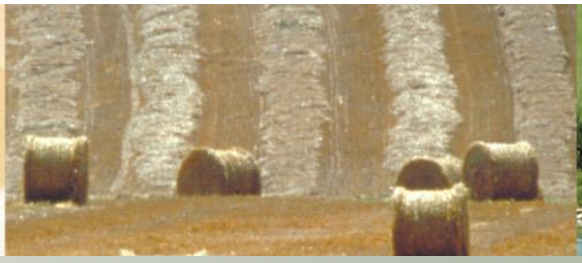
Development of a pipeline for SNPs detection

Colloque EPGV 2010:

Détection, Gestion et Analyse du Polymorphisme Des Génomes Végétaux






9, 10 et 11 Juin

ARNOUX Sandie
BRAS Marc

I. Background and objectives of the pipeline

Setting up a pipeline of SNPs detection to meet the needs of various projects.

ARNOUX Sandie
BRAS Marc

II. Rated software

• **BWA:** software of Mapping, particularly suitable for alignment of Short Reads against one sequence of reference (Burrows - Wheeler Alignment tool).
→ open source !!!

• **SAM tools:** toolkit for working on the output file of BWA.

• **VarScan:** software used to filter SNPs and indels from a BAM file, obtained from SAM tools.
→ predict SNPs / Indels / Seq. Cns.

• **Tablet:** software used to view differents file formats (GFF3, ACE, AFG, MAQ, SOAP, SAM et BAM).

• **GenomeView:** software used to view differents file formats (BAM, GFF, FASTA et annotation file).

II. Software selected
Format : Pileup vs VarScan

Pileup

C10HBa0111D09_LR276	99	T	24	,,,,.,,.,,,,,,,.,,,,,,^],^],	ZZZZTZZZXZZZZZZZZZZVXVUV
C10HBa0111D09_LR276	100	G	24	,$,,,.,,.,,,,,,,.,,,,,,,,	ZZZZZZZZZZZZZZZZZTTZVVVX
C10HBa0111D09_LR276	101	C	24	,,,.,$,.,,,,,,,.,,,,,,,^].	ZZZZZZZZZZZYLXZSQWZTQSQZ
C10HBa0111D09_LR276	102	C	23	,,,.,.,,,,,,.,,,,,,,.	ZZZZZZZYZZRZWZWJJROQUKZ
C10HBa0111D09_LR276	103	T	24	,$,,.,,,,,,,.,,,,,,,.^],	ZZZQZZZZZZZZZZZZZVVVVZU
C10HBa0111D09_LR276	104	T	27	,,.,,,,,,,.,,,,,,,.,^],^],^],^],	ZZTZZZZZZZZZZZZZZZVTZXVUVV
C10HBa0111D09_LR276	105	A	28	,$,$.,.,,,,,,,.,,,,,,,.,,,,,^F,	ZZKZTZZZZZZZZZZZZZZZXVUNXU
C10HBa0111D09_LR276	106	A	28	T,.,,,,,,,.,,,,,,,.,,,,,,^],^],	AZZZZZZZZZZZZZZZZZZXVZXXVTT

Reference Seq. Name

Position In Ref.

Base In this Position

Number Of Short Reads In this Position

Informations For All Short Reads

Base Quality

ARNOUX Sandie
BRAS Marc

II. Software selected
Format : Pileup vs VarScan

VarScan

Chrom	Position	Ref	Var	Reads1	Reads2	VarFreq	Strands1	Strands2	Qual1	Qual2	Pvalue
scaffold_1	7936	C	T	4	3	42,86%	2	1	28	35	0.1923076923076922
scaffold_1	14721	T	G	1	5	83,33%	1	2	39	30	0.015151515151515102
scaffold_1	28988	G	A	6	2	25%	2	2	30	38	0.4666666666666698
scaffold_1	59708	A	G	13	2	13,33%	2	1	29	22	0.48275862068965925
scaffold_1	60144	T	C	22	2	8,33%	2	1	35	34	0.48936170212766006
scaffold_1	60150	A	C	20	4	16,67%	2	2	30	31	0.10921985815602651
scaffold_1	60187	A	C	13	2	13,33%	2	1	34	36	0.48275862068965925
scaffold_1	60210	T	C	12	2	14,29%	2	1	34	26	0.48148148148148034

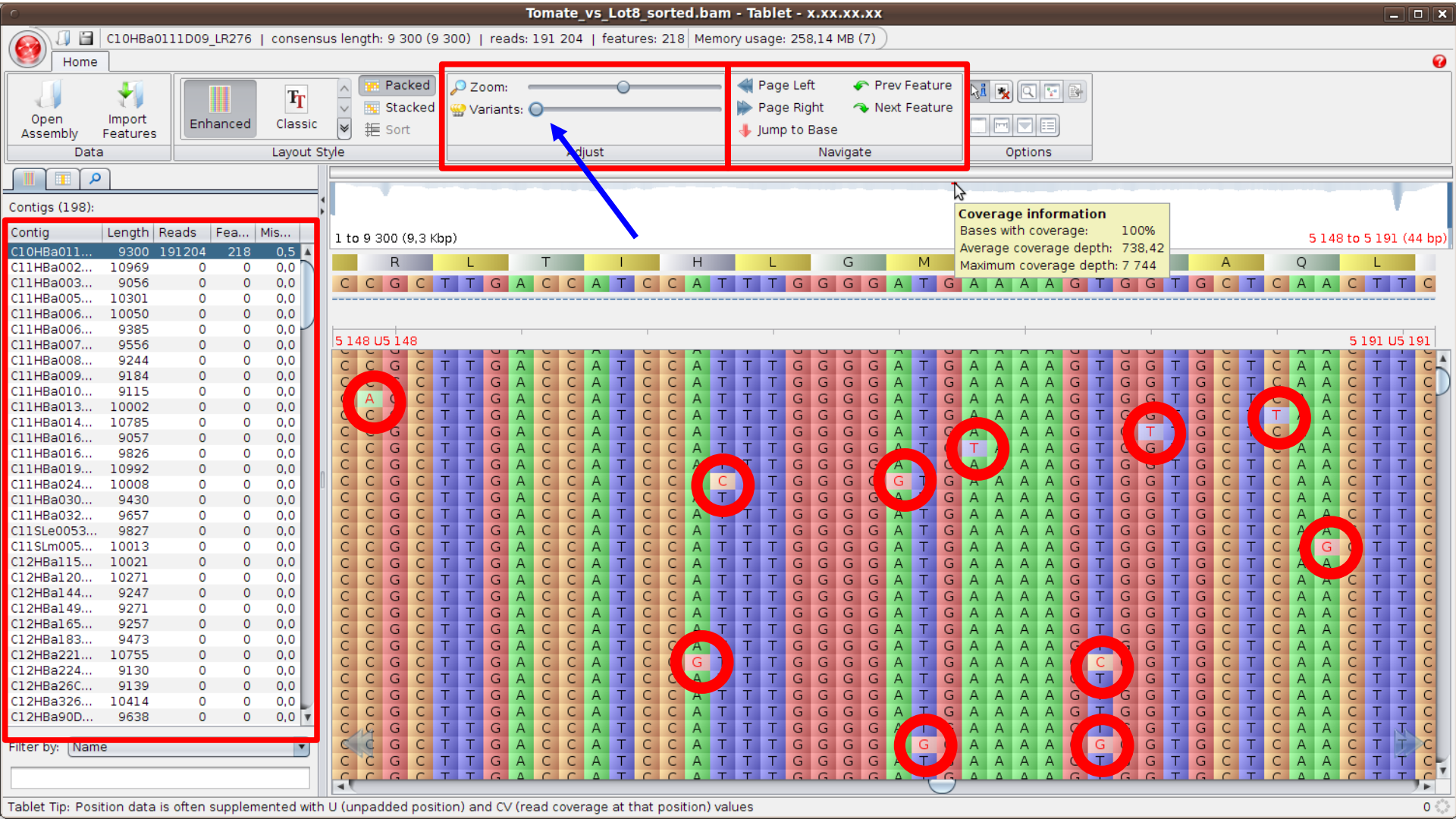
Min coverage: 8
Min reads2: 2
Min var freq: 0.01
Min avg qual: 15
P-value thresh: 0.99
Reading input from STDIN
332191236 bases in pileup file
7698459 met minimum coverage of 8x
86479 SNPs predicted

Different Filters

→ Choose your values !

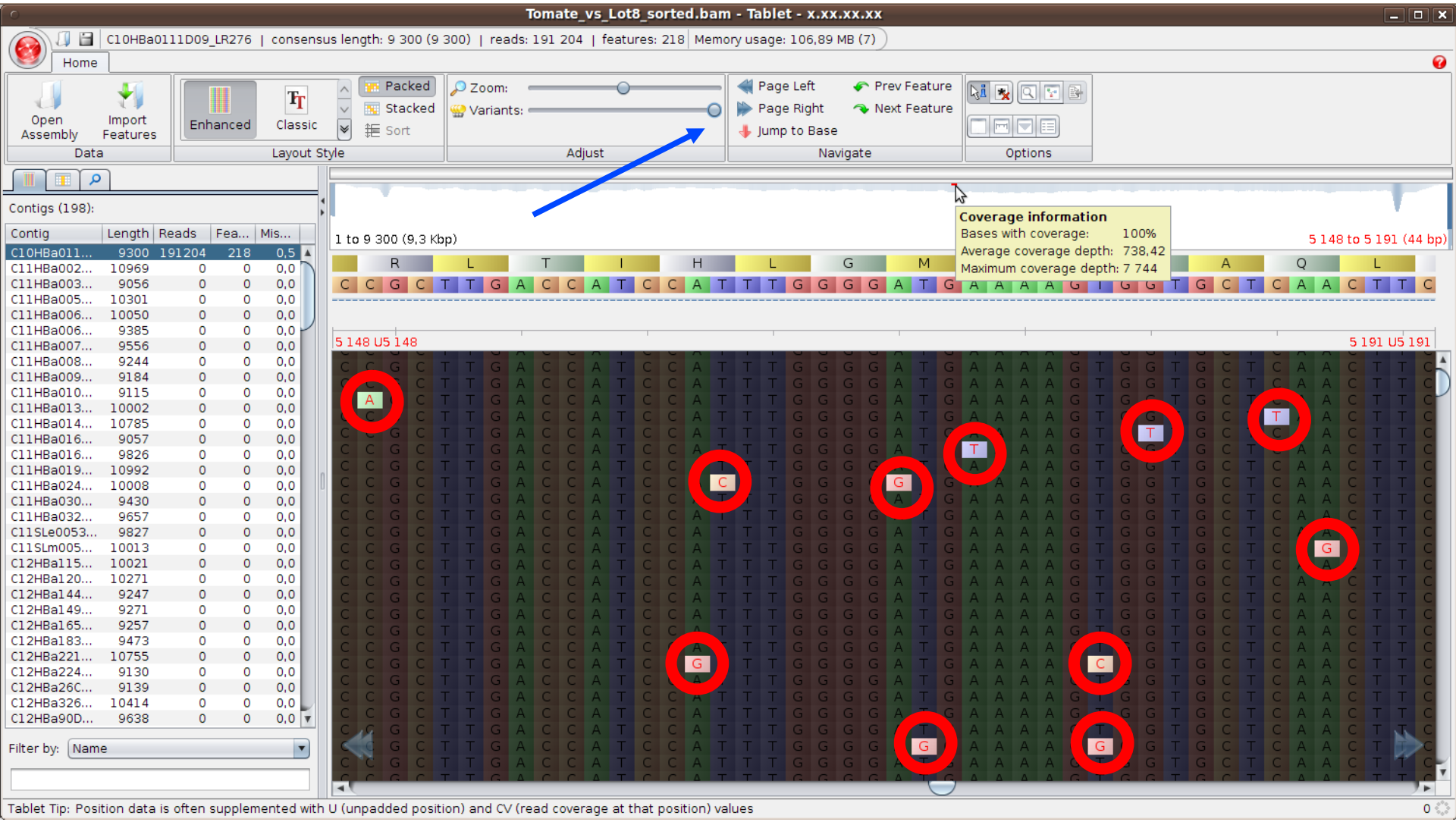
ARNOUX Sandie
BRAS Marc

II. Software selected
Tablet: SNPs view



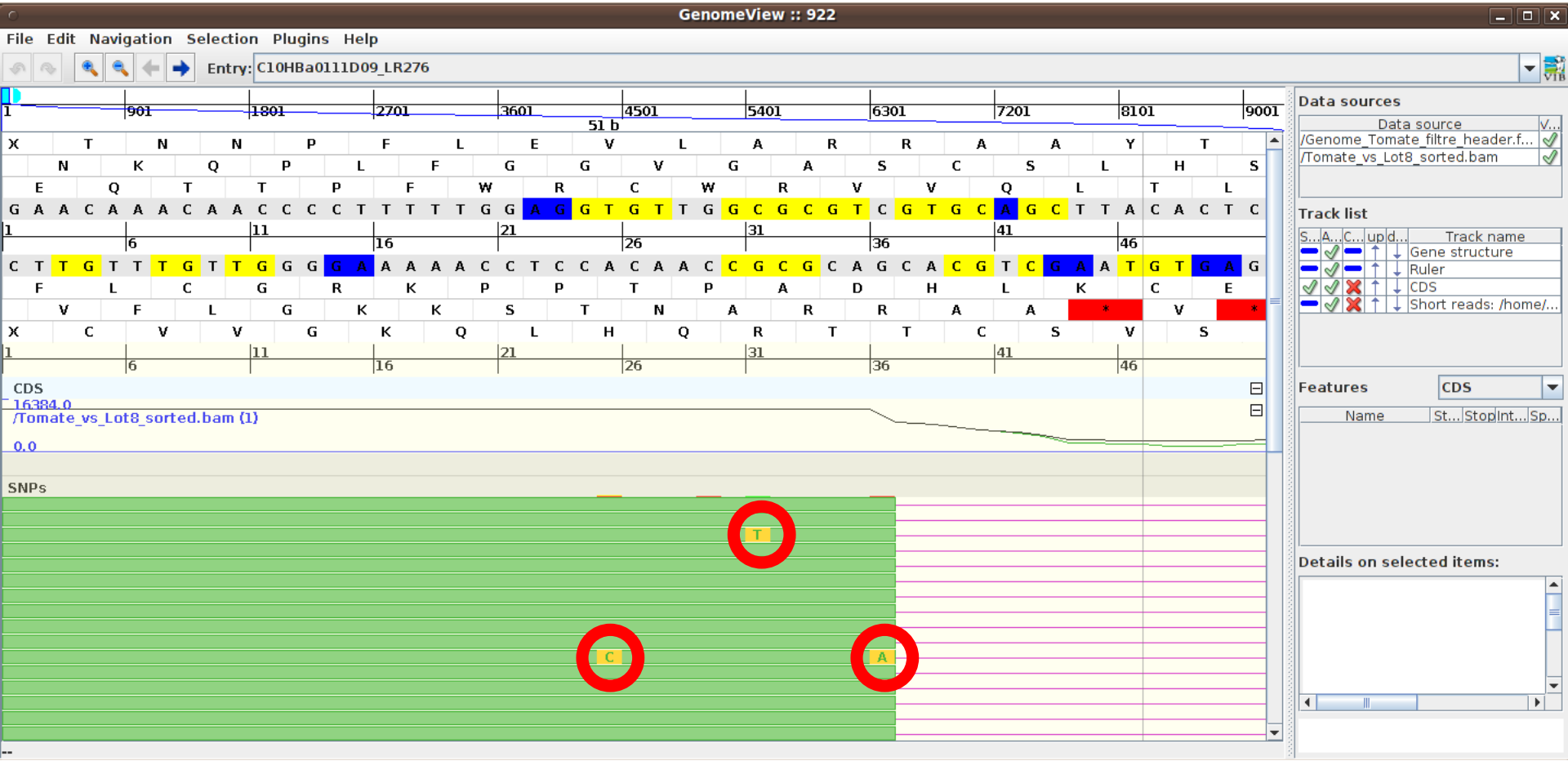
ARNOUX Sandie
BRAS Marc

II. Software selected
Tablet: SNPs view



ARNOUX Sandie
BRAS Marc

II. Software selected
GenomeView: SNPs view



ARNOUX Sandie
BRAS Marc

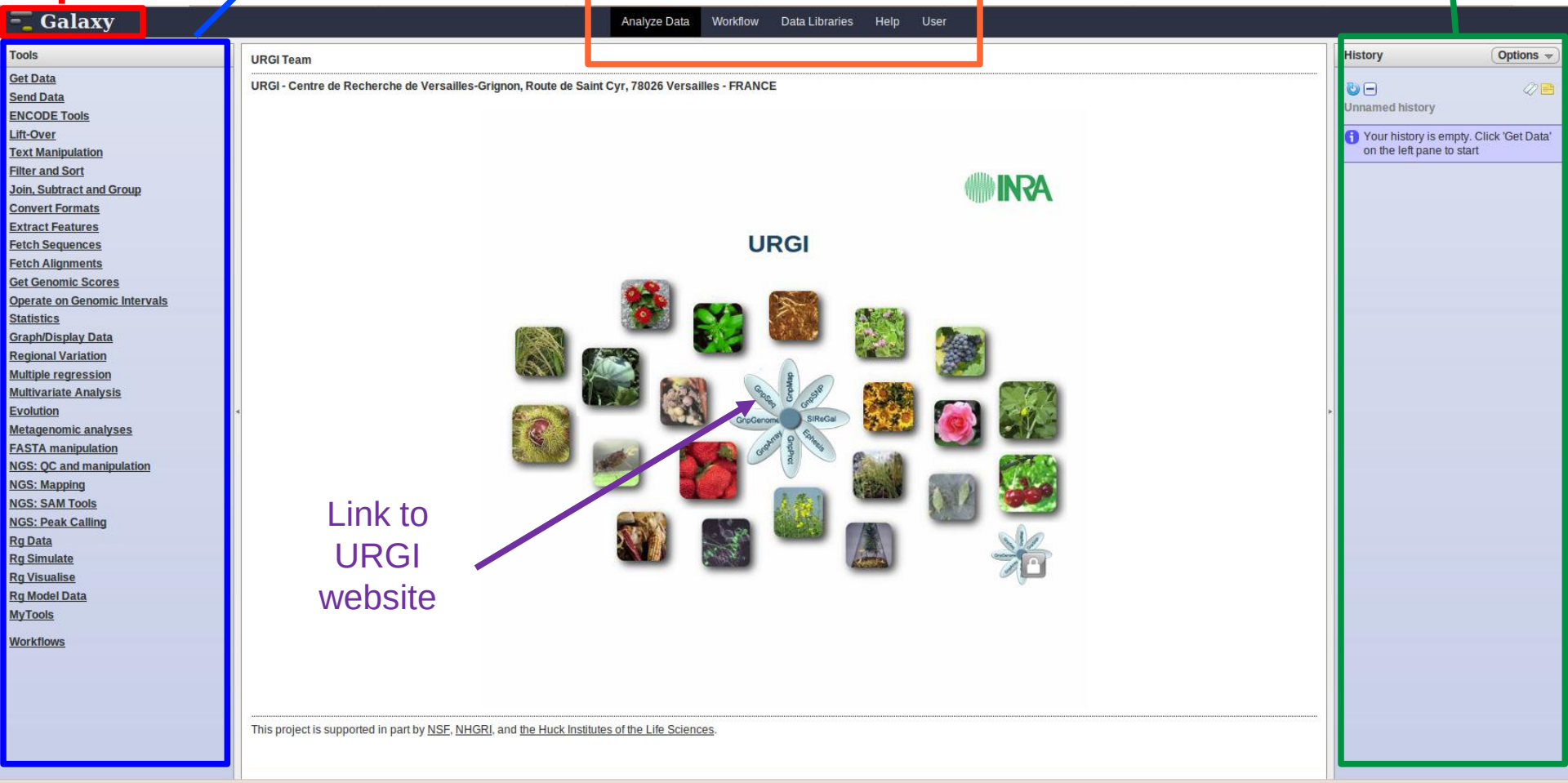
III. Integration of tools in Galaxy

Return to homepage

Tools

Managment of Galaxy

History

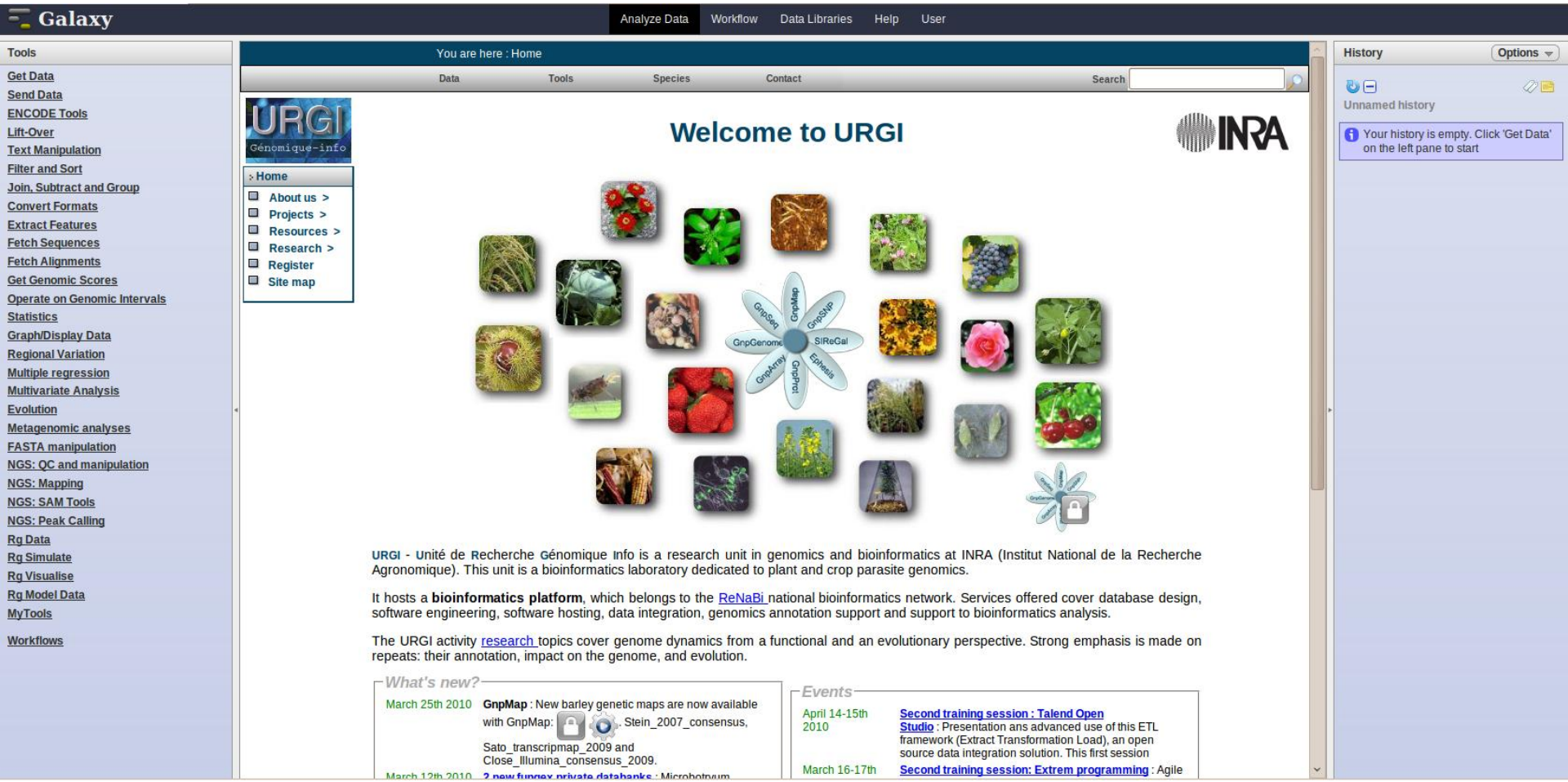


Link to URGI website

ARNOUX Sandie
BRAS Marc

URGI public site

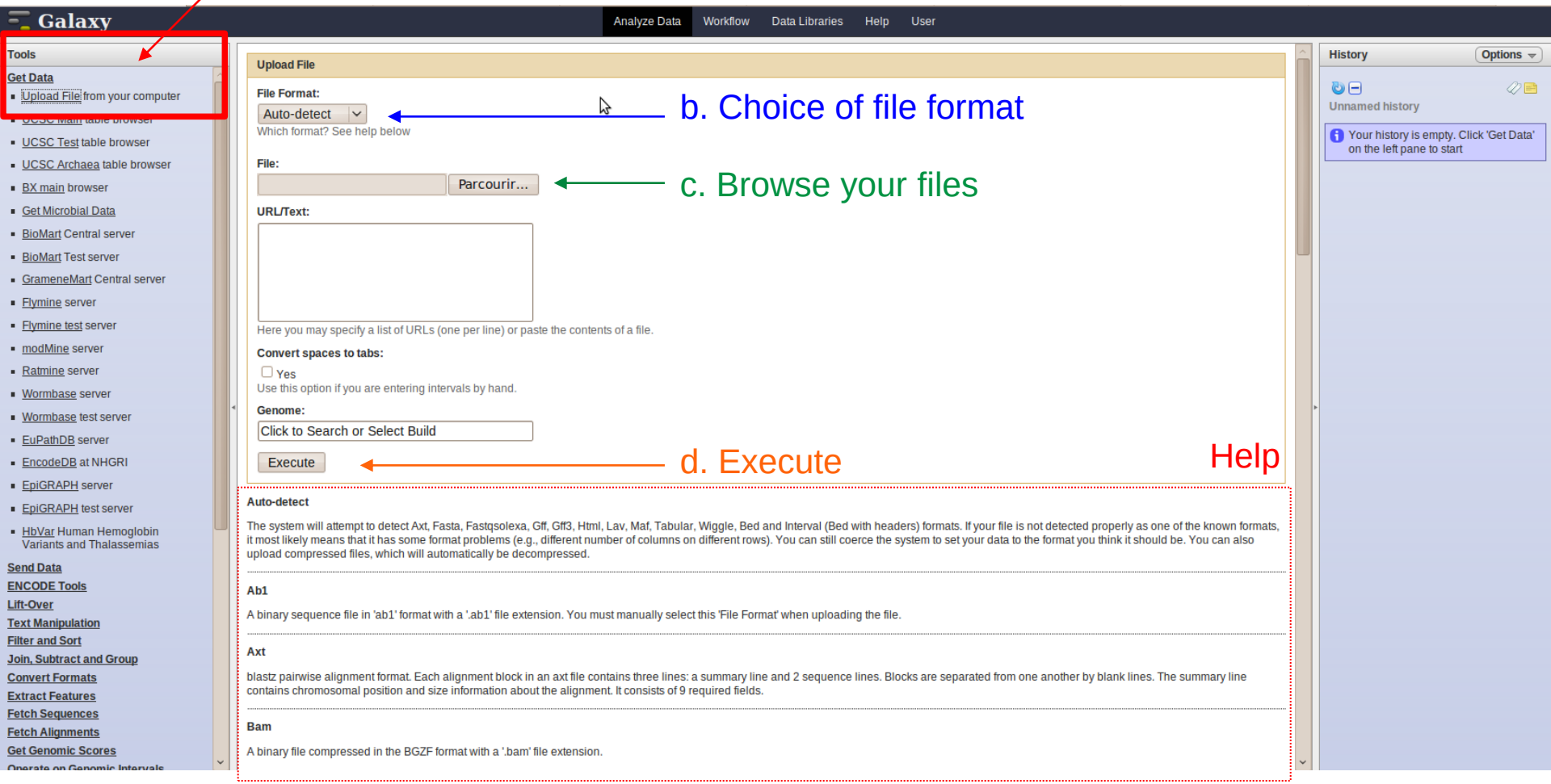
http://urgi.versailles.inra.fr/



ARNOUX Sandie
BRAS Marc

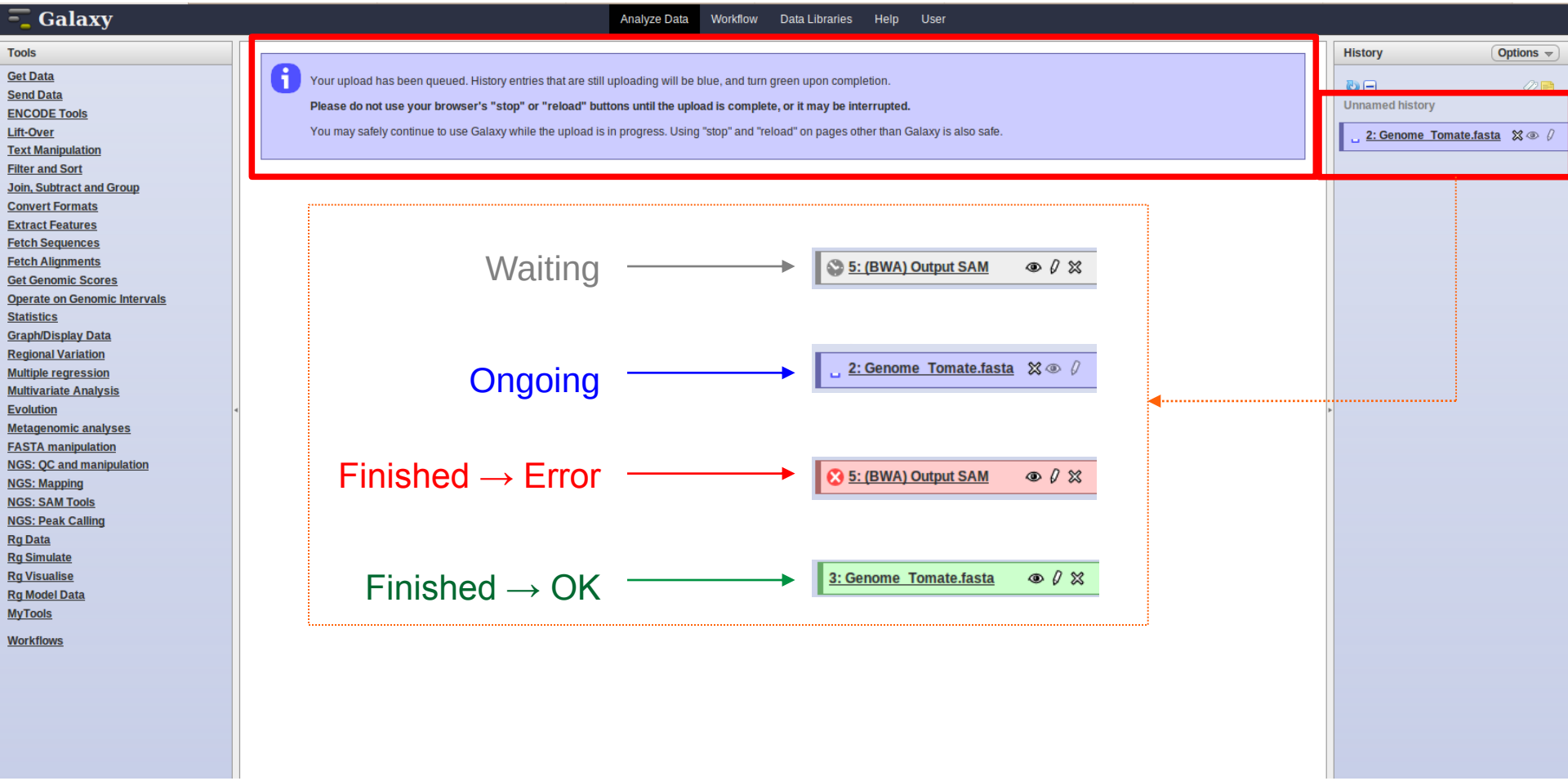
III.1. How to provide your data to Galaxy ?

a. Tools → Get Data → Upload File



ARNOUX Sandie
BRAS Marc

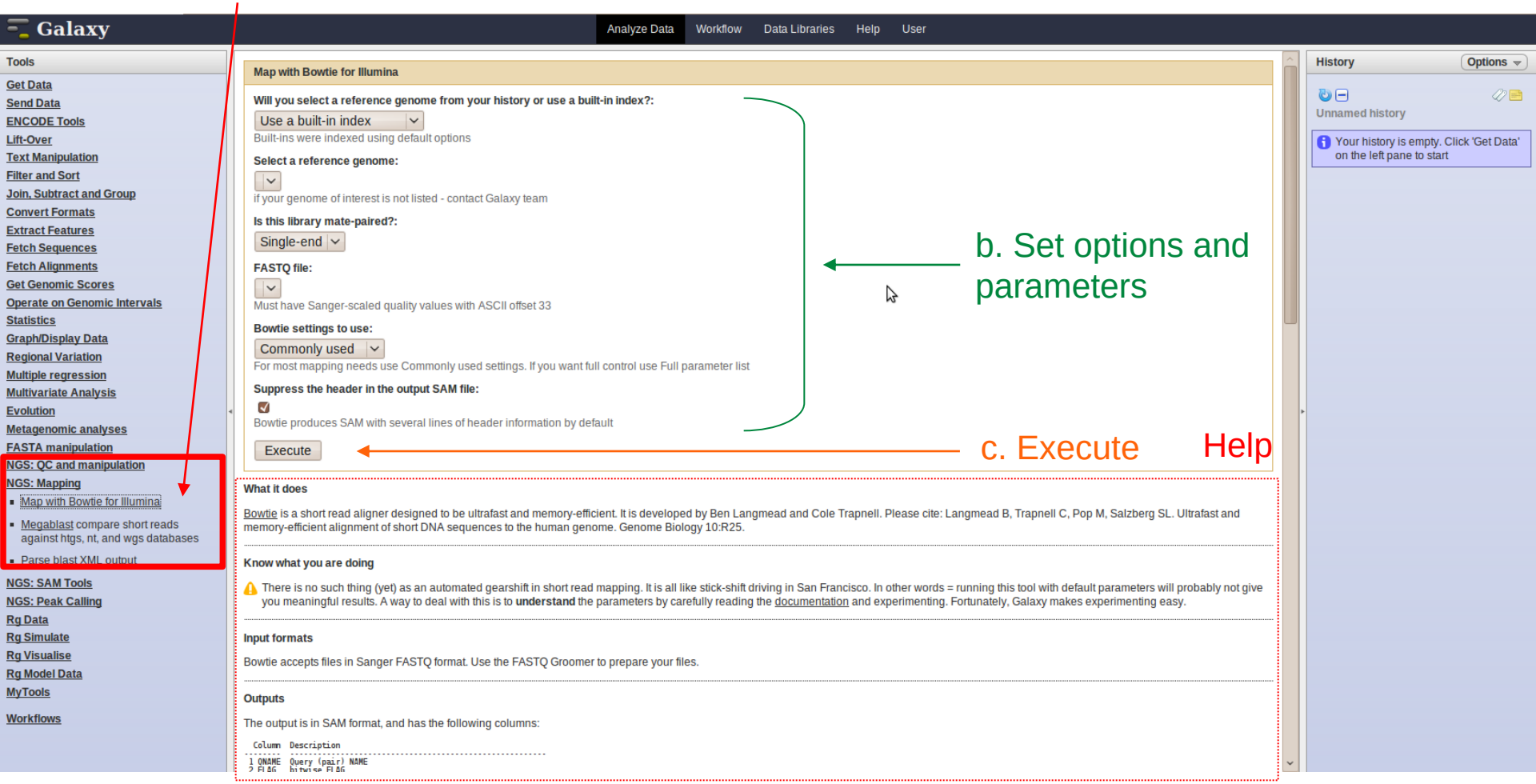
III.1. How to provide your data to Galaxy ?



ARNOUX Sandie
BRAS Marc

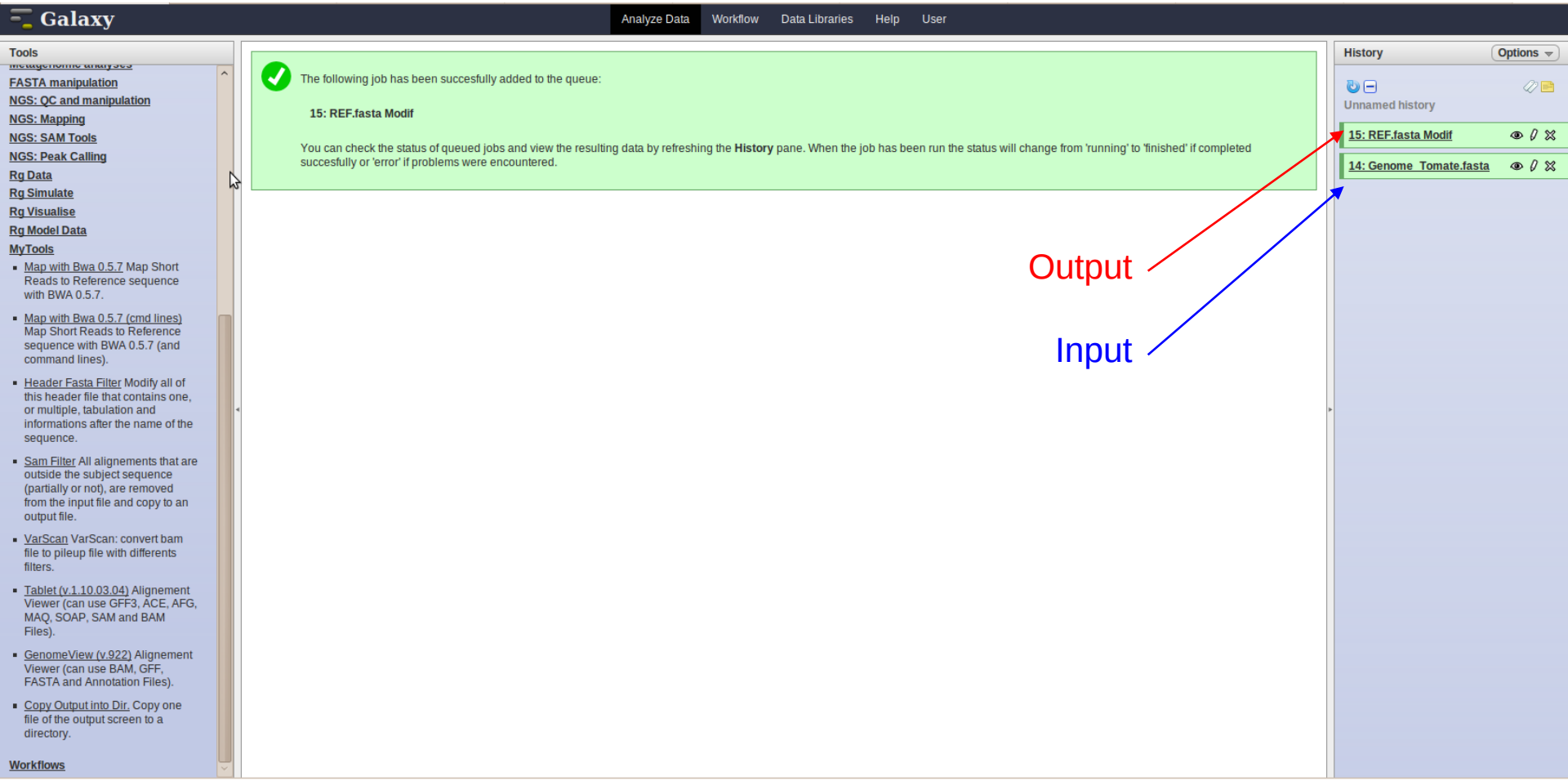
III.2. How to use a tool on Galaxy ?

a. Choose a tool from the list



ARNOUX Sandie
BRAS Marc

III.2. How to use a tool on Galaxy ?



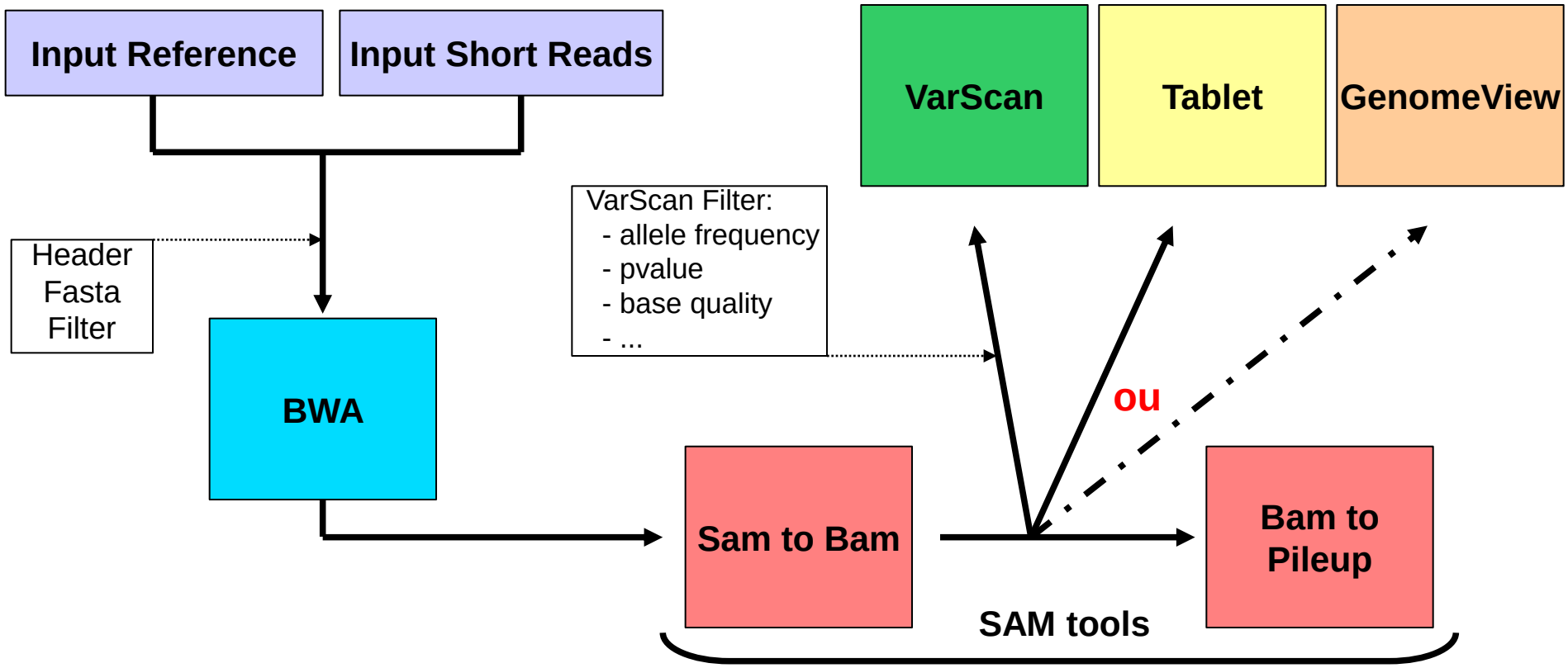
ARNOUX Sandie
BRAS Marc

III.3. How to integrate a tool in Galaxy ?

We can easily integrate a new tool !!!

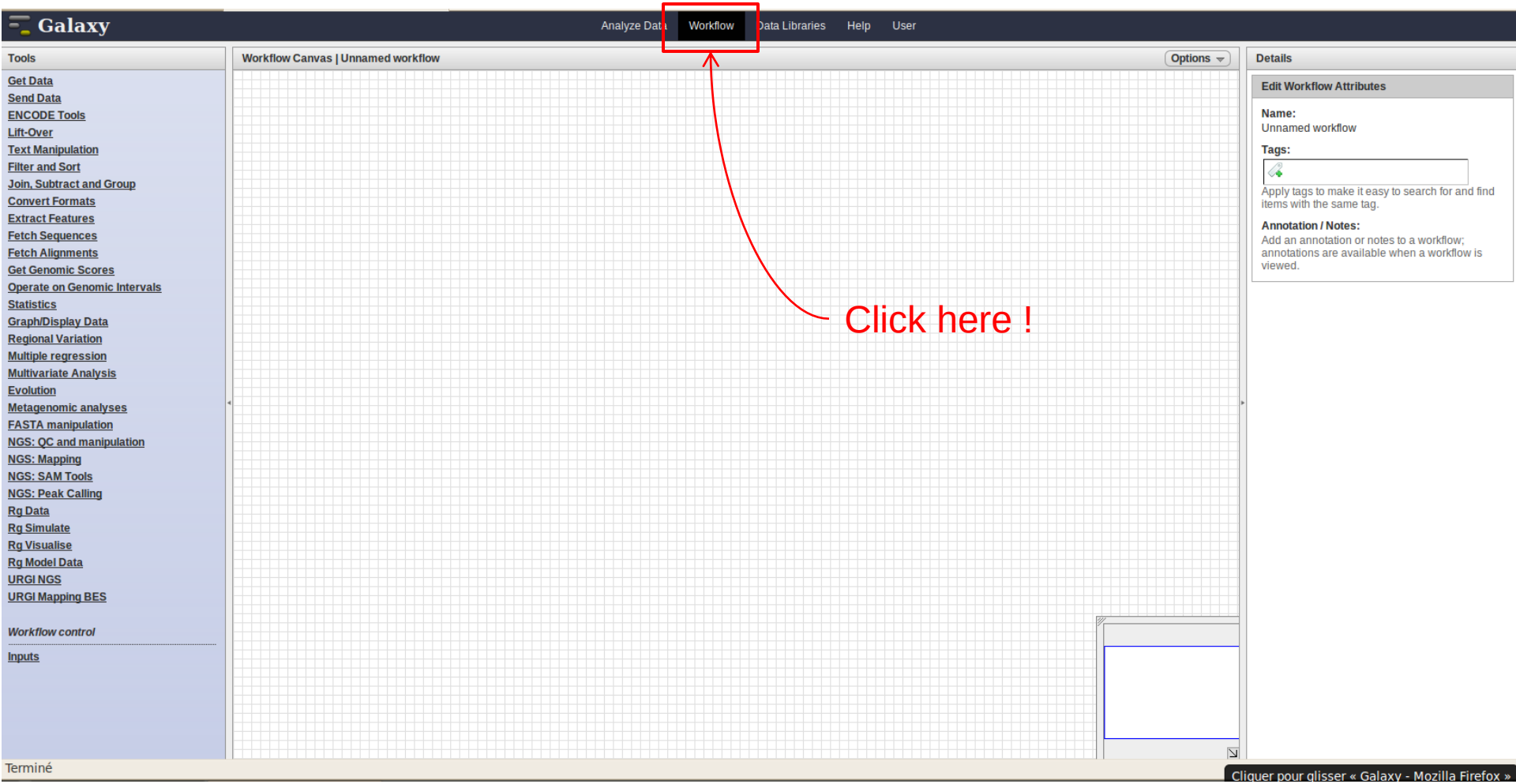
ARNOUX Sandie
BRAS Marc

III.4. Our worklow for SNPs detection.



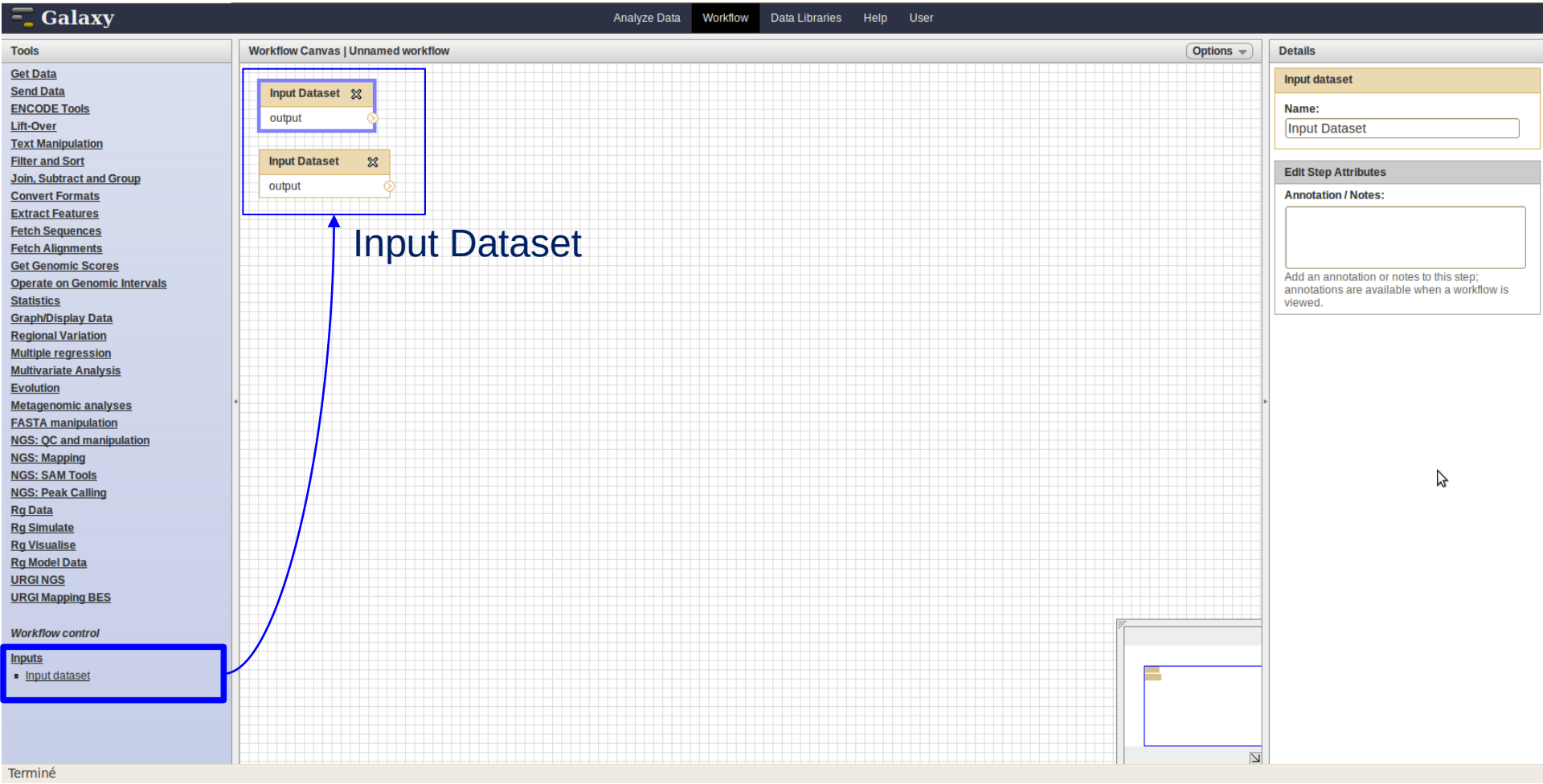
ARNOUX Sandie
BRAS Marc

III.5. How to do this workflow with Galaxy ?



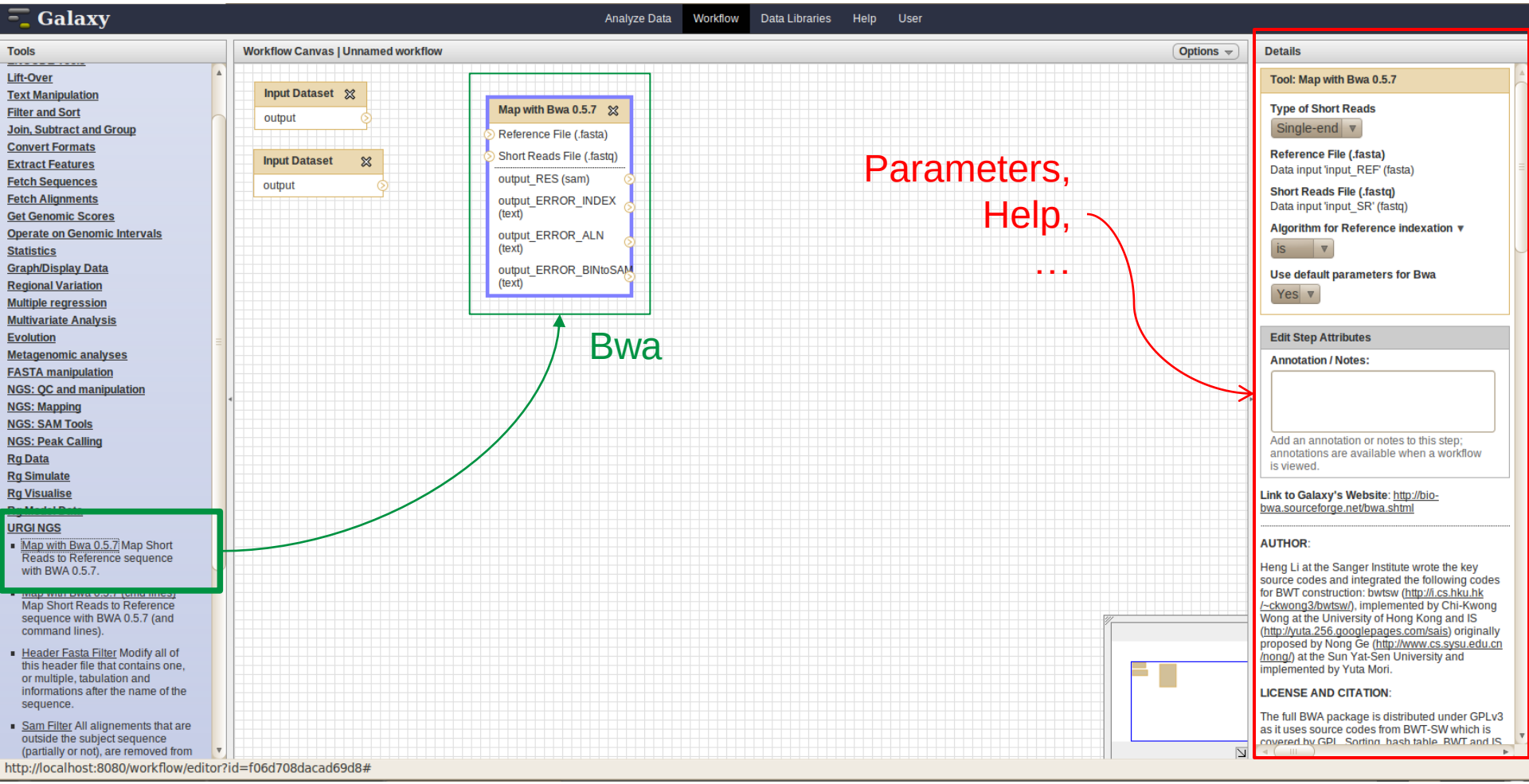
ARNOUX Sandie
BRAS Marc

III.5. How to do this workflow with Galaxy ?



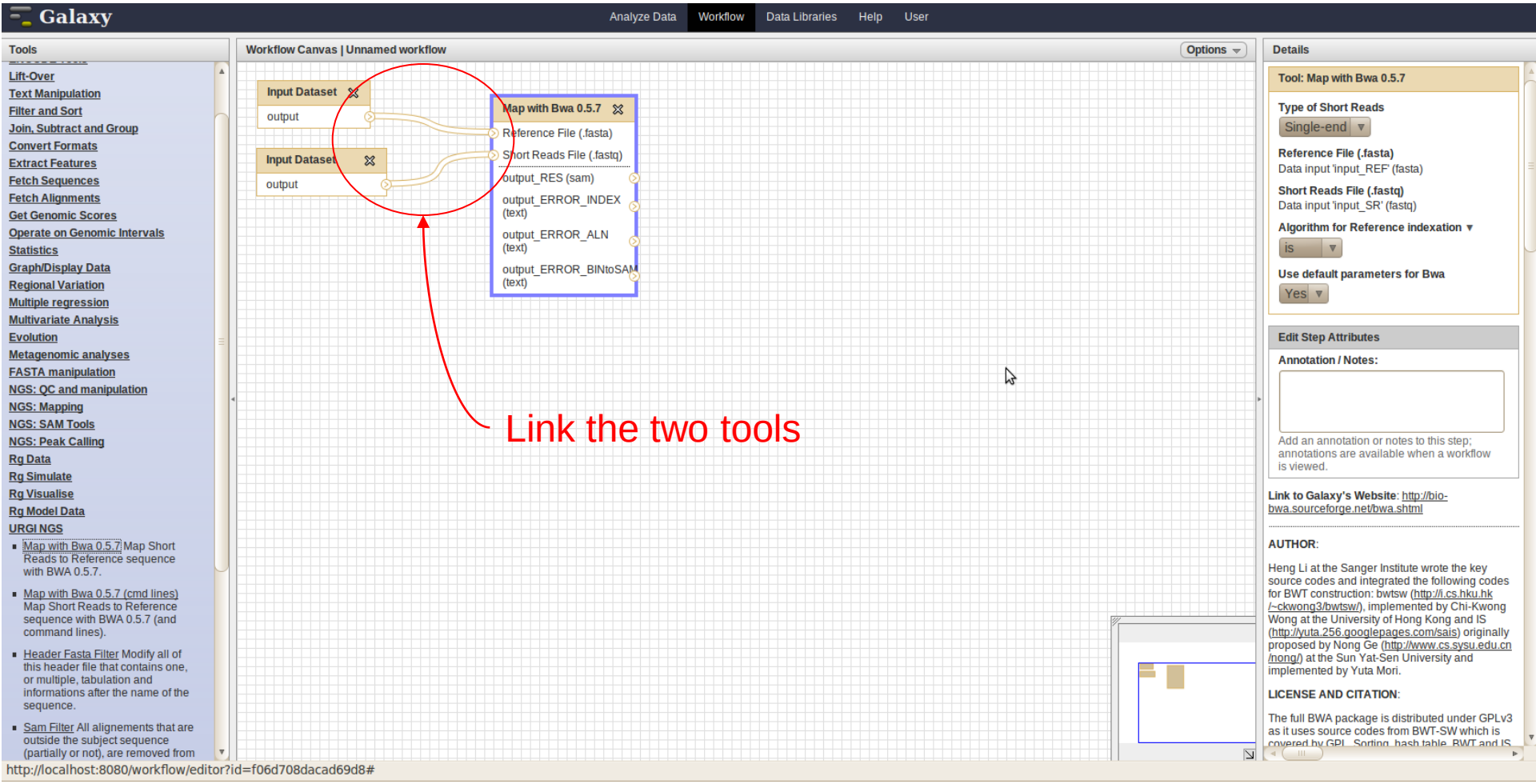
ARNOUX Sandie
BRAS Marc

III.5. How to do this workflow with Galaxy ?



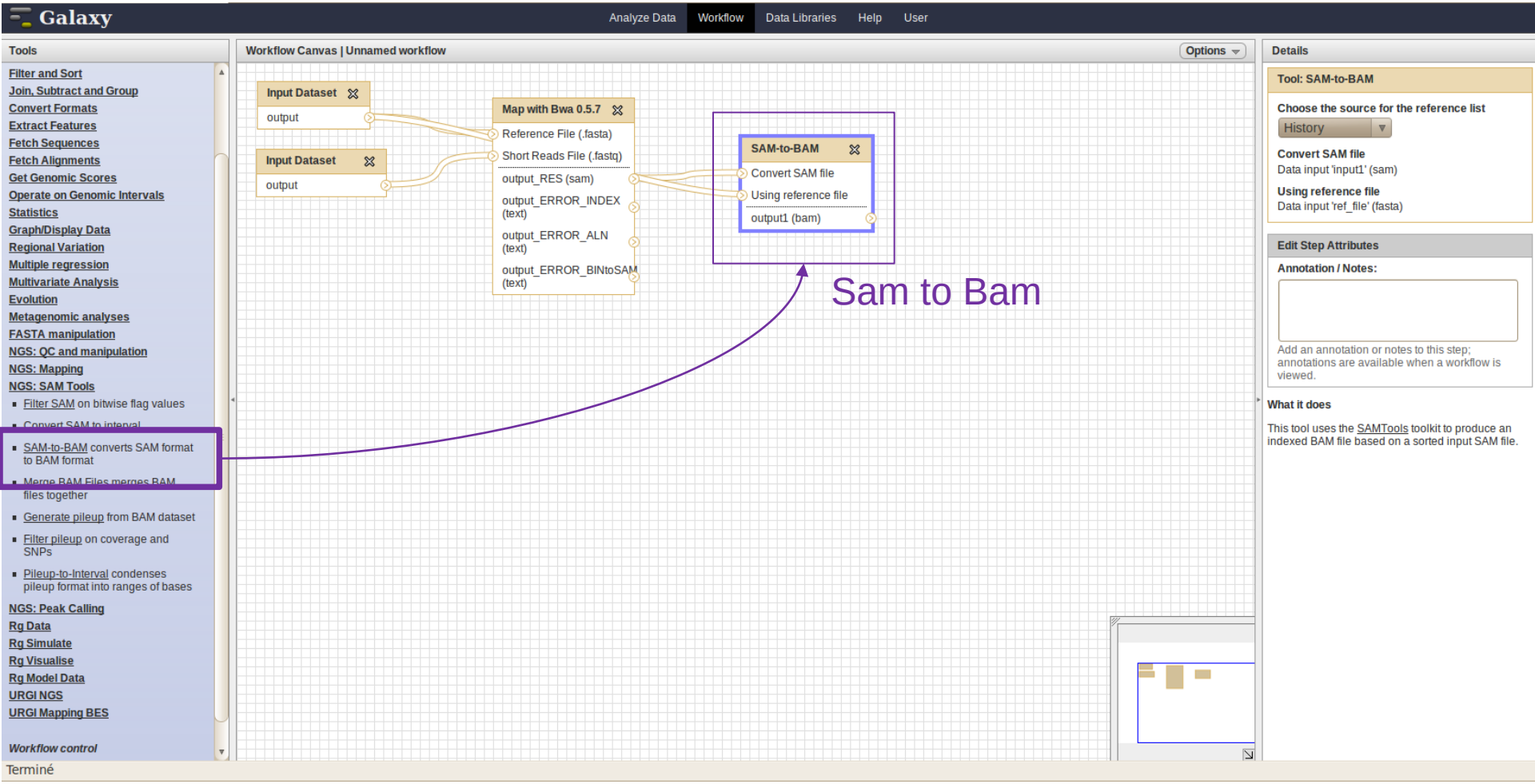
ARNOUX Sandie
BRAS Marc

III.5. How to do this workflow with Galaxy ?



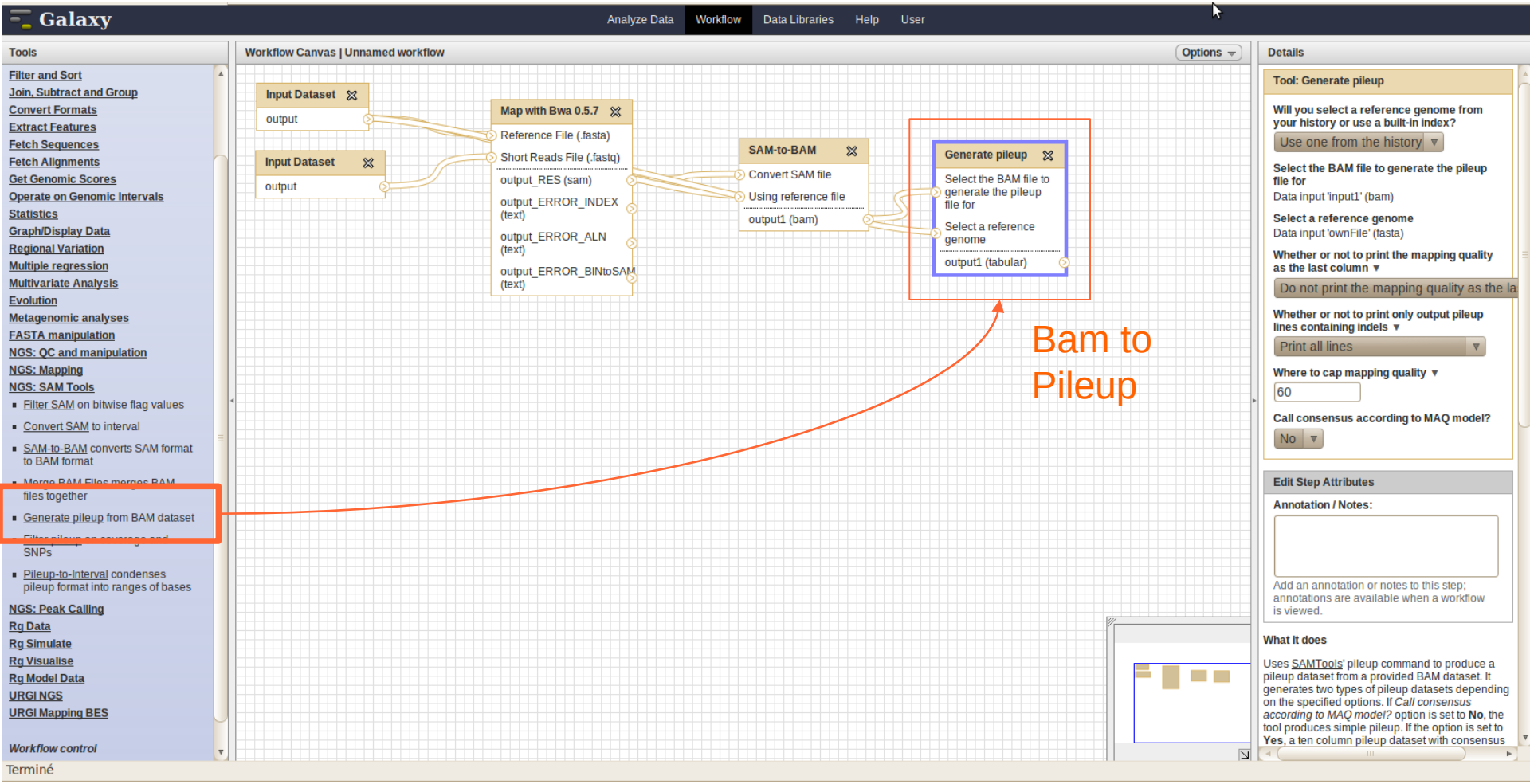
ARNOUX Sandie
BRAS Marc

III.5. How to do this workflow with Galaxy ?



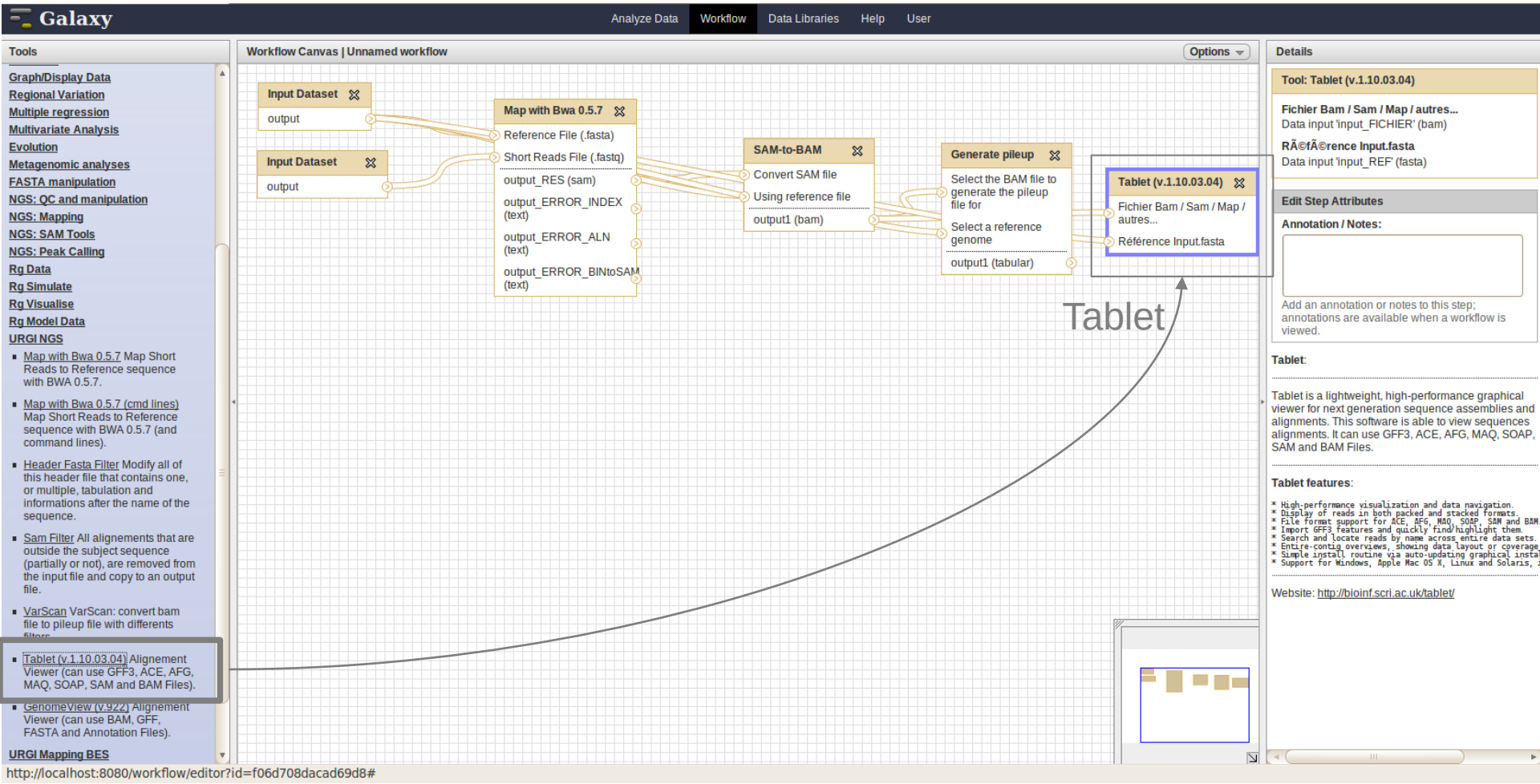
ARNOUX Sandie
BRAS Marc

III.5. How to do this workflow with Galaxy ?



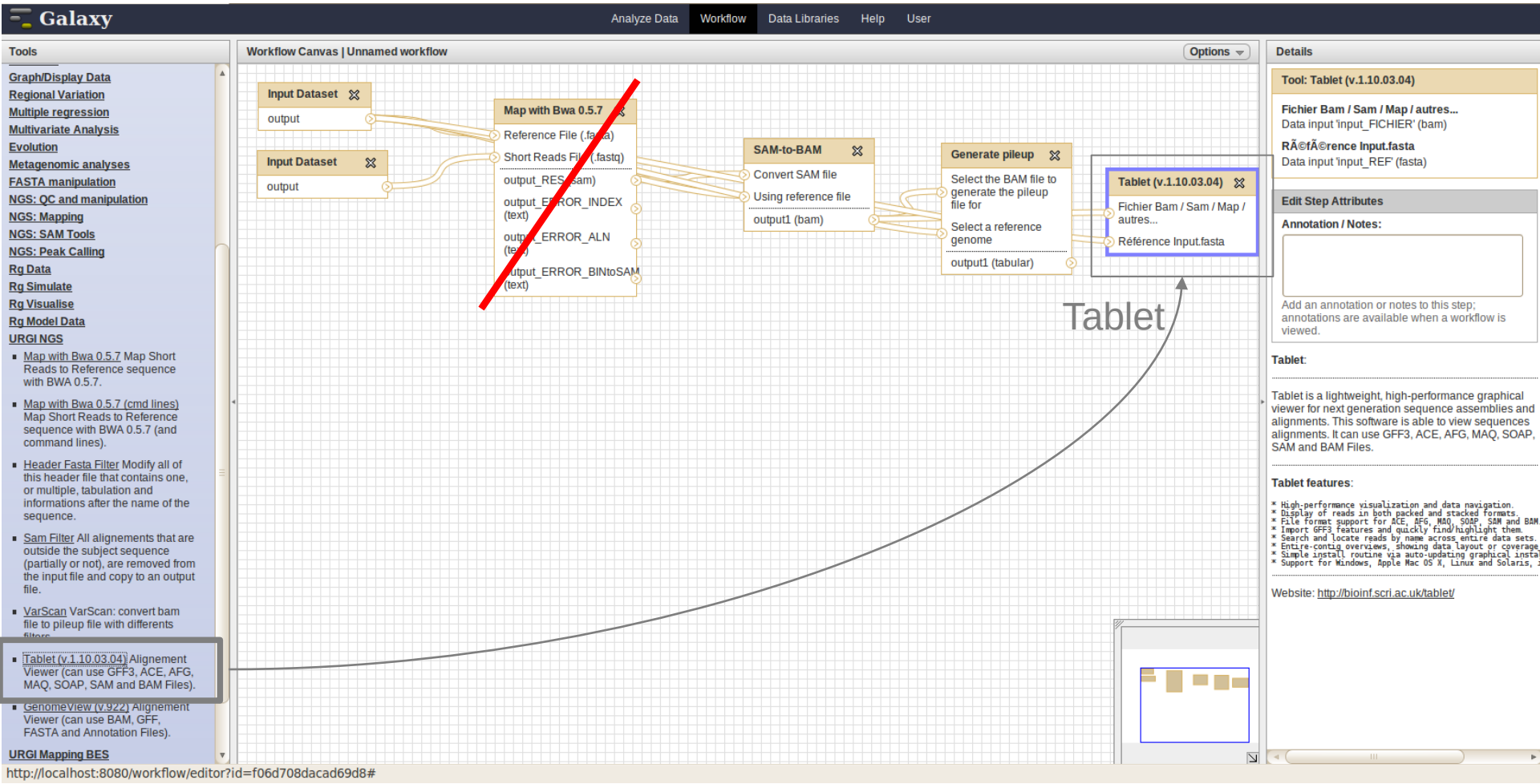
ARNOUX Sandie
BRAS Marc

III.5. How to do this workflow with Galaxy ?



ARNOUX Sandie
BRAS Marc

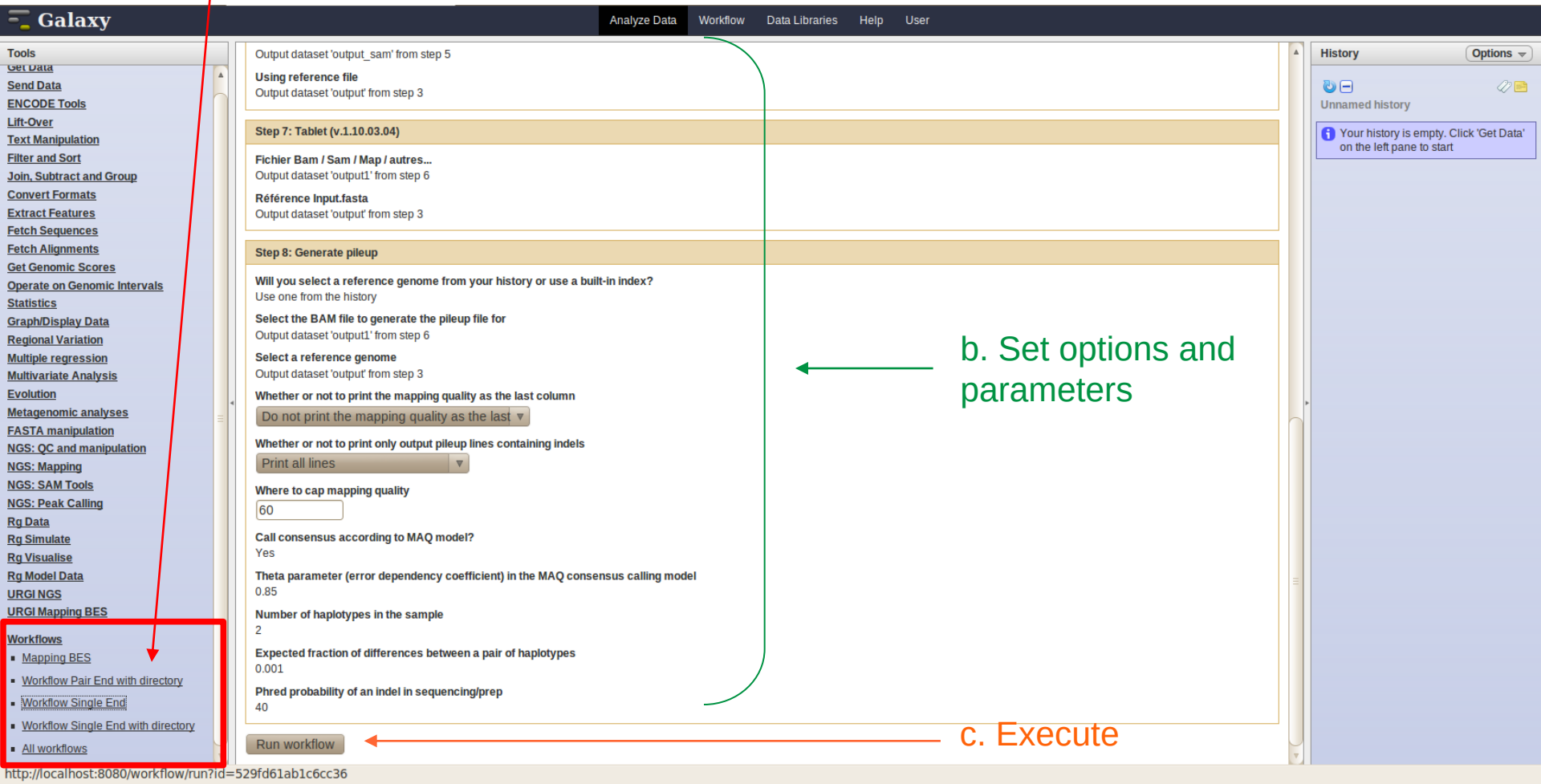
III.5. How to do this workflow with Galaxy ?



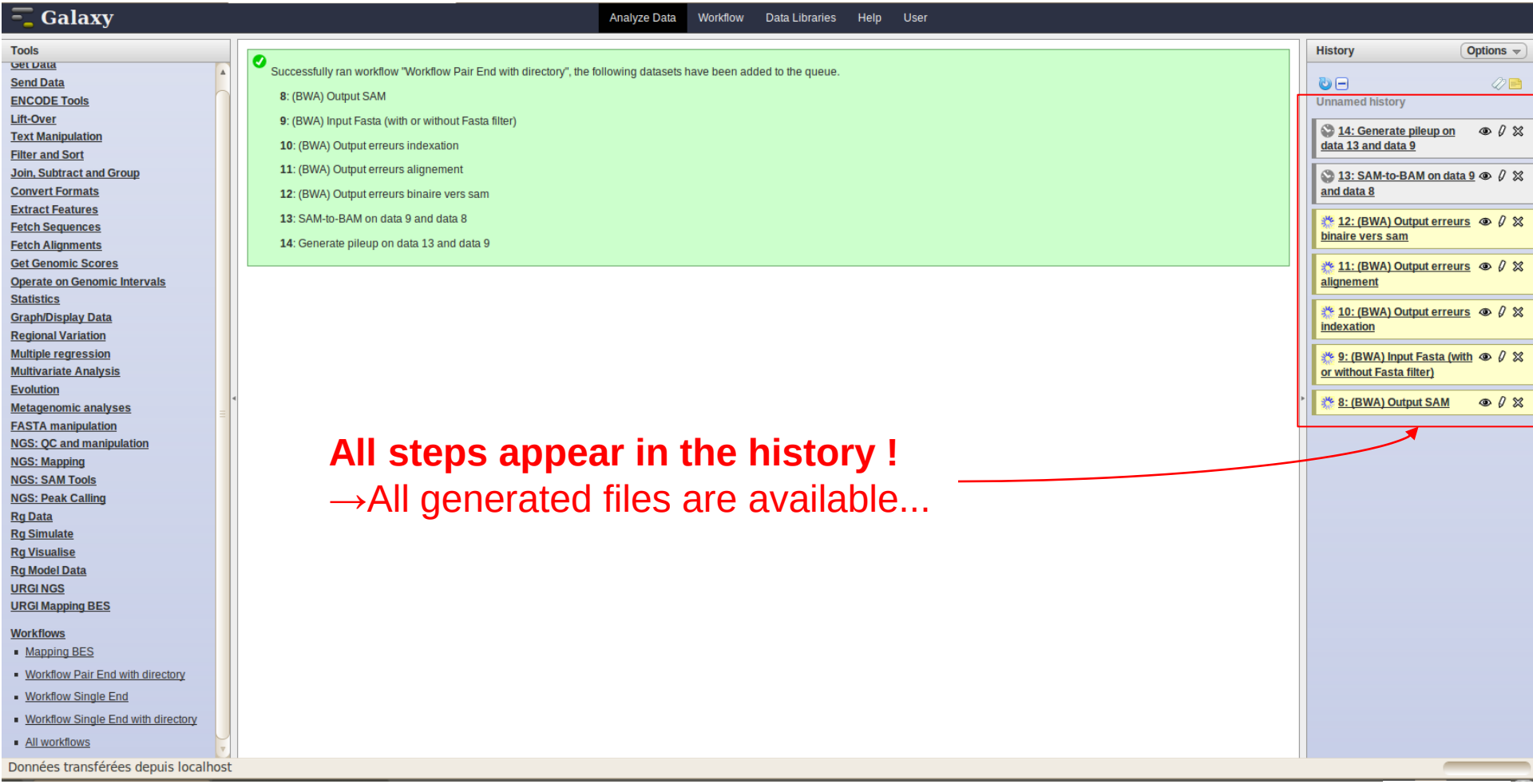
ARNOUX Sandie
BRAS Marc

III.6. How to launch the workflow with Galaxy ?

a. Choose the workflow from the list



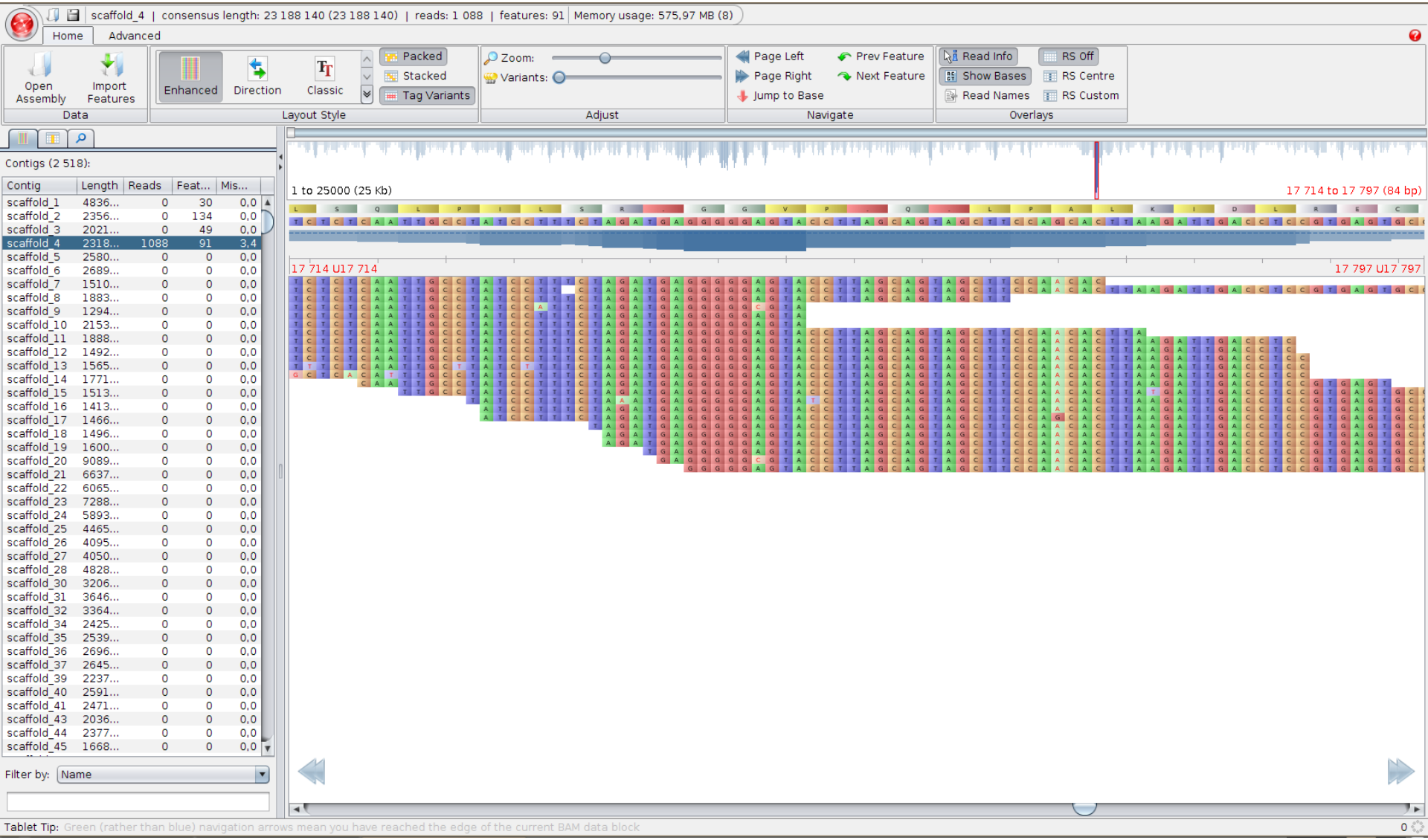
III.6. How to launch the workflow with Galaxy ?



All steps appear in the history !

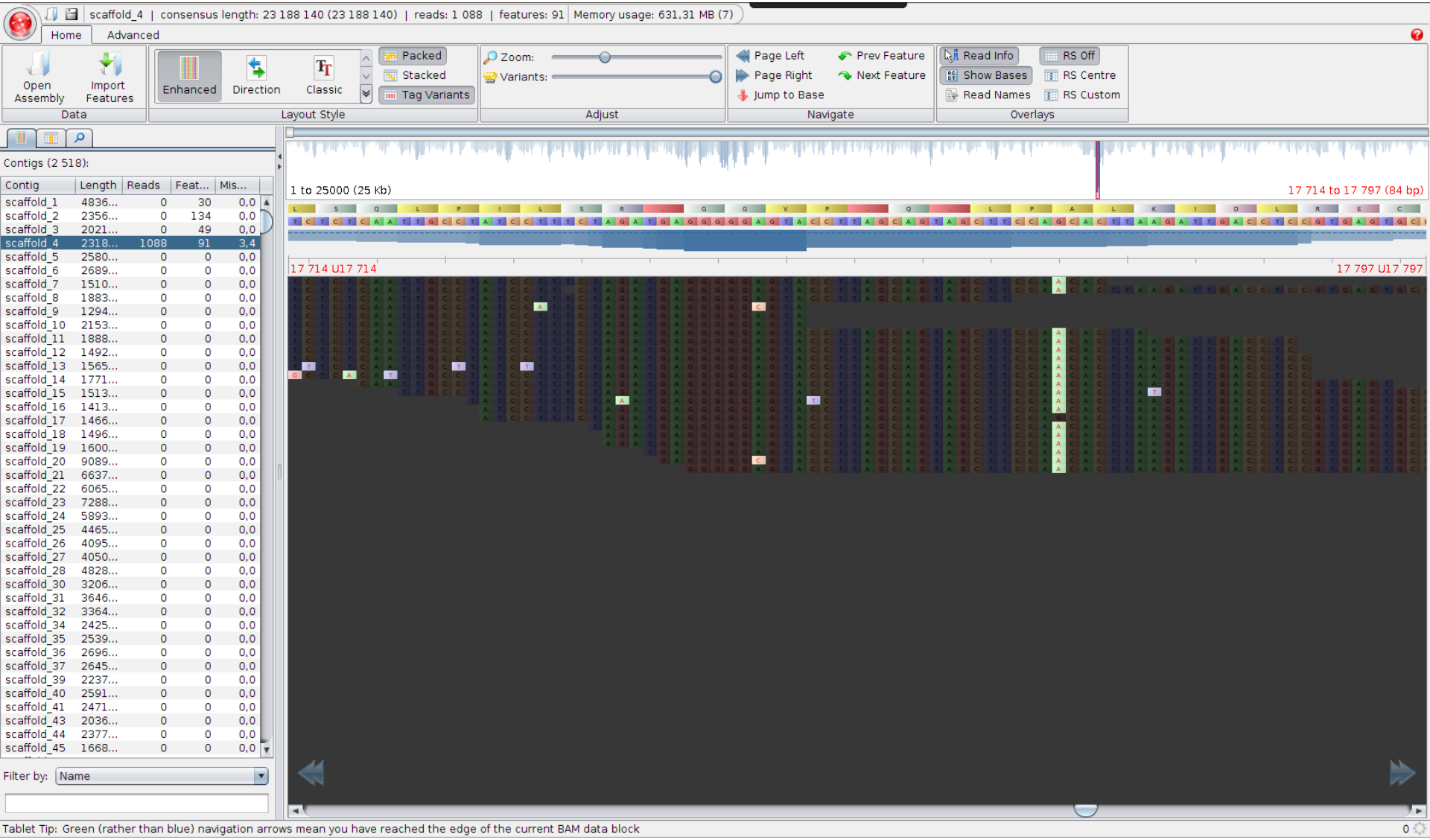
→All generated files are available...

III.7. Results in Tablet



ARNOUX Sandie
BRAS Marc

III.7. Results in Tablet



ARNOUX Sandie
BRAS Marc

IV. Validation Process

1. Tests on public data (NCBI → *A. thaliana* and *V. vinifera*).

2. Tests on data from *Tomato* (pair ends, 36 bps) and *Poplar* (pair ends, 70 bps) with comparison of EPGV and GAFL results (project PlanteReseq INRA).

3. Validation / modification of pipeline steps with the help of users.

4. Implementation and final validation of the pipeline with the ANR projects (GrapeReseq and Muscares).

ARNOUX Sandie
BRAS Marc

V. Outlook

- Try other tools for mapping / SNPs calling (according to validation).

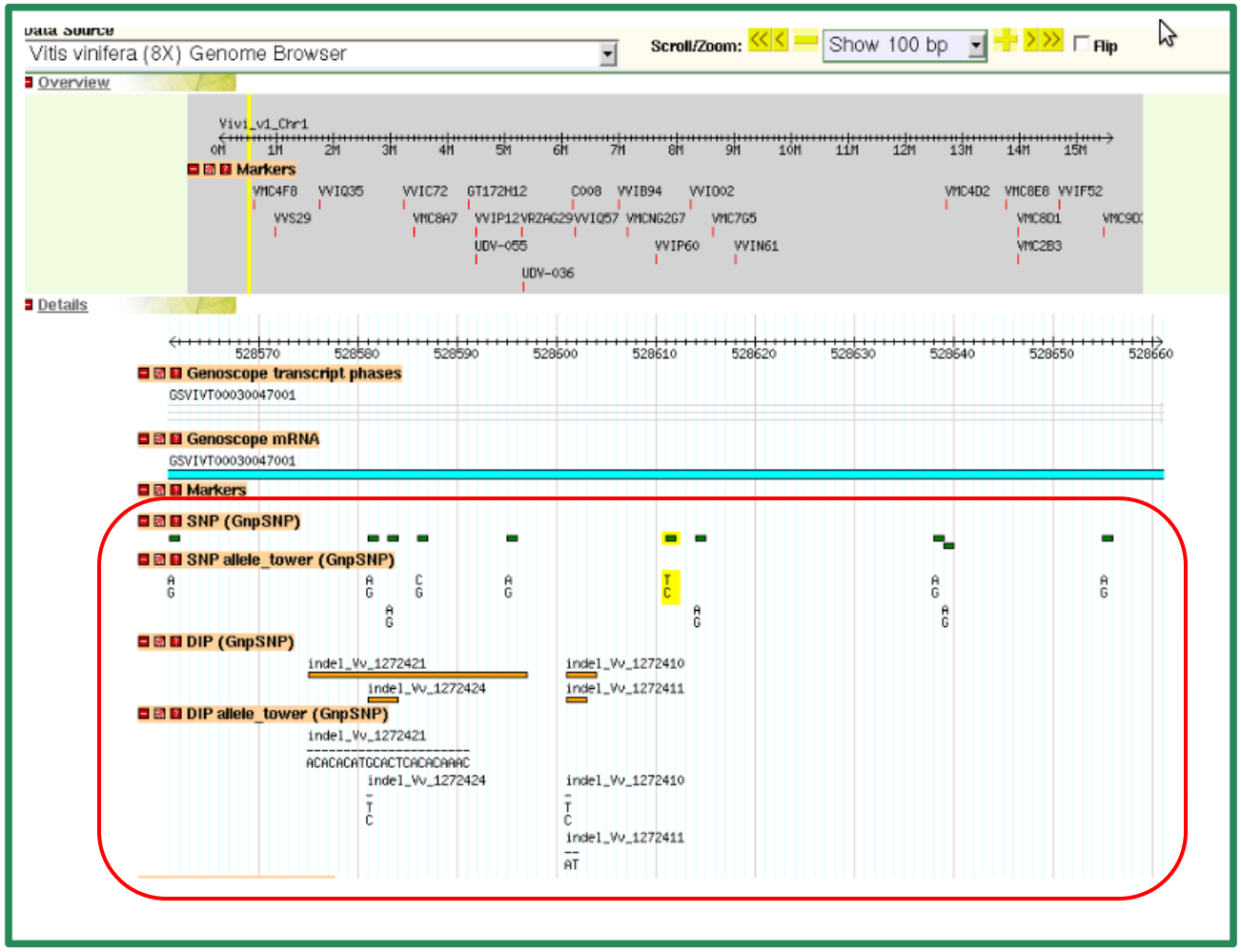
- Insert data into our databases:
  → SNPs and Indels in GnpSNP - GnpGenome
  → Contigs and Clusters in GnpSeq

- Put the pipeline on our website !
  → BWA Parallelization on our clusters.

- Visualization of results in Gbrowse.

ARNOUX Sandie
BRAS Marc

V. Outlook

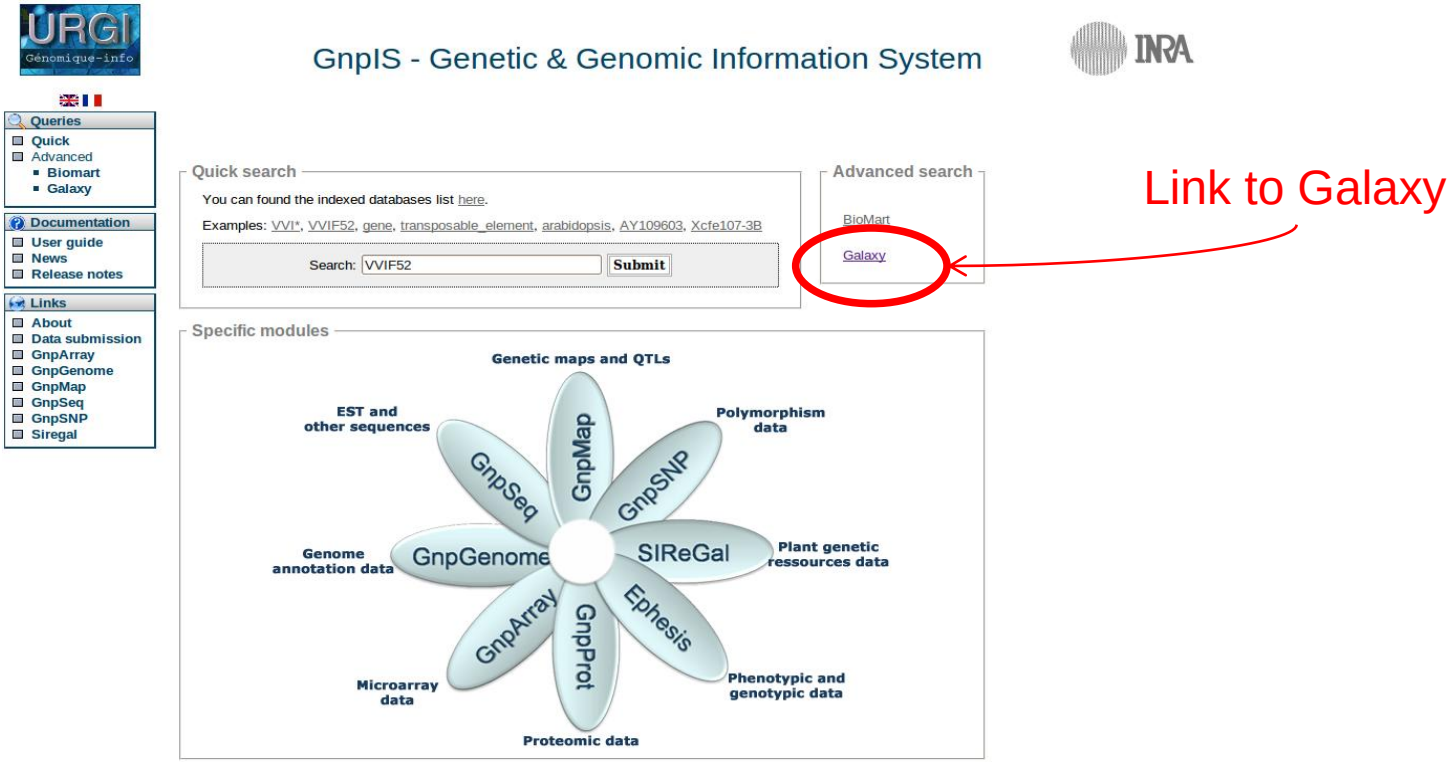


ARNOUX Sandie
BRAS Marc

V. Outlook beyond SNP detection

- Interoperability with the BioMart GnpIS URGI server.
  → http://urgi.versailles.inra.fr/gnpis/

- Extension of the system with other tools and pipelines (IBISA 2010 ...).



ARNOUX Sandie
BRAS Marc

Acknowledgement

URGV	Anne-Francoise ADAM		
URGV	Patricia FAIVRE RAMPANT		
URGI	Nathalie CHOISNE	**EPGV**	Dominique BRUNEL
URGI	Sebastien REBOUX	**EPGV**	Marie-Christine LE PASLIER
URGI	Olivier INIZAN	**EPGV**	Stéphane SCHLUB
URGI	Nacer MOHELLIBI		
URGI	Delphine STEINBACH	**GAFL**	Stéphane MUNOS
URGI	Hadi QUESNEVILLE	**GAFL**	Jean-Paul BOUCHET