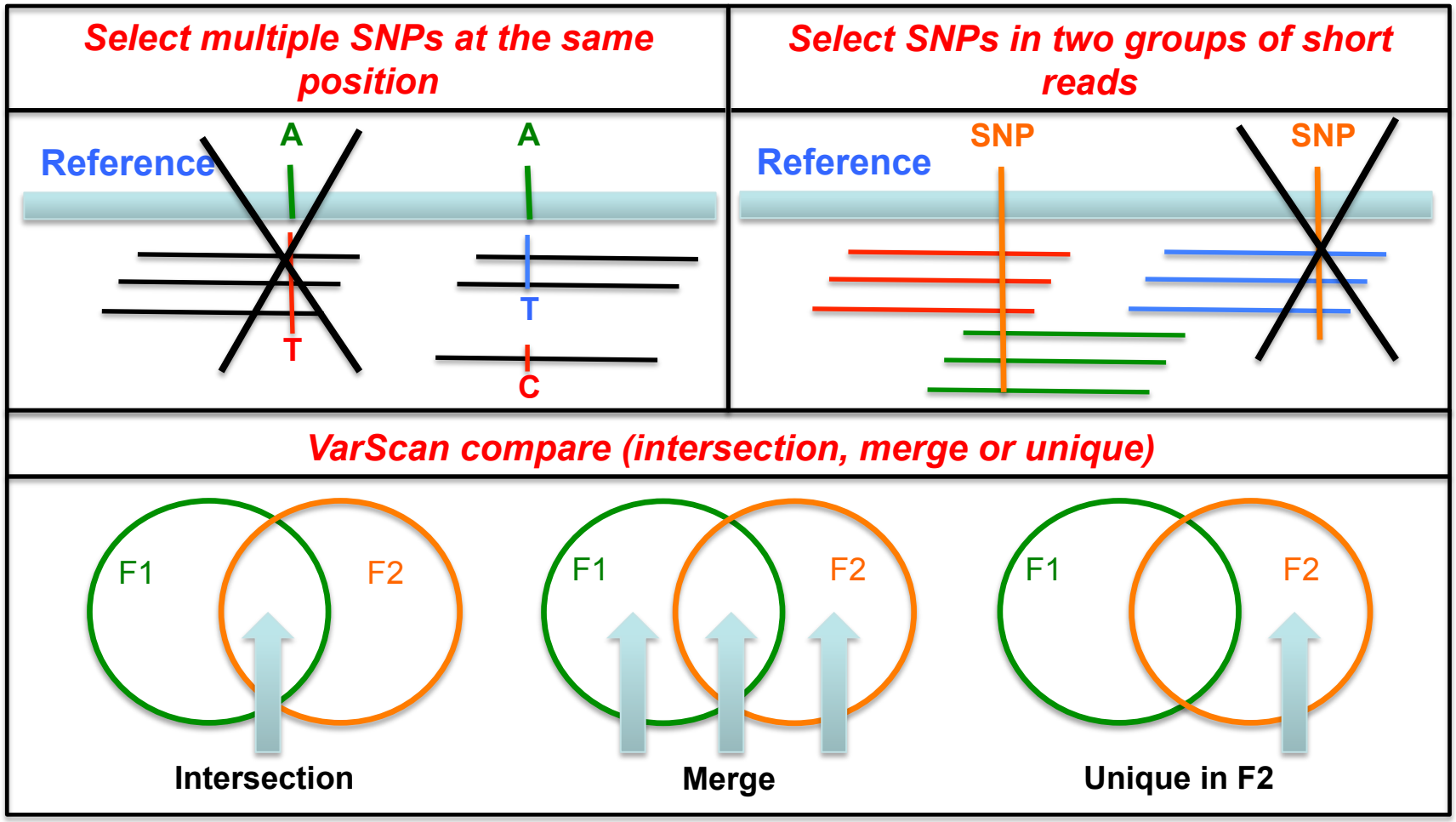
- **Preprocessing tools:**
  - Remove duplicated short-reads
  - Remove short reads not in paired-ends
  - Remove short reads > ‘N’%
  - Remove informations in each FASTA file header

- **Postprocessing tools:**
  - Count multiple hits from the results of BWA
  - Extract short reads from SAM file
  - VarScan compare (multiple files)
  - VarScan filter
  - VarScan to Gff3

B.3. MAPHiTS Development Tools



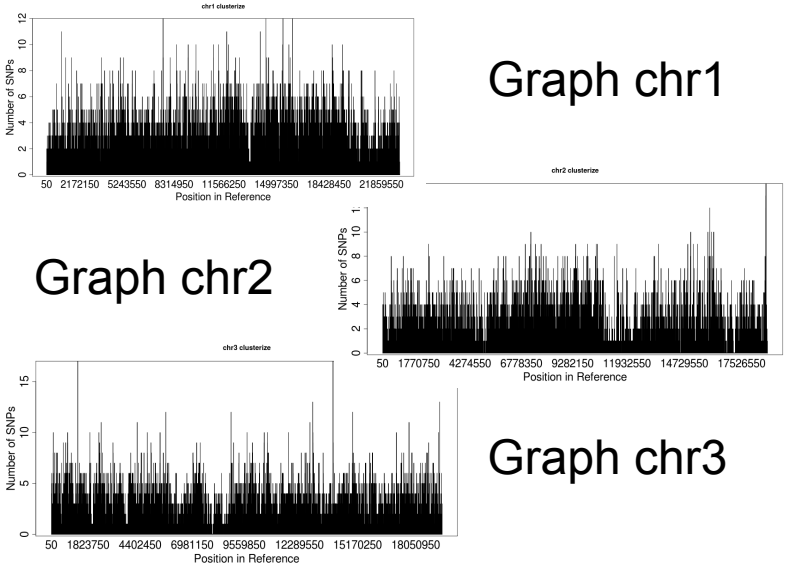
B.3. MAPHiTS Development Tools



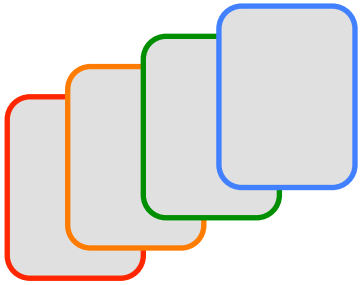
B.3. MAPHiTS Development Tools

Get SNPs distribution by chromosome

VarScan File



Give conscensus nucleotides (with different VarScan analysis)



VarScan Files

HTML File

INRA - URGI Team

MAPHiTS Genotype Card

Ref	Position	RefAllele	Trayshed	Vitis-Rotundifolia	Regale_run2	Regale_run1	Fry	Carlos	Consensus
chr18	801	G	T	T	G	G	T	T	NA
chr18	807	C	T	T	C	C	T	T	NA
chr18	977	A	A	A	A	A	A	T	NA
chr18	1060	T	T	T	T	T	T	A	NA
chr18	1751	A	A	A	A	A	A	G	NA
chr18	1752	A	A	A	A	A	A	T	NA
chr18	1786	A	A	A	A	A	A	G	NA
chr18	1890	T	T	T	T	T	T	G	NA
chr18	2586	T	T	T	T	T	T	G	NA
chr18	2591	T	G	G	T	T	T	G	NA
chr18	4499	A	G	G	G	G	G	G	G 6/6
chr18	4534	T	A	A	A	T	A	A	NA
chr18	4585	G	C	G	C	G	C	C	NA
chr18	4612	C	T	C	T	T	T	C	NA
chr18	4618	T	C	T	C	C	C	C	NA
chr18	4665	A	G	G	G	G	G	G	G 6/6
chr18	4754	T	C	C	C	C	C	C	C 6/6
chr18	4785	C	A	C	A	A	A	A	NA
chr18	4813	A	G	G	G	G	G	G	G 6/6
chr18	4824	A	T	T	T	T	T	T	T 6/6
chr18	4910	A	G	G	G	A	A	G	NA

INRA - URGI Team

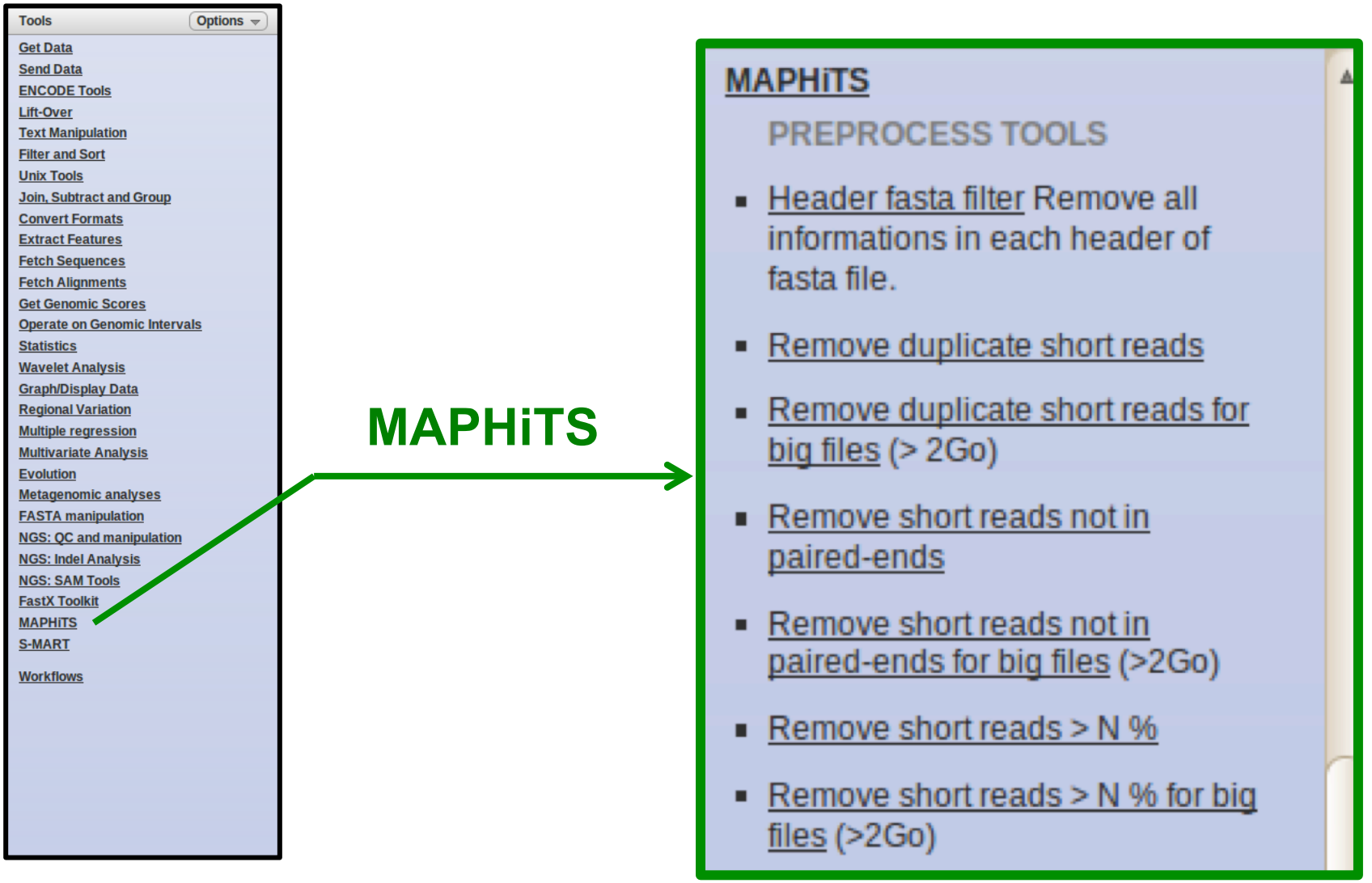
B.3. MAPHiTS Development Tools

MAPHiTS Genotype Card

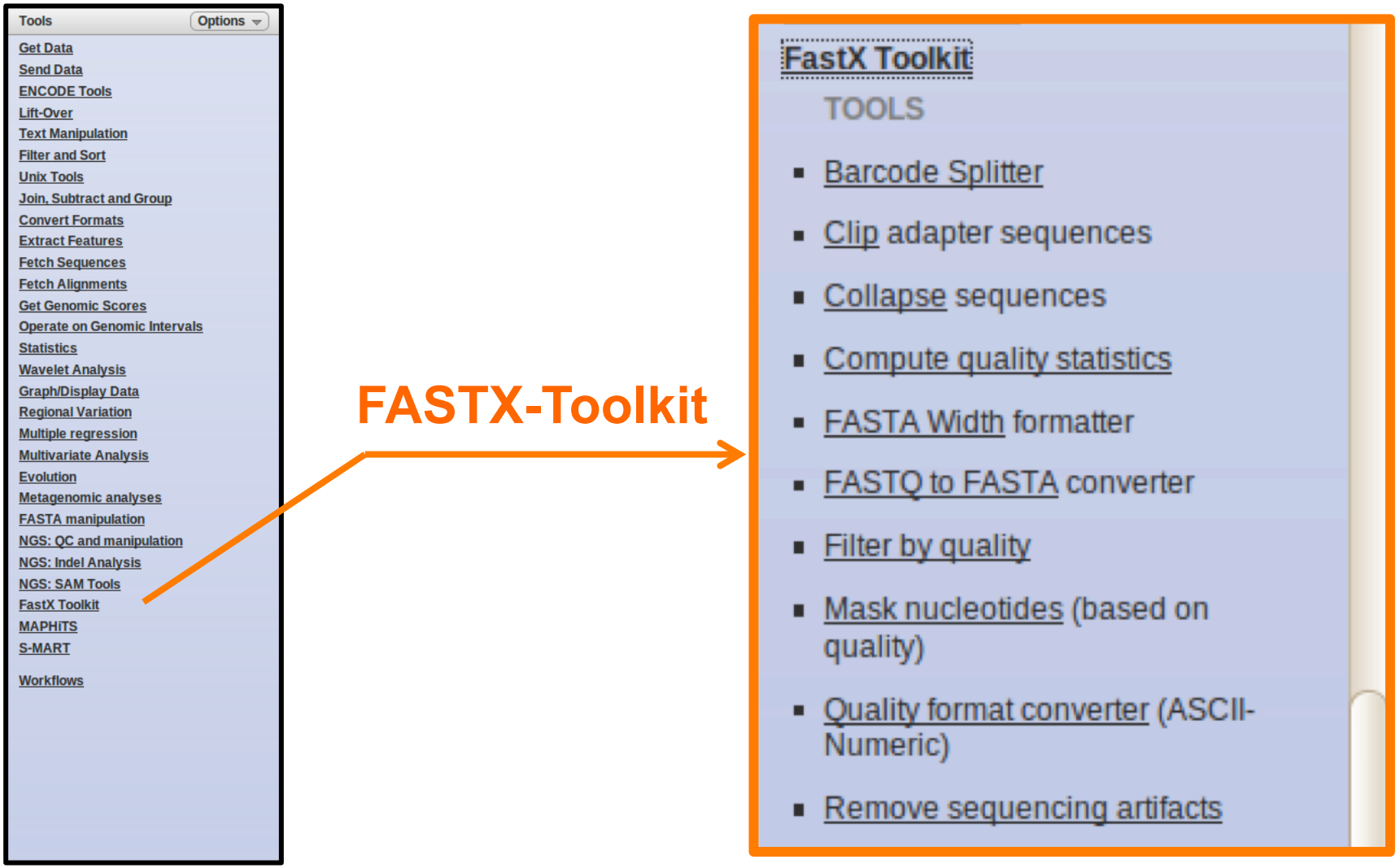
Ref	Position	RefAllele	Trayshed	Vitis-Rotundifolia	Regale_run2	Regale_run1	Fry	Carlos	Consensus
chr18	801	G	T	T	G	G	T	T	NA
chr18	807	C	T	T	C	C	T	T	NA
chr18	977	A	A	A	A	A	A	T	NA
chr18	1060	T	T	T	T	T	T	A	NA
chr18	1751	A	A	A	A	A	A	G	NA
chr18	1752	A	A	A	A	A	A	T	NA
chr18	1786	A	A	A	A	A	A	G	NA
chr18	1890	T	T	T	T	T	T	G	NA
chr18	2586	T	T	T	T	T	T	G	NA
chr18	2591	T	G	G	T	T	T	G	NA
chr18	4499	A	G	G	G	G	G	G	G 6/6
chr18	4534	T	A	A	A	T	A	A	NA
chr18	4585	G	C	G	C	G	C	C	NA
chr18	4612	C	T	C	T	T	T	C	NA
chr18	4618	T	C	T	C	C	C	C	NA
chr18	4665	A	G	G	G	G	G	G	G 6/6
chr18	4754	T	C	C	C	C	C	C	C 6/6
chr18	4785	C	A	C	A	A	A	A	NA
chr18	4813	A	G	G	G	G	G	G	G 6/6
chr18	4824	A	T	T	T	T	T	T	T 6/6
chr18	4910	A	G	G	G	A	A	G	NA

C. MAPHiTS in Galaxy

C.2. New URGI Integrated tools

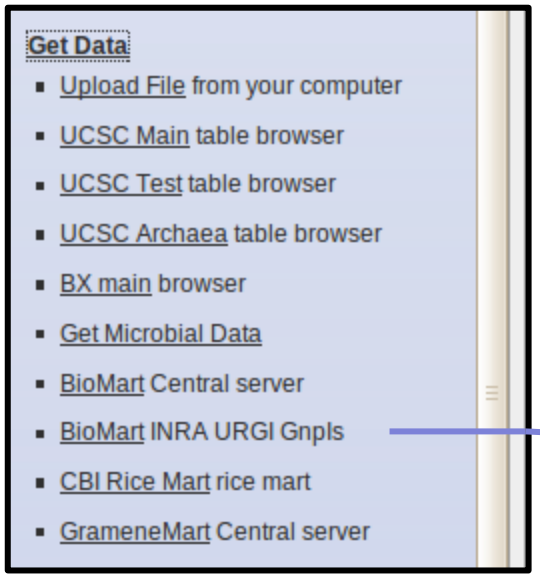


C.2. New Integrated tools

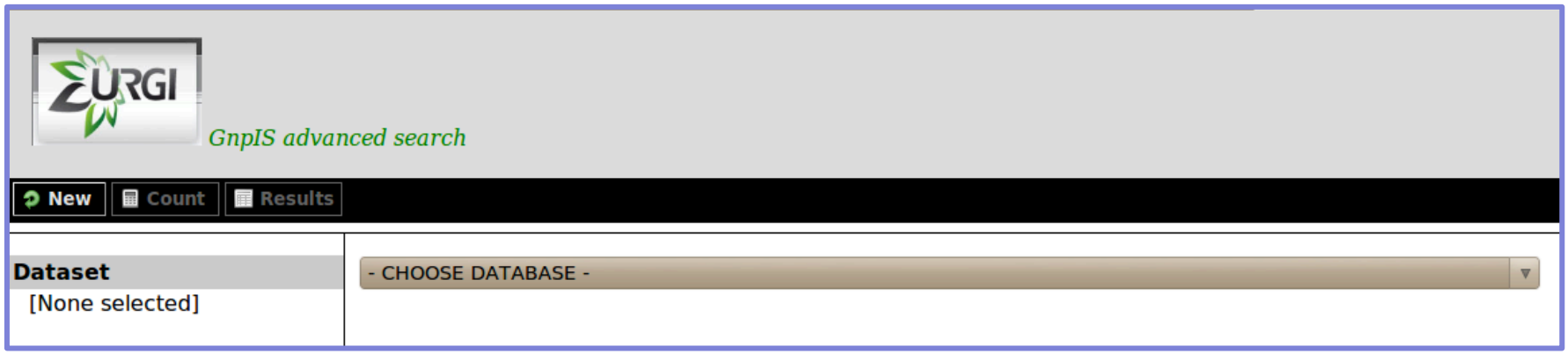


C.2. New Others URGI Integrated tools

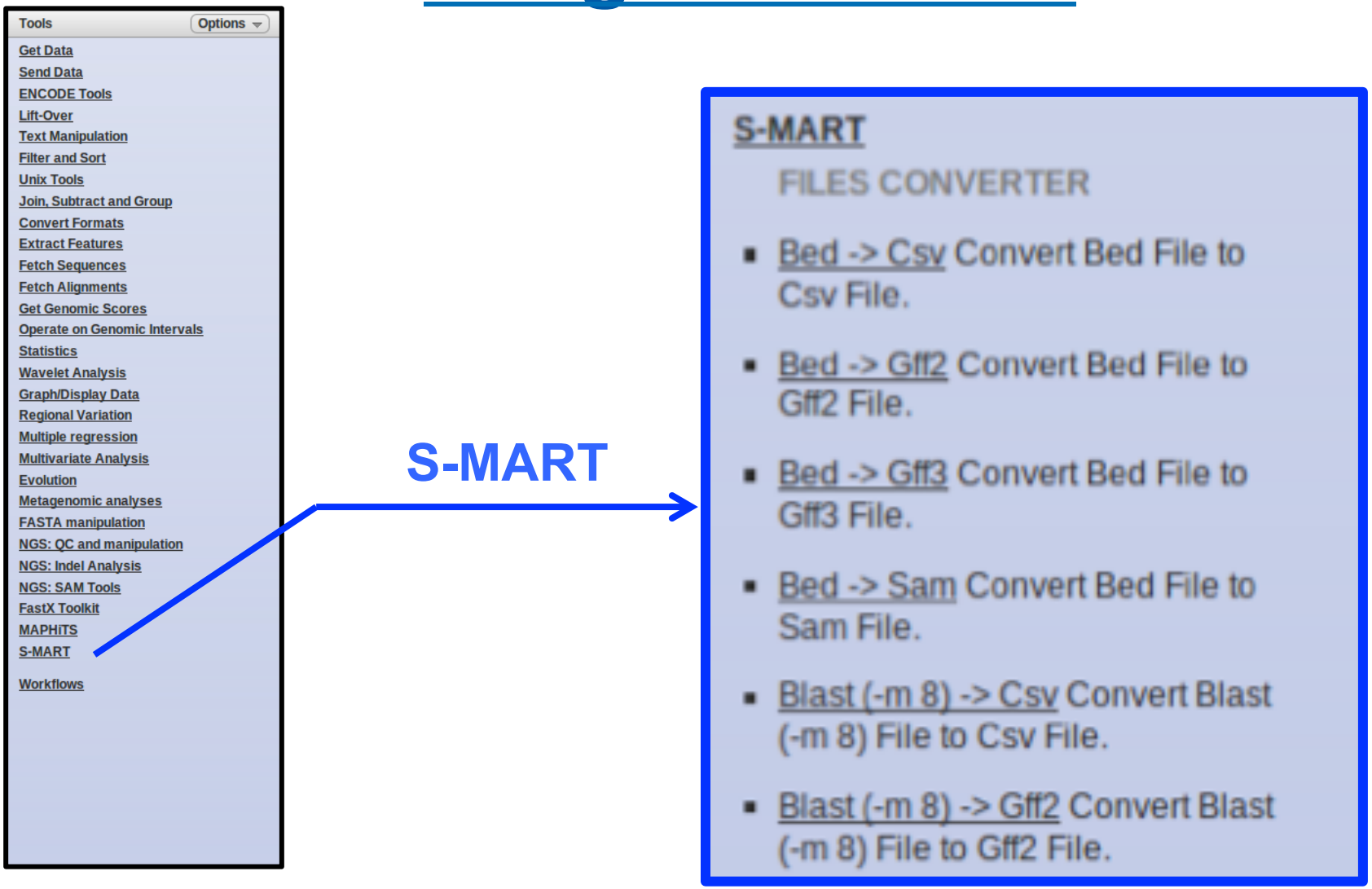
Access to URGI Information System via **BioMart** software



BioMart URGI Gnpls



C.2. New Others URGI Integrated tools



C.2. New Others URGI Integrated tools

What can I do with all this RNA-Seq data ?

S-MART:

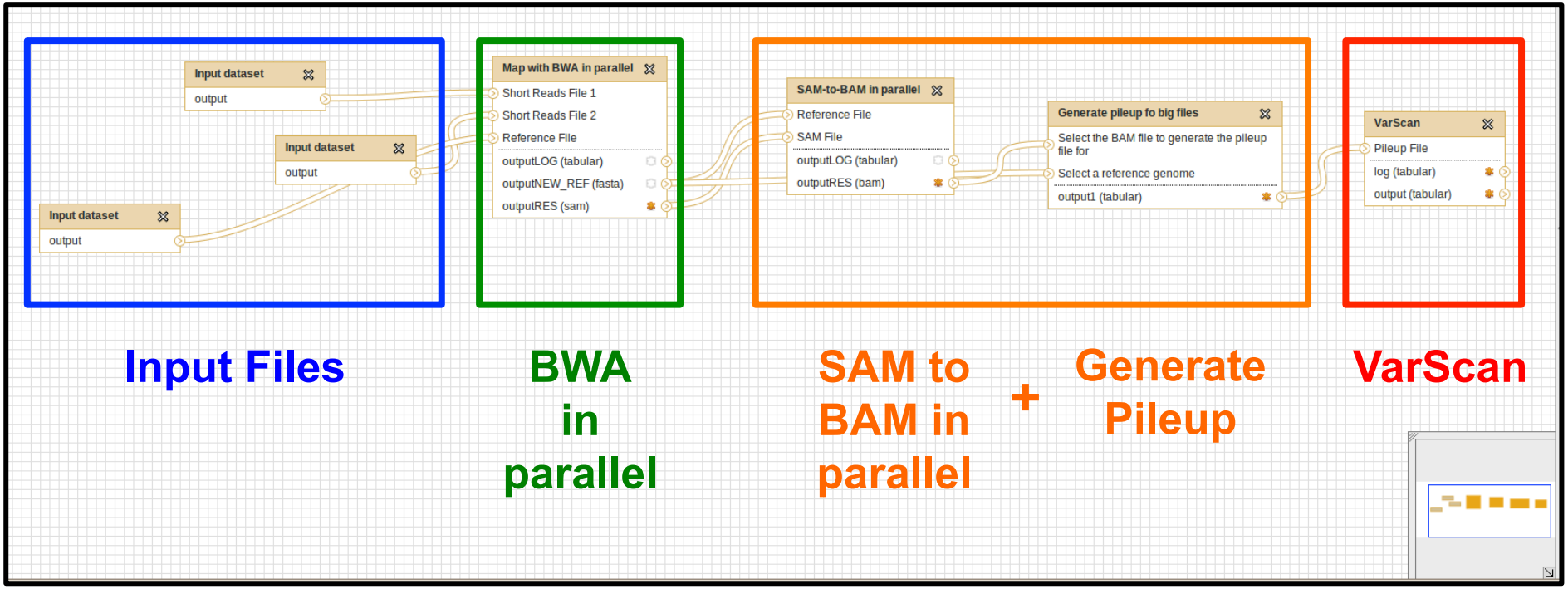
- is a set of independant tools
- works on a standard PC and with Galaxy
- can be installed and used easily

Use S-MART for data manipulation, data visualization, differential expression, …

Link: *http://urgi.versailles.inra.fr/Tools/S-MART*

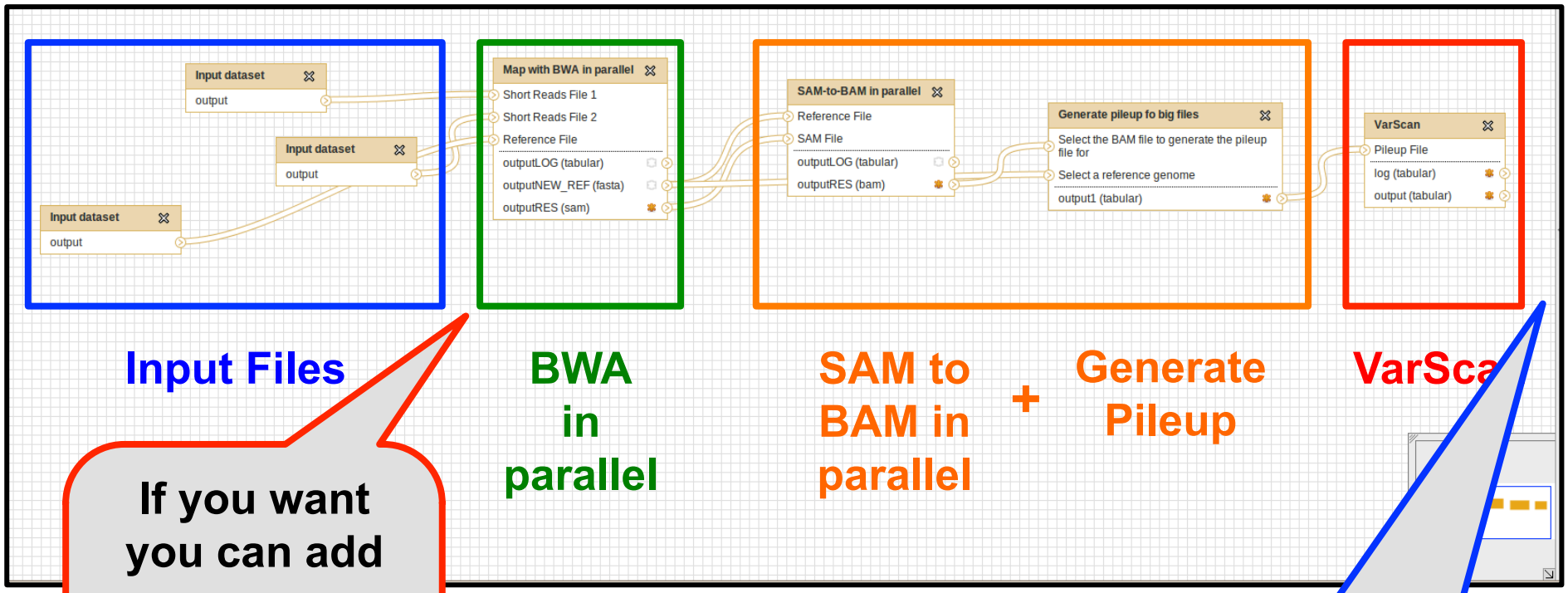
Contact: *matthias.zytnicki@versailles.inra.fr*

C.3. MAPHiTS: Build

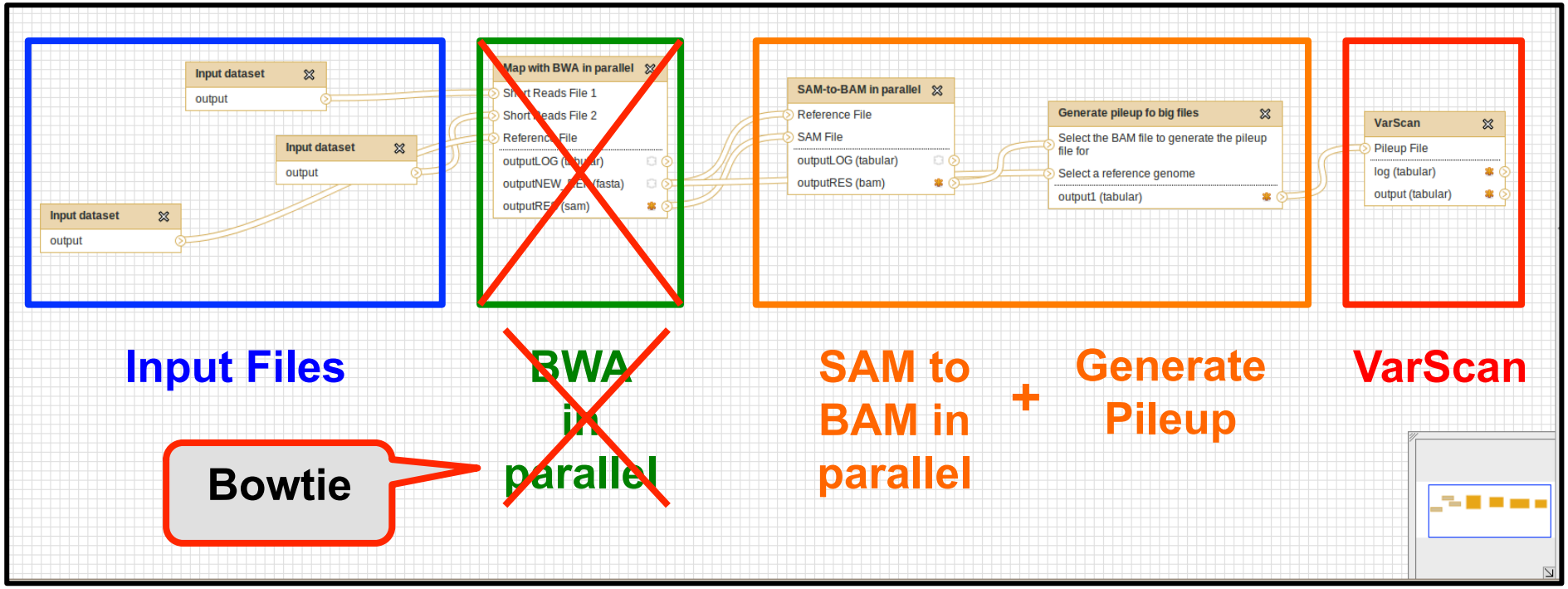


MAPHiTS is build using the graphical interface of Galaxy.

C.3. MAPHiTS: Build

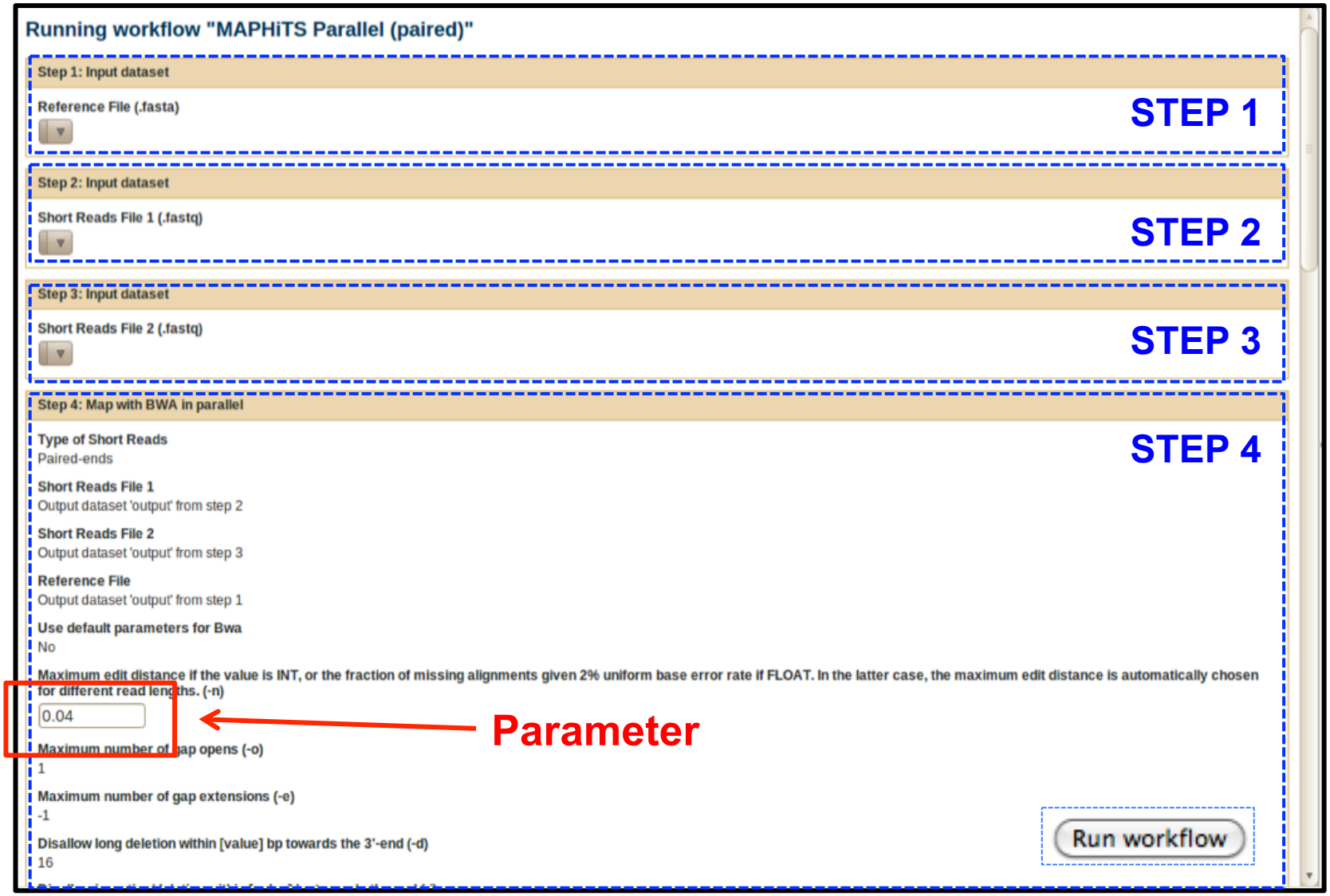


C.3. MAPHiTS: Build

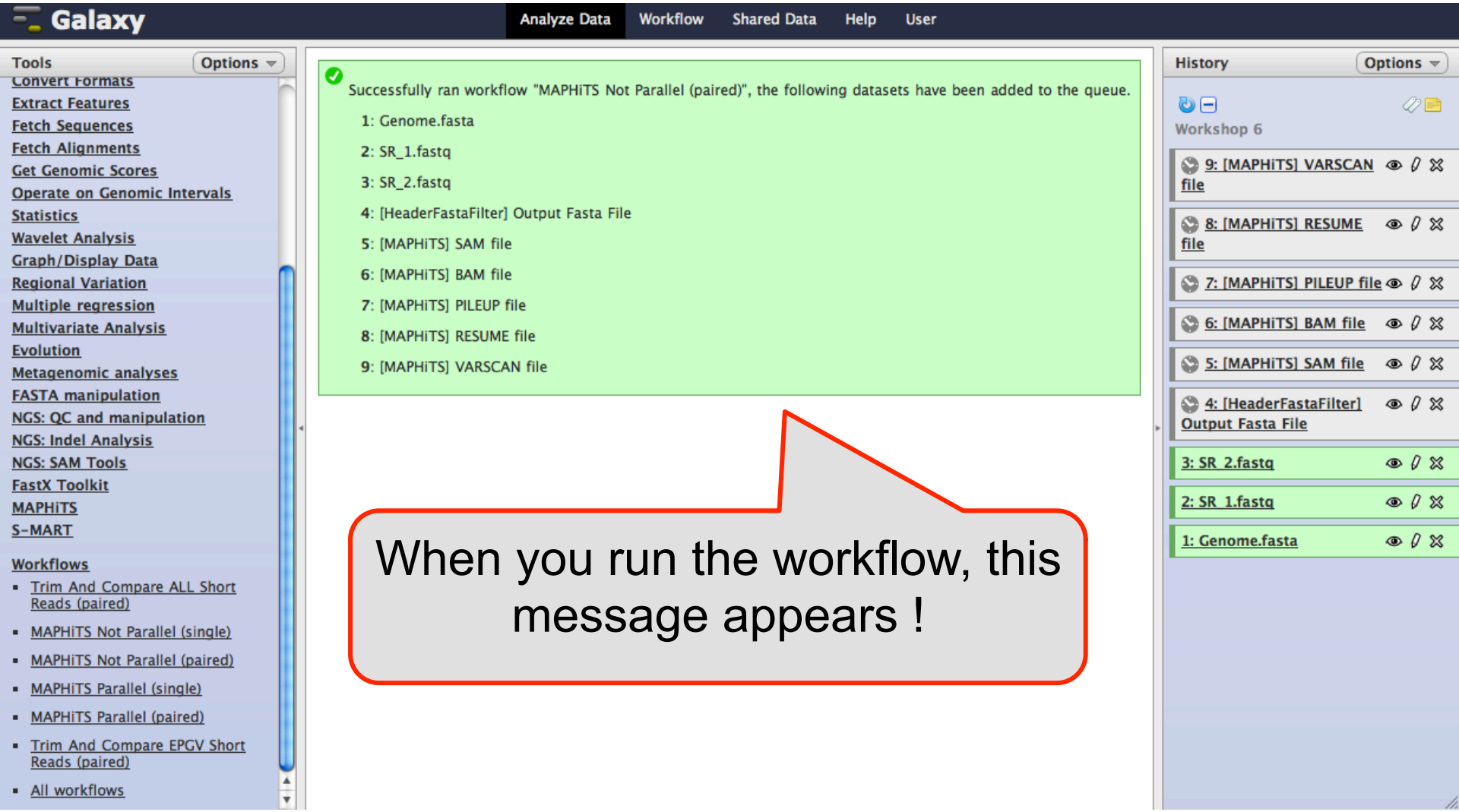


You can **remove** one tool **and replace** it by an other tool very **quickly**.

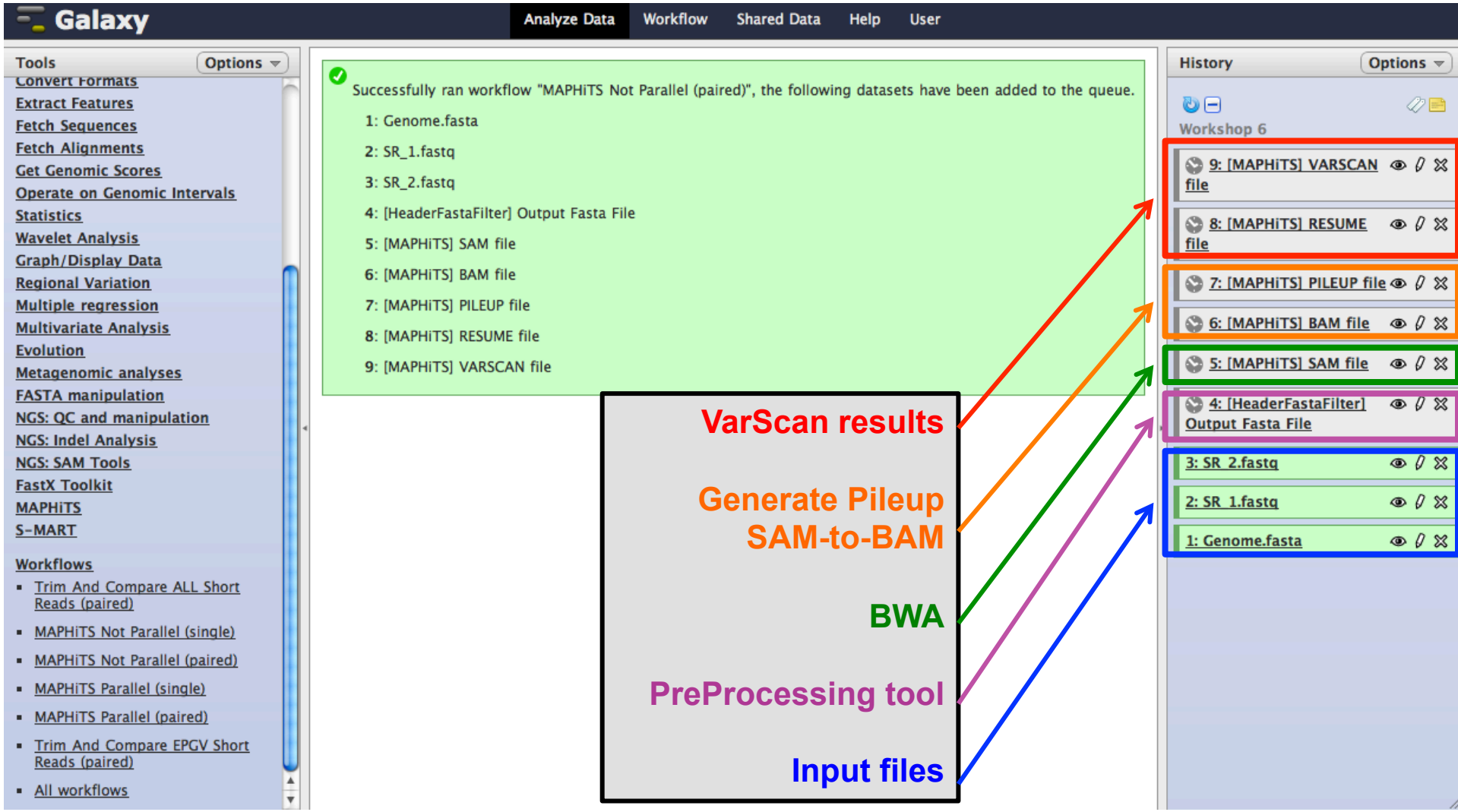
C.4. MAPHiTS: Launch



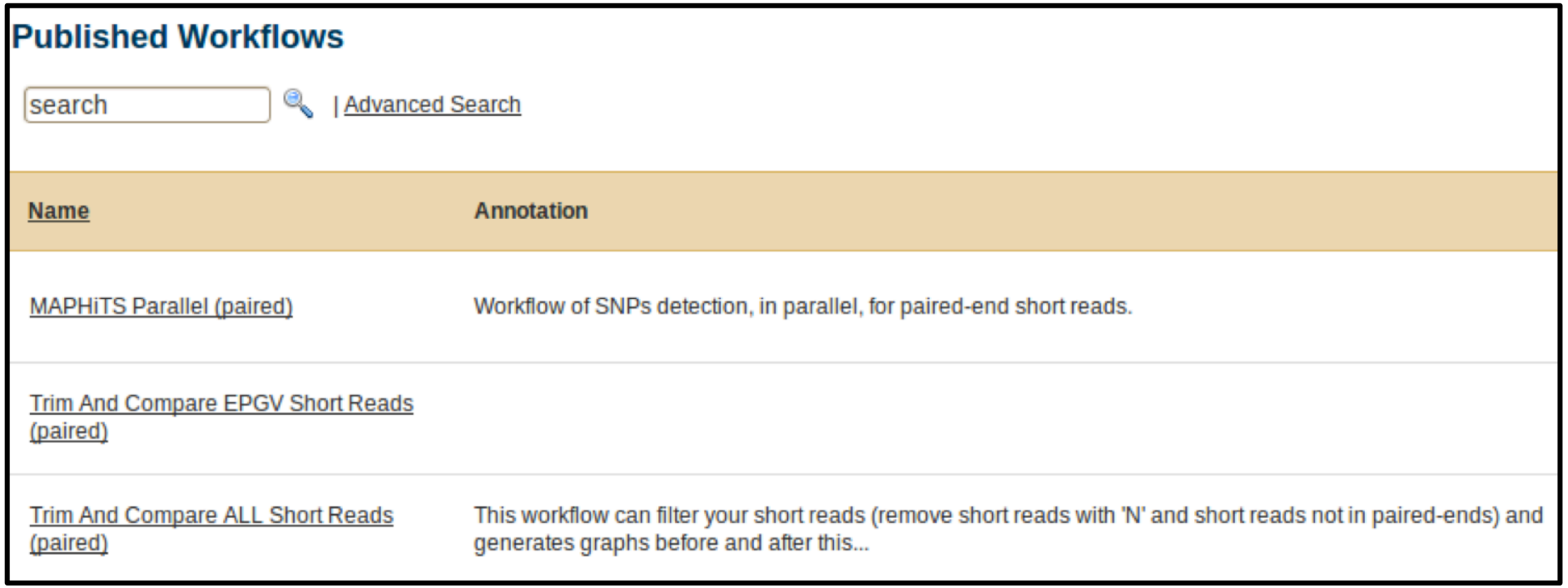
C.4. MAPHiTS: Launch



C.4. MAPHiTS: Launch



C.5. Shared Workflows



Published Workflows

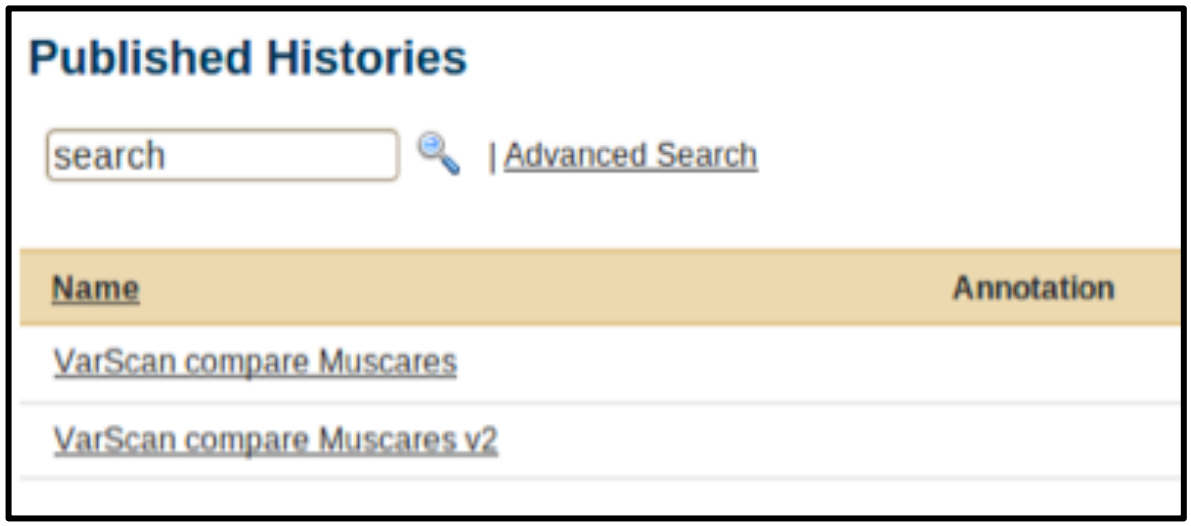
search | Advanced Search

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.

C.6. Shared your History

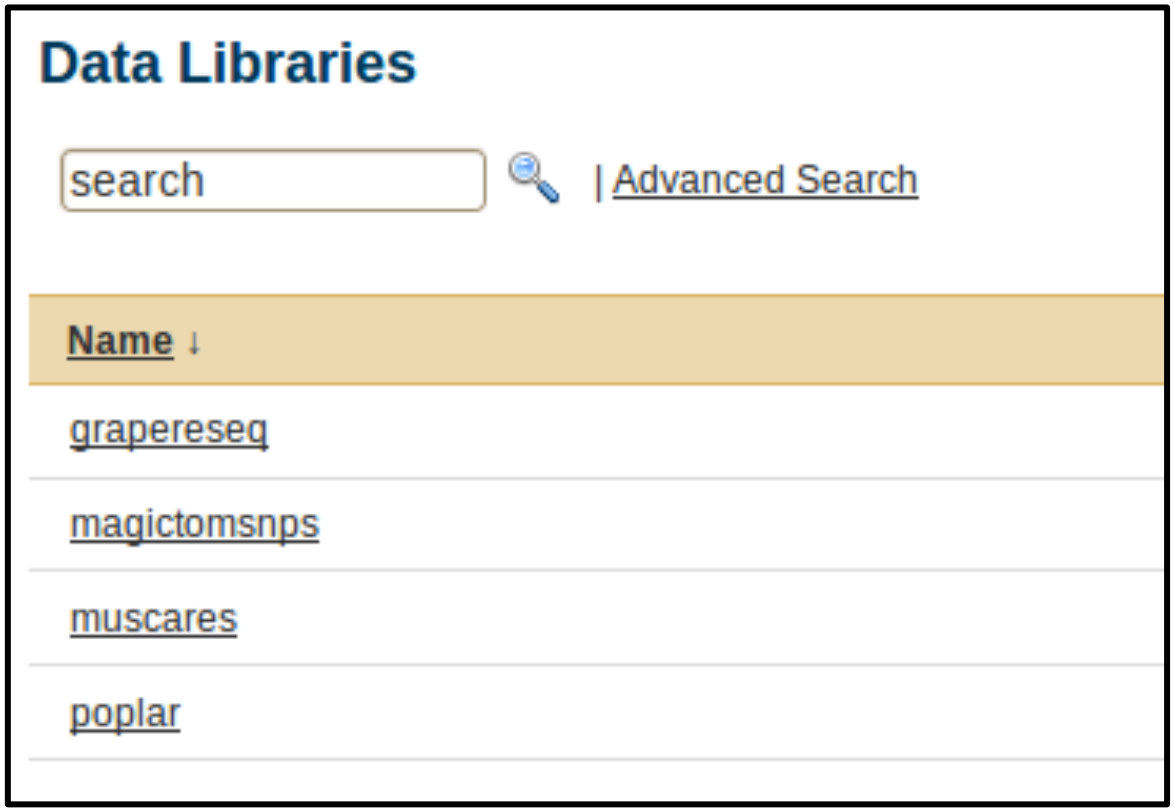
If a user wants to share its results with other users or a specific user, it's possible !



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

C.7. Shared Data

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



C.7. Shared Data

They can import their data into the history quickly.

➡ Usefull for **NGS** !

D. Perspectives

D. 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

- EPGV Team
- URGI Team

- Galaxy developers
- Galaxy community

Thank you for your attention !!!

BRAS Marc
Colloque EPGV 2011 – Monastir, Tunisie

ALIMENTATION
AGRICULTURE
ENVIRONNEMENT