

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

**MAPHiTS: Mapping Analysis Pipeline for High-Throughput Sequences**



# Overview

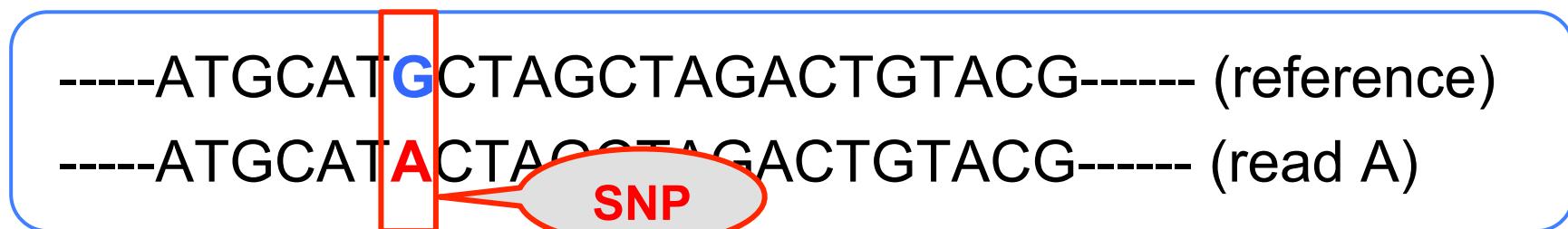
- I. Background and objectives of the pipeline
- II. Existing Tools
- III. MAPHiTS Development Tools
- IV. Integration of tools in Galaxy
- V. Preliminary Results
- VI. Perspectives

## 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 resistance 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 alignment (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/>

The screenshot shows the URGI website homepage. A red arrow points to the "GNPIIS PORTAL" section, which is highlighted with a red border. This section contains links to various bioinformatics tools: GnpMap, GnpSeq, GnpSnp, GnpGenome, SIREGal, GnpArray, Ephesis, GnpProt, and GnpSNP. Below this is a large image of a plant larva. To the right of the portal, there is a "WHAT'S NEW?" section listing updates for June 2011 and May 2011, each with a green gear icon and a link to a private data server.

URGI - Unité de Recherche Génomique Info is a research unit in genomics and bioinformatics at Institut National de la Recherche Agronomique (INRA), dedicated to plants and crop parasites. The URGI [research activity](#) covers genome structure and dynamics. URGI hosts a [bioinformatics platform](#), which belongs to the French national network of bioinformatics platforms ([ReNaBi](#)).

L'URGI est une unité de recherche en génomique et bio-informatique de l'Institut National de la Recherche Agronomique (INRA), dédiée à la génomique des plantes et de leurs pathogènes. Son [activité de recherche](#) porte sur la structure et la dynamique du génome. L'unité héberge une [plate-forme bioinformatique](#) appartenant au REseau NAtional des plateformes Bio-informatiques ([ReNaBi](#)).

EVENTS

06 Jul 2011 Transposable Elements in Marine Stramenopiles ...  
COM (talks) CNET, XVIIe edition 4th - 6th july 2011, Lyon France ...

04 Jun 2011 Vitis vinifera annotation jamboree : Apollo training and annotation jamboree on Vitis vinifera genome sequence We organize an Apollo ...

SEARCH  OK

WHAT'S NEW ? RSS

24 Jun 2011 GnPIIS update  
[GnPIIS 1.6.7](#) is now available.   
Also available on [private data server](#)  
[GnpMap 2.5.4](#) is now available.   
Also available on [private data server](#)  
[GnpSNP 1.9.7](#) is now available.   
Also available on [private data server](#)

03 May 2011 GnPIIS update  
[GnPIIS 1.6.6](#) is now available.   
Also available on [private data server](#)  
[GnpArray 1.8.7](#) is now available.   
Also available on [private data server](#)  
[GnpMap 2.5.3](#) is now available.   
Also available on [private data server](#)  
[GnpSnp 1.9.6](#) is now available.   
Also available on [private data server](#)  
[Siregal 1.6.6](#) is now available.   
Also available on [private data server](#)

# IV. Integration of tools in Galaxy

**GnplIS - Genetic & Genomic Information System**

Quick search  
You can find the indexed databases list [here](#).  
Examples: [VVI\\*](#), [VVIF52](#), [gene](#), [arabidopsis](#), [AY109603](#), [Xwmc430-3B](#)

Advanced search  
BioMart  
**Galaxy**

Category  Submit

Specific modules –

Genetic maps and QTLs

EST and other sequences

Polymorphism data

Plant genetic resource data

Transcriptome data

Proteomic data

Genome annotation data

GnpSeq GnpMap GnpSNP SIReGal Ephesia GnpProt GnpArray GnpGenome



# IV. Integration of tools in Galaxy

Galaxy

Analyze Data Workflow Shared Data Help User

Tools Options

- Get Data
- Send Data
- ENCODE Tools
- Lift-Over
- Text Manipulation
- Filter and Sort
- Unix Tools
- Join, Subtract and Group
- Convert Formats
- Extract Features
- Fetch Sequences
- Fetch Alignments
- Get Genomic Scores
- Operate on Genomic Intervals
- Statistics
- Wavelet Analysis
- Graph/Display Data
- Regional Variation
- Multiple regression
- Multivariate Analysis
- Evolution
- Metagenomic analyses
- FASTA manipulation
- NGS: QC and manipulation
- NGS: Indel Analysis
- NGS: SAM Tools
- FastX Toolkit
- MAPHITS
- S-MART
- Workflows

Unité Recherche Génomique Info



The Galaxy team is a part of BX at Penn State.  
This project is supported in part by NSF, NHGRI, and the Huck Institutes of the Life Sciences.

INRA

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

History Options

Unnamed history

Your history is empty. Click 'Get Data' on the left pane to start

# IV. Integration of tools in Galaxy

The screenshot shows the Galaxy web interface. On the left, a red box highlights the 'Tools' sidebar, which lists various genomic analysis tools. In the center, a green box highlights the main workspace where the URGI logo and INRA branding are displayed. On the right, a blue box highlights the 'History' panel, which is currently empty.

**TOOLS LIST**

**Unité Recherche Génomique Info**

**URGI**

**INRA**

**HISTORY**

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

## IV.1. Installation of URGI Galaxy

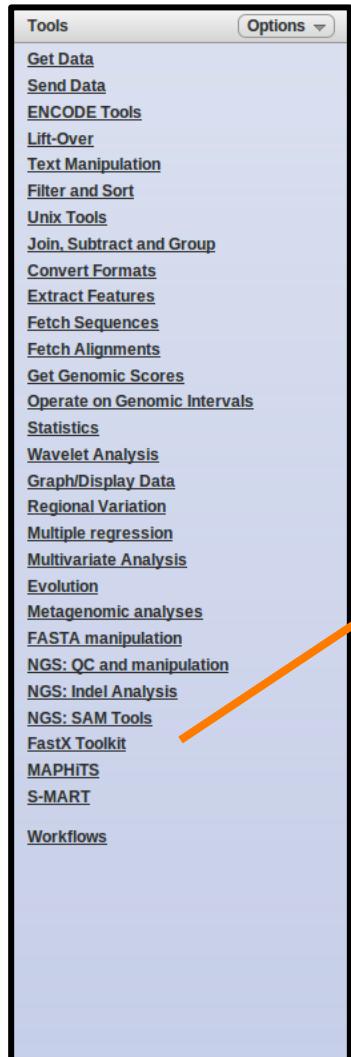
Galaxy is installed on URGI cluster with:

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



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

## IV.2. New Integrated tools



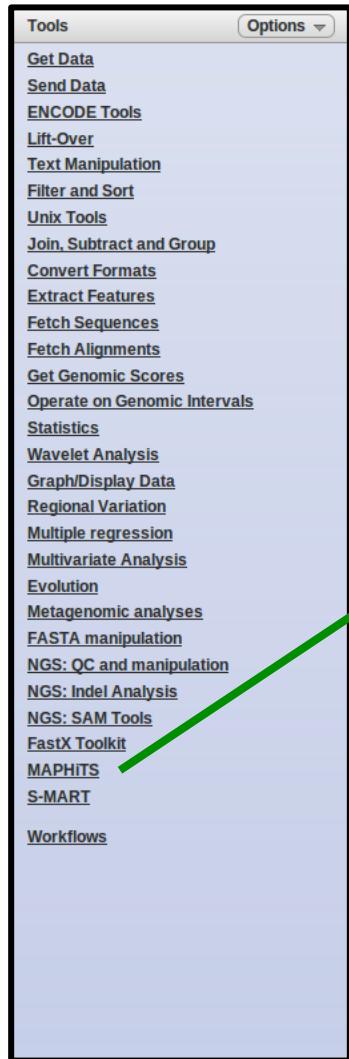
### FASTX-Toolkit

**FastX Toolkit**

**TOOLS**

- [Barcode Splitter](#)
- [Clip adapter sequences](#)
- [Collapse sequences](#)
- [Compute quality statistics](#)
- [FASTA Width formatter](#)
- [FASTQ to FASTA converter](#)
- [Filter by quality](#)
- [Mask nucleotides \(based on quality\)](#)
- [Quality format converter \(ASCII-Numeric\)](#)
- [Remove sequencing artifacts](#)

## IV.2. New URGI Integrated tools



### MAPHiTS

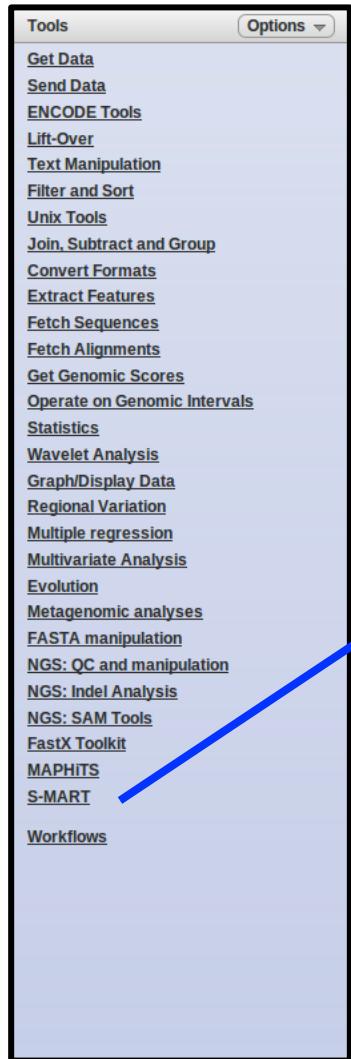
**MAPHiTS**

#### PREPROCESS TOOLS

- Header fasta filter Remove all informations in each header of fasta file.
- Remove duplicate short reads
- Remove duplicate short reads for big files (> 2Go)
- Remove short reads not in paired-ends
- Remove short reads not in paired-ends for big files (>2Go)
- Remove short reads > N %
- Remove short reads > N % for big files (>2Go)

## IV.2. New Others URGI

# Integrated tools



S-MART

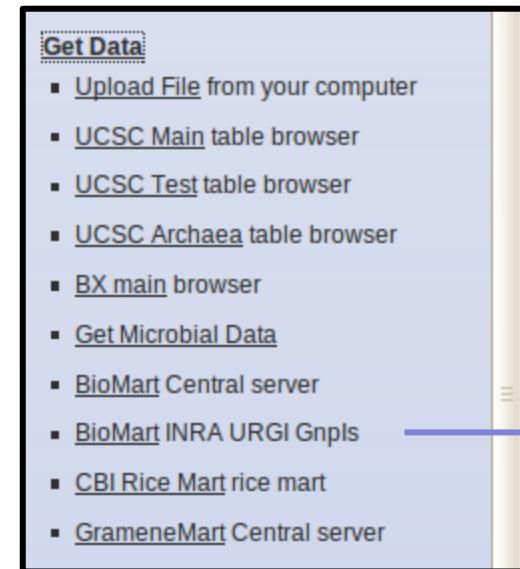
### S-MART

#### FILES CONVERTER

- [Bed -> Csv](#) Convert Bed File to Csv File.
- [Bed -> Gff2](#) Convert Bed File to Gff2 File.
- [Bed -> Gff3](#) Convert Bed File to Gff3 File.
- [Bed -> Sam](#) Convert Bed File to Sam File.
- [Blast \(-m 8\) -> Csv](#) Convert Blast (-m 8) File to Csv File.
- [Blast \(-m 8\) -> Gff2](#) Convert Blast (-m 8) File to Gff2 File.

## IV.2. New Others URGI Integrated tools

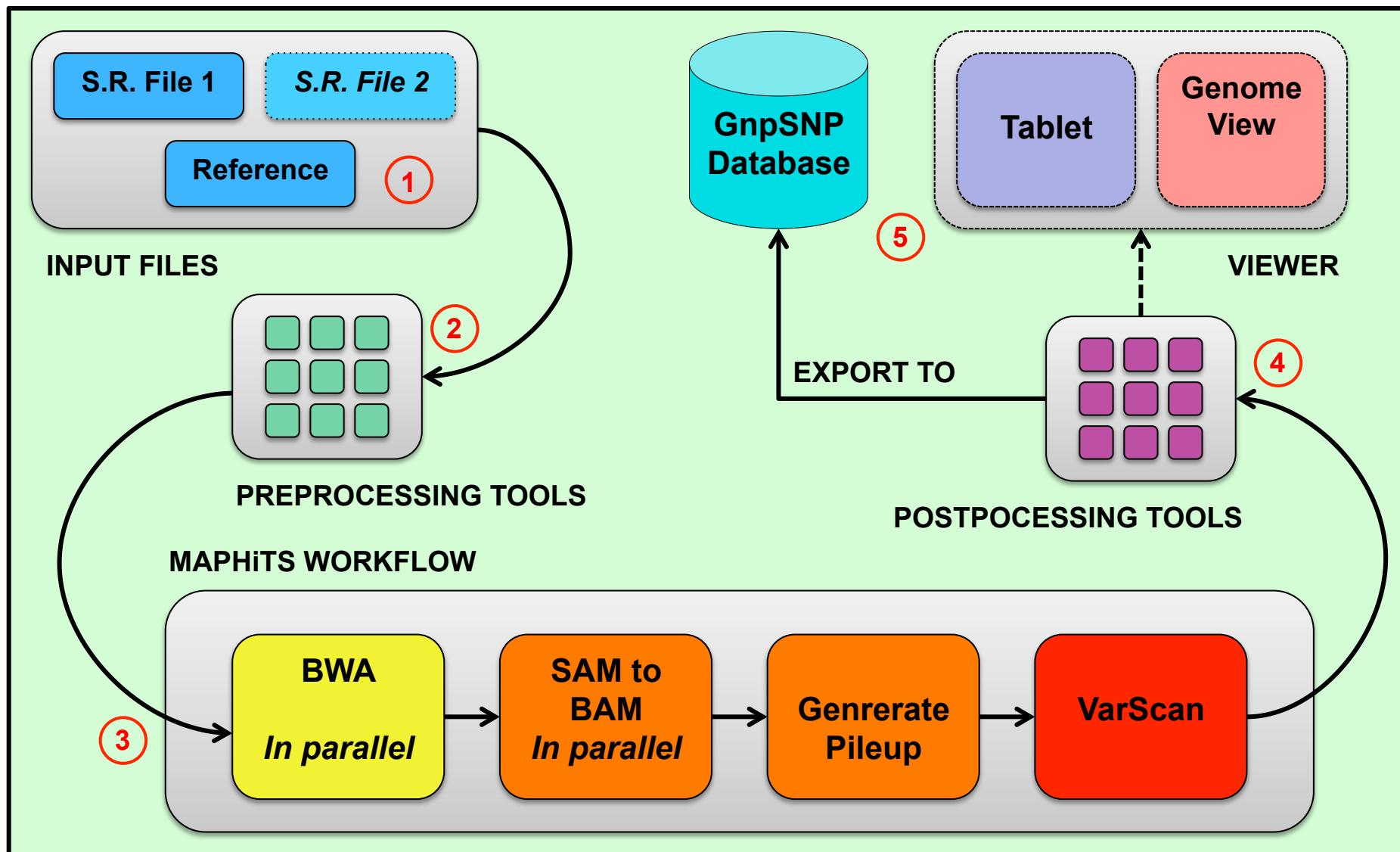
Access to URGI  
Information System  
via **BioMart** software



**BioMart**  
**URGI**  
**Gnpls**

The screenshot shows the 'GnpIS advanced search' interface. At the top left is the URGI logo. Below it is a navigation bar with buttons for 'New', 'Count', and 'Results'. On the left, there is a 'Dataset' dropdown menu showing '[None selected]'. To the right of the dropdown is a 'CHOOSE DATABASE' dropdown menu. The main area is currently empty.

# IV.3. MAPHiTS: Resume

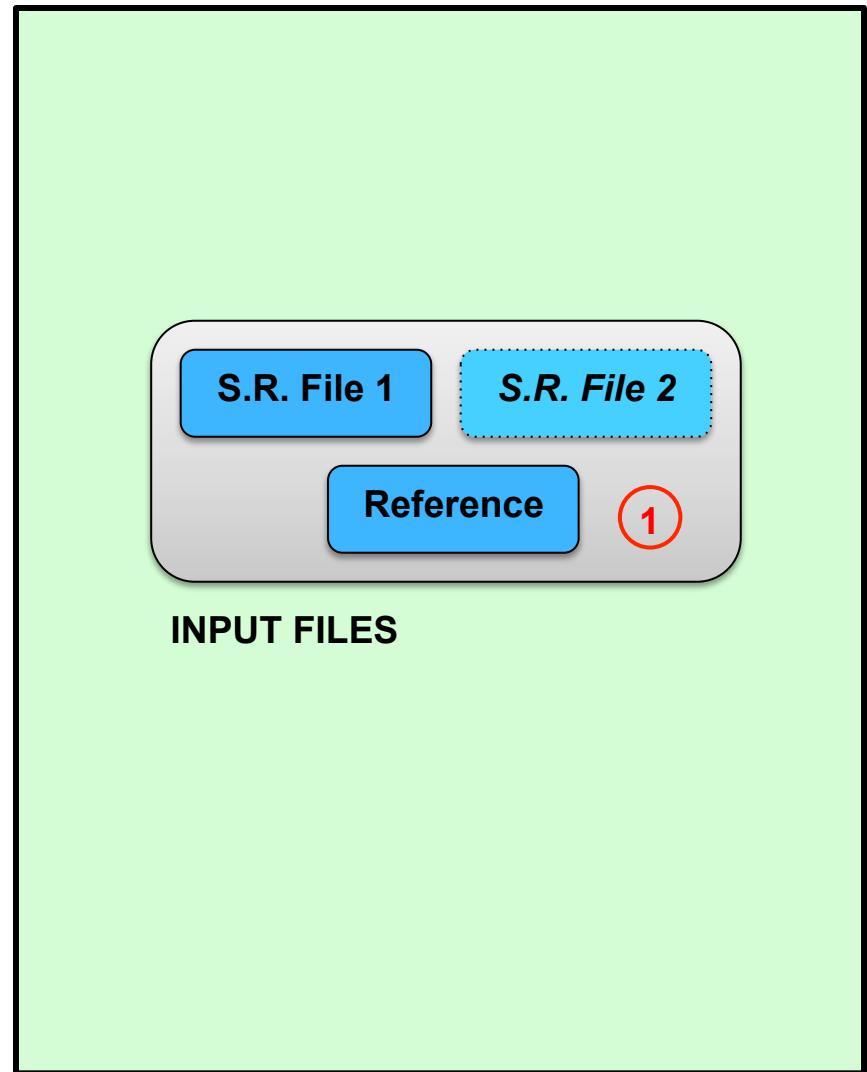


## IV.3. MAPHiTS: Resume

### Step 1

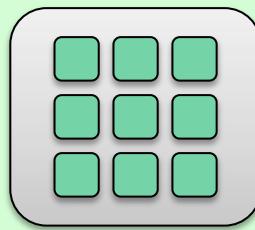
Upload your input files:

- 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



PREPROCESSING TOOLS

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

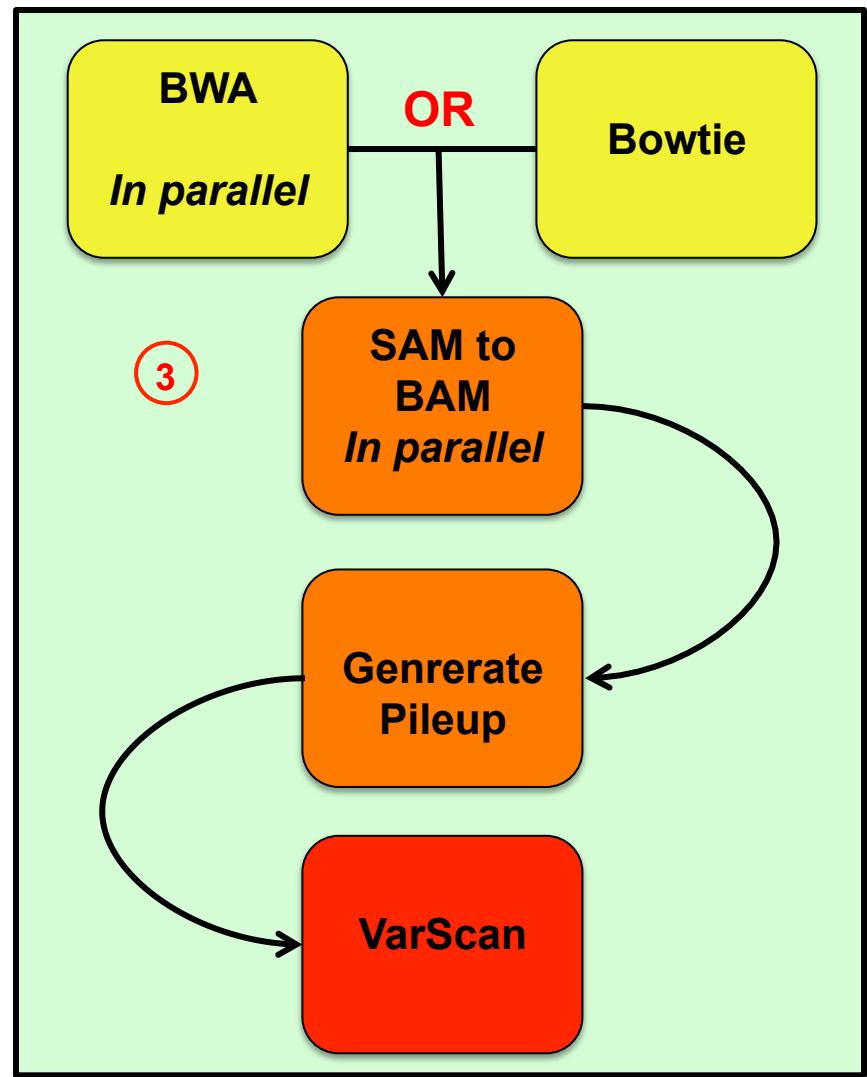
#### Examples:

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

## IV.3. MAPHiTS: Resume

### Step 3

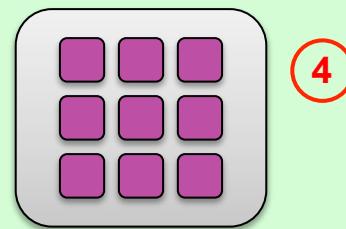
You have to launch  
MAPHiTS workflow.



## IV.3. MAPHiTS: Resume

### Step 4

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



POSTPROCESSING TOOLS

#### Examples:

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

## IV.3. MAPHiTS: Resume

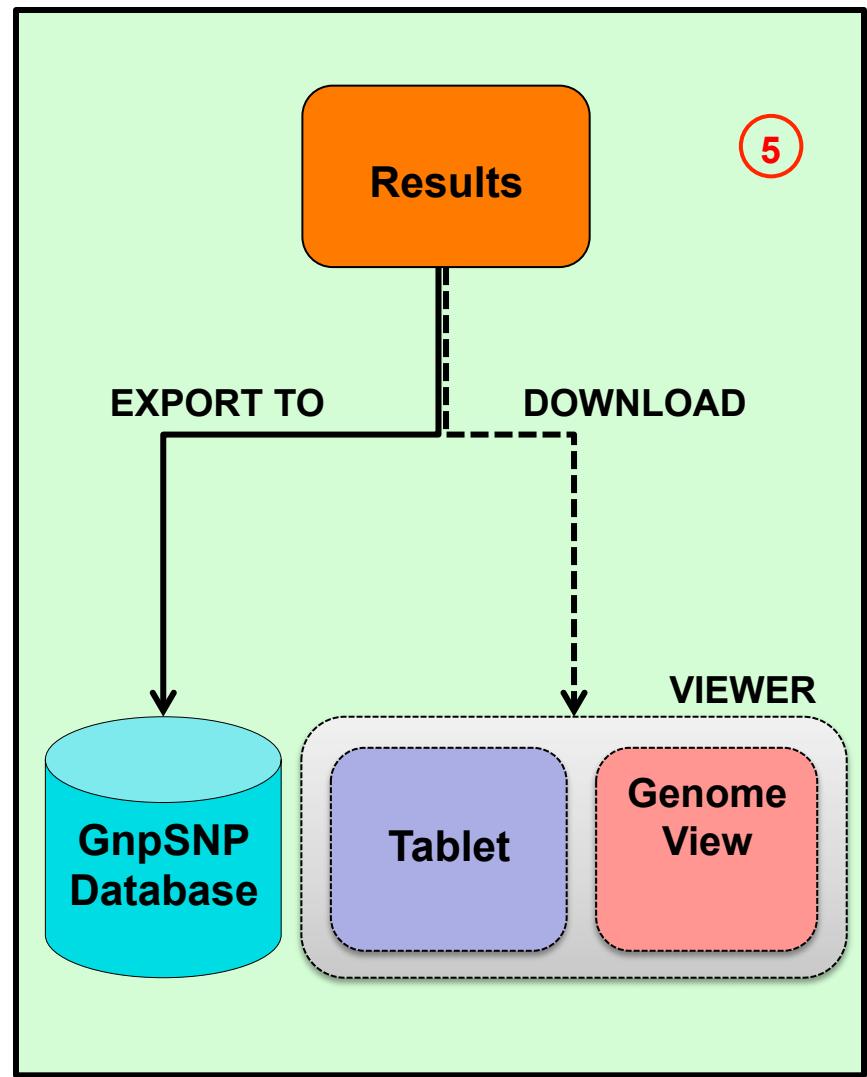
### Step 5

Finally, you can :

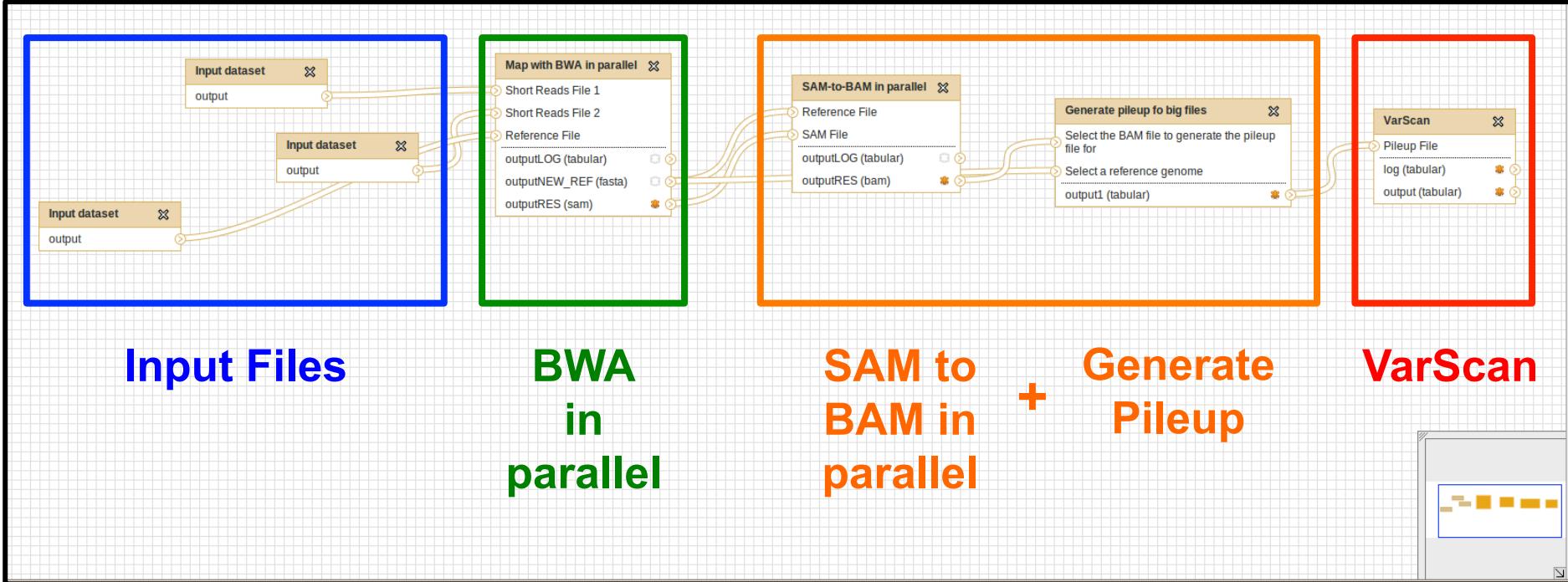
- Export your results to our GnpSNP database

**OR / AND**

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

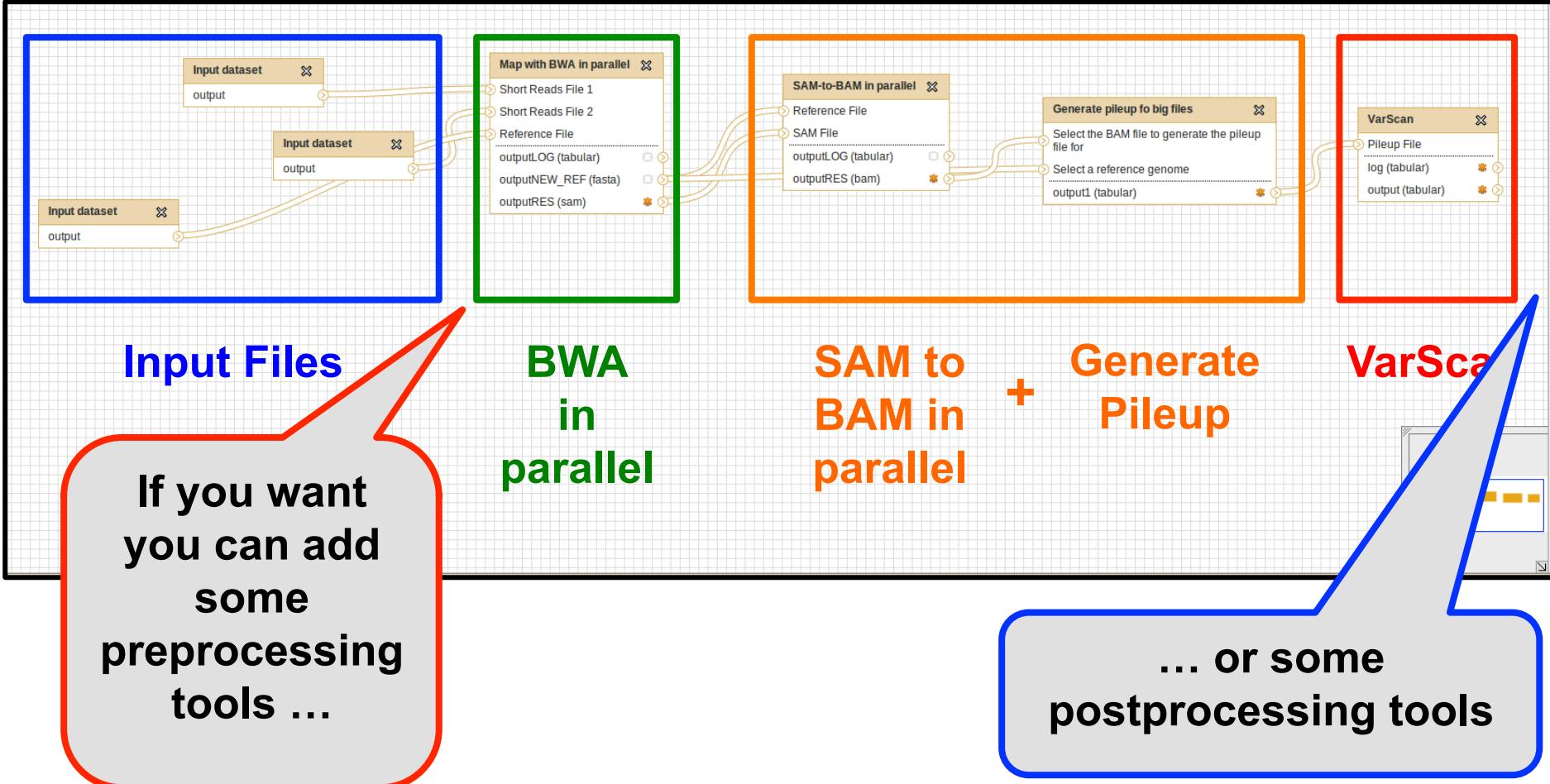


# IV.3. MAPHiTS: Build

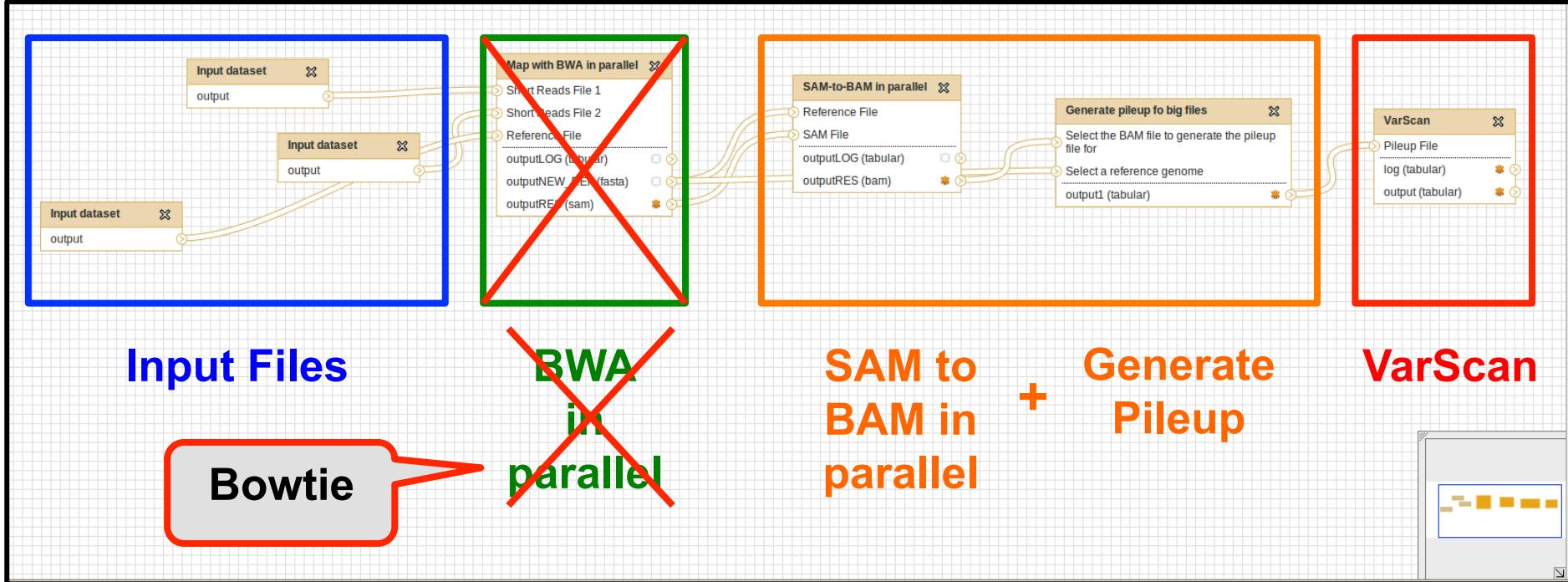


MAPHiTS is build using the graphical interface of Galaxy.

# 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

Running workflow "MAPHiTS Parallel (paired)"

Step 1: Input dataset

Reference File (.fasta)

▼

STEP 1

Step 2: Input dataset

Short Reads File 1 (.fastq)

▼

STEP 2

Step 3: Input dataset

Short Reads File 2 (.fastq)

▼

STEP 3

Step 4: Map with BWA in parallel

Type of Short Reads  
Paired-ends

Short Reads File 1  
Output dataset 'output' from step 2

Short Reads File 2  
Output dataset 'output' from step 3

Reference File  
Output dataset 'output' from step 1

Use default parameters for Bwa  
No

Maximum edit distance if the value is INT, or the fraction of missing alignments given 2% uniform base error rate if FLOAT. In the latter case, the maximum edit distance is automatically chosen for different read lengths. (-n)

0.04

Parameter

Maximum number of gap opens (-o)  
1

Maximum number of gap extensions (-e)  
-1

Disallow long deletion within [value] bp towards the 3'-end (-d)  
16

Run workflow

## IV.3. MAPHiTS: Launch

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

Step 4: Map with BWA in parallel

Type of Short Reads  
Paired-ends

Short Reads File 1  
Output dataset 'output' from step 2

Short Reads File 2  
Output dataset 'output' from step 3

Reference File  
Output dataset 'output' from step 1

Use default parameters for Bwa  
No

Maximum edit distance if the value is INT, or the fraction of missing alignments given 2% uniform base error rate if FLOAT. In the latter case, the maximum edit distance is automatically chosen for different read lengths. (-n)

0.04

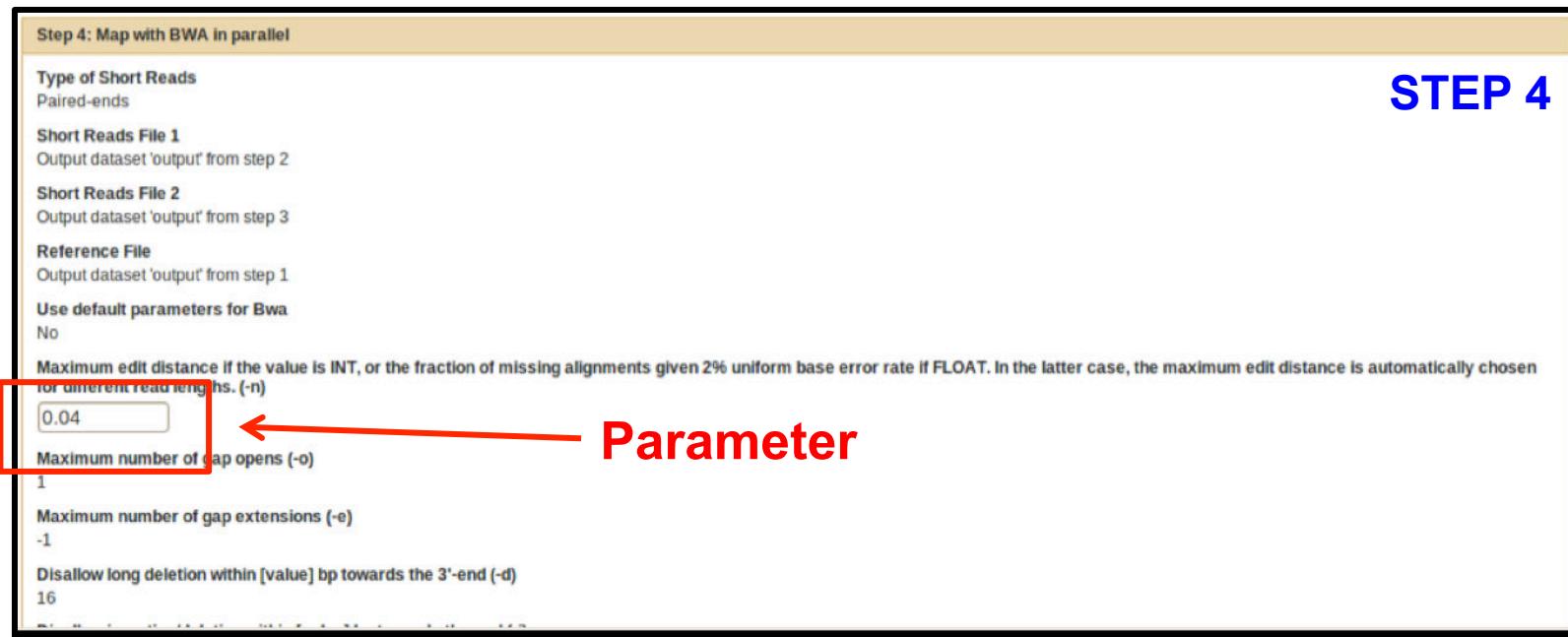
Maximum number of gap opens (-o)  
1

Maximum number of gap extensions (-e)  
-1

Disallow long deletion within [value] bp towards the 3'-end (-d)  
16

**STEP 4**

**Parameter**



# IV.3. MAPHiTS: Launch

Galaxy

Analyze Data Workflow Shared Data Help User

Tools

- [Convert Formats](#)
- [Extract Features](#)
- [Fetch Sequences](#)
- [Fetch Alignments](#)
- [Get Genomic Scores](#)
- [Operate on Genomic Intervals](#)
- [Statistics](#)
- [Wavelet Analysis](#)
- [Graph/Display Data](#)
- [Regional Variation](#)
- [Multiple regression](#)
- [Multivariate Analysis](#)
- [Evolution](#)
- [Metagenomic analyses](#)
- [FASTA manipulation](#)
- [NGS: QC and manipulation](#)
- [NGS: Indel Analysis](#)
- [NGS: SAM Tools](#)
- [FastX Toolkit](#)
- [MAPHiTS](#)
- [S-MART](#)
- [Workflows](#)
  - [Trim And Compare ALL Short Reads \(paired\)](#)
  - [MAPHiTS Not Parallel \(single\)](#)
  - [MAPHiTS Not Parallel \(paired\)](#)
  - [MAPHiTS Parallel \(single\)](#)
  - [MAPHiTS Parallel \(paired\)](#)
  - [Trim And Compare EPGV Short Reads \(paired\)](#)
  - [All workflows](#)

Options

History Options

Workshop 6

9: [MAPHiTS] VARSCAN file

8: [MAPHiTS] RESUME file

7: [MAPHiTS] PILEUP file

6: [MAPHiTS] BAM file

5: [MAPHiTS] SAM file

4: [HeaderFastaFilter] Output Fasta File

3: SR\_2.fastq

2: SR\_1.fastq

1: Genome.fasta

✓ Successfully ran workflow "MAPHiTS Not Parallel (paired)", the following datasets have been added to the queue.

1: Genome.fasta  
2: SR\_1.fastq  
3: SR\_2.fastq  
4: [HeaderFastaFilter] Output Fasta File  
5: [MAPHiTS] SAM file  
6: [MAPHiTS] BAM file  
7: [MAPHiTS] PILEUP file  
8: [MAPHiTS] RESUME file  
9: [MAPHiTS] VARSCAN file

When you run the workflow, this message appears !

# IV.3. MAPHiTS: Launch

Galaxy

Analyze Data Workflow Shared Data Help User

Tools Options

- Convert Formats
- Extract Features
- Fetch Sequences
- Fetch Alignments
- Get Genomic Scores
- Operate on Genomic Intervals
- Statistics
- Wavelet Analysis
- Graph/Display Data
- Regional Variation
- Multiple regression
- Multivariate Analysis
- Evolution
- Metagenomic analyses
- FASTA manipulation
- NGS: QC and manipulation
- NGS: Indel Analysis
- NGS: SAM Tools
- FastX Toolkit
- MAPHiTS
- S-MART

Workflows

- Trim And Compare ALL Short Reads (paired)
- MAPHiTS Not Parallel (single)
- MAPHiTS Not Parallel (paired)
- MAPHiTS Parallel (single)
- MAPHiTS Parallel (paired)
- Trim And Compare EPGV Short Reads (paired)
- All workflows

Successfully ran workflow "MAPHiTS Not Parallel (paired)", the following datasets have been added to the queue.

- 1: Genome.fasta
- 2: SR\_1.fastq
- 3: SR\_2.fastq
- 4: [HeaderFastaFilter] Output Fasta File
- 5: [MAPHiTS] SAM file
- 6: [MAPHiTS] BAM file
- 7: [MAPHiTS] PILEUP file
- 8: [MAPHiTS] RESUME file
- 9: [MAPHiTS] VARSCAN file

History Options

Workshop 6

9: [MAPHiTS] VARSCAN file  
8: [MAPHiTS] RESUME file  
7: [MAPHiTS] PILEUP file  
6: [MAPHiTS] BAM file  
5: [MAPHiTS] SAM file  
4: [HeaderFastaFilter] Output Fasta File  
3: SR\_2.fastq  
2: SR\_1.fastq  
1: Genome.fasta

# IV.4. Shared Workflows / Data

## Published Workflows

search  | [Advanced Search](#)

Name	Annotation
<a href="#">MAPHiTS Parallel (paired)</a>	Workflow of SNPs detection, in parallel, for paired-end short reads.
<a href="#">Trim And Compare EPGV Short Reads (paired)</a>	
<a href="#">Trim And Compare ALL Short Reads (paired)</a>	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.

**Data Libraries**

search  | [Advanced Search](#)

Name ↓

grapereseq

magictomsnps

muscares

poplar

# IV.4. Shared Workflows / Data

Data Library “grapereseq”

Name

➤  short reads ▾ All short reads

VVinifera\_v5.1\_chr\_05Jan2010.fasta ▾ Reference Genome

For selected items:  ▾



They can import their data into the history quickly.

→ Useful 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 !

The screenshot shows a web-based application interface titled 'Published Histories'. At the top, there is a search bar with the placeholder 'search' and a magnifying glass icon, followed by a link to 'Advanced Search'. Below the search area is a table with two columns: 'Name' and 'Annotation'. The 'Name' column contains two entries: 'VarScan compare Muscares' and 'VarScan compare Muscares v2'. The 'Annotation' column is empty for both entries.

Name	Annotation
<a href="#">VarScan compare Muscares</a>	
<a href="#">VarScan compare Muscares v2</a>	

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

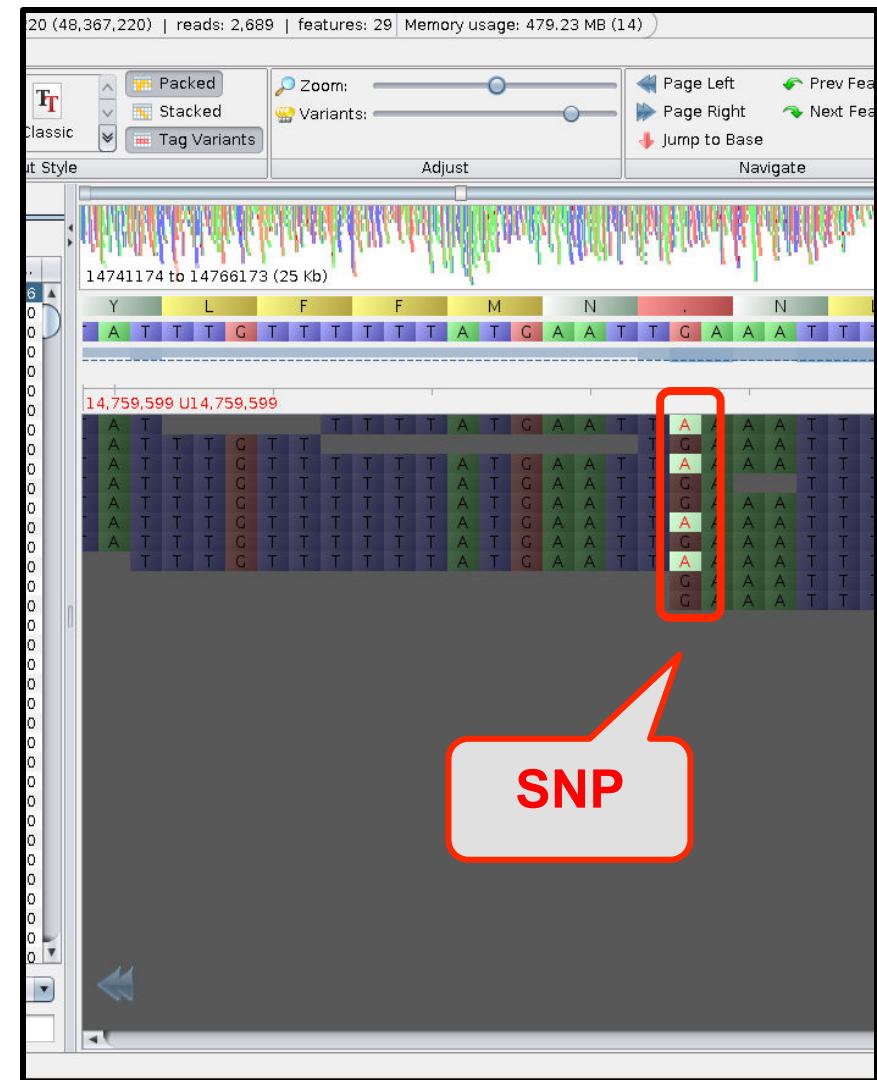
## V. Preliminary Results



# 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  $\leq 1e10^{-3}$



# 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)**

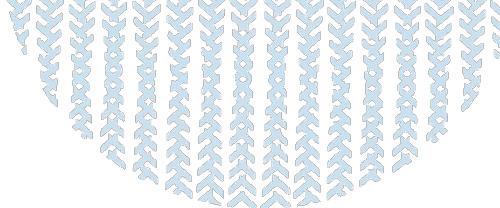
# Acknowledgements

- **Dave Clements**
- **Galaxy developers**
- **Galaxy community**



# Acknowledgements





# Thank you for your attention !!!

