

Development of a pipeline for SNPs detection

Colloque EPGV 2010:

Détection, Gestion et Analyse du Polymorphisme
Des Génomes Végétaux



I. Background and objectives of the pipeline

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

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 different file formats (GFF3, ACE, AFG, MAQ, SOAP, SAM et BAM).
- **GenomeView**: software used to view different 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	,\$,,,,,,,,,,,,,,,,,,,,,	ZZZZZZZZZZZZZZZZTTZVVVX
C10HBa0111D09_LR276	101	C	24	,,,\$,,,,,,,,,,,,,,,,,,,,^.	ZZZZZZZZZZZZYLXZSQWZTQSQZ
C10HBa0111D09_LR276	102	C	23	,,,,,,,,,,,,,,,,,,,,,	ZZZZZZZYZZRZWZWJJROQUKZ
C10HBa0111D09_LR276	103	T	24	,\$,,,,,,,,,,,,,,,,,,,,^,	ZZZQZZZZZZZZZZZZVWVVZU
C10HBa0111D09_LR276	104	T	27	,,,,,,,,,,,,,,,,,,,,^,^,^,^,	ZZTZZZZZZZZZZZZZZVTZXVUVV
C10HBa0111D09_LR276	105	A	28	,\$,\$,,,,,,,,,,,,,,,,,,,,^F,	ZZKZTZZZZZZZZZZZZZZZXVUNXU
C10HBa0111D09_LR276	106	A	28	T,,,,,,,,,,,,,,,,,,,,^,^,	AZZZZZZZZZZZZZZZZZXVZXXVTT

Reference
Seq. Name

Base
In this
Position

Informations
For All
Short Reads

Base
Quality

Position
In Ref.

Number
Of Short Reads
In this Position

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	0.46666666666666698
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 !

II. Software selected

Tablet: SNPs view

Tomate_vs_Lot8_sorted.bam - Tablet - x.xx.xx.xx

C10HBa0111D09_LR276 | consensus length: 9 300 (9 300) | reads: 191 204 | features: 218 | Memory usage: 258,14 MB (7)

Home

Open Assembly Import Features

Enhanced Classic

Data Layout Style

Packed Stacked Sort

Zoom: Variants: Adjust

Page Left Page Right Jump to Base

Prev Feature Next Feature

Navigate Options

Contigs (198):

Contig	Length	Reads	Fea...	Mis...
C10HBa011...	9300	191204	218	0,5
C11HBa002...	10969	0	0	0,0
C11HBa003...	9056	0	0	0,0
C11HBa005...	10301	0	0	0,0
C11HBa006...	10050	0	0	0,0
C11HBa006...	9385	0	0	0,0
C11HBa007...	9556	0	0	0,0
C11HBa008...	9244	0	0	0,0
C11HBa009...	9184	0	0	0,0
C11HBa010...	9115	0	0	0,0
C11HBa013...	10002	0	0	0,0
C11HBa014...	10785	0	0	0,0
C11HBa016...	9057	0	0	0,0
C11HBa016...	9826	0	0	0,0
C11HBa019...	10992	0	0	0,0
C11HBa024...	10008	0	0	0,0
C11HBa030...	9430	0	0	0,0
C11HBa032...	9657	0	0	0,0
C11SLe0053...	9827	0	0	0,0
C11SLm005...	10013	0	0	0,0
C12HBa115...	10021	0	0	0,0
C12HBa120...	10271	0	0	0,0
C12HBa144...	9247	0	0	0,0
C12HBa149...	9271	0	0	0,0
C12HBa165...	9257	0	0	0,0
C12HBa183...	9473	0	0	0,0
C12HBa221...	10755	0	0	0,0
C12HBa224...	9130	0	0	0,0
C12HBa26C...	9139	0	0	0,0
C12HBa326...	10414	0	0	0,0
C12HBa90D...	9638	0	0	0,0

1 to 9 300 (9,3 Kbp)

Coverage information

Bases with coverage: 100%

Average coverage depth: 738,42

Maximum coverage depth: 7 744

5 148 to 5 191 (44 bp)

5 148 U5 148

5 191 U5 191

Tablet Tip: Position data is often supplemented with U (unpadded position) and CV (read coverage at that position) values

II. Software selected

Tablet: SNPs view

Tomate_vs_Lot8_sorted.bam - Tablet - x.xx.xx.xx

C10HBa0111D09_LR276 | consensus length: 9 300 (9 300) | reads: 191 204 | features: 218 | Memory usage: 106,89 MB (7)

Home

Open Assembly Import Features

Enhanced Classic

Packed Stacked Sort

Zoom: [Slider] Variants: [Slider] Adjust

Page Left Page Right Jump to Base

Prev Feature Next Feature

Options

Contigs (198):

Contig	Length	Reads	Fea...	Mis...
C10HBa011...	9300	191204	218	0,5
C11HBa002...	10969	0	0	0,0
C11HBa003...	9056	0	0	0,0
C11HBa005...	10301	0	0	0,0
C11HBa006...	10050	0	0	0,0
C11HBa006...	9385	0	0	0,0
C11HBa007...	9556	0	0	0,0
C11HBa008...	9244	0	0	0,0
C11HBa009...	9184	0	0	0,0
C11HBa010...	9115	0	0	0,0
C11HBa013...	10002	0	0	0,0
C11HBa014...	10785	0	0	0,0
C11HBa016...	9057	0	0	0,0
C11HBa016...	9826	0	0	0,0
C11HBa019...	10992	0	0	0,0
C11HBa024...	10008	0	0	0,0
C11HBa030...	9430	0	0	0,0
C11HBa032...	9657	0	0	0,0
C11SLe0053...	9827	0	0	0,0
C11SLm005...	10013	0	0	0,0
C12HBa115...	10021	0	0	0,0
C12HBa120...	10271	0	0	0,0
C12HBa144...	9247	0	0	0,0
C12HBa149...	9271	0	0	0,0
C12HBa165...	9257	0	0	0,0
C12HBa183...	9473	0	0	0,0
C12HBa221...	10755	0	0	0,0
C12HBa224...	9130	0	0	0,0
C12HBa26C...	9139	0	0	0,0
C12HBa326...	10414	0	0	0,0
C12HBa90D...	9638	0	0	0,0

1 to 9 300 (9,3 kbp)

Coverage information
 Bases with coverage: 100%
 Average coverage depth: 738,42
 Maximum coverage depth: 7 744

5 148 to 5 191 (44 bp)

5 148 U5 148

5 191 U5 191

Tablet Tip: Position data is often supplemented with U (unpadded position) and CV (read coverage at that position) values

II. Software selected

GenomeView: SNPs view

GenomeView :: 922

File Edit Navigation Selection Plugins Help

Entry: C10HBa0111D09_LR276

1 901 1801 2701 3601 4501 5401 6301 7201 8101 9001

51 b

X T N N P F L E V L A R R A A Y T
 N K Q P L F F G V V G A S C S L H S
 E Q T T P P F W R C W R V V Q L T L
 G A A C A A A C A A C C C T T T T T G G A G G T G T T G G C G C G T C G T G C A G C T T A C A C T C
 1 6 11 16 21 26 31 36 41 46
 C T T G T T T G T T G G G A A A A A C C T C C A C A A C C G C G C A G C A C G T C G A A T G T G A G
 F L C G R K P P T P A D H L K C E
 V F L G K K S T N A R R A A * V *
 X C V V G K Q L H Q R T T C S V S
 1 6 11 16 21 26 31 36 41 46

CDS
 16384.0
 /Tomate_vs_Lot8_sorted.bam (1)
 0.0

SNPs

T
 C
 A

Data sources

Data source	V...
/Genome_Tomate_filtre_header.f...	✓
/Tomate_vs_Lot8_sorted.bam	✓

Track list

S...	A...	C...	upd...	Track name
✓	✓	✓	↑ ↓	Gene structure
✓	✓	✓	↑ ↓	Ruler
✓	✓	✗	↑ ↓	CDS
✓	✓	✗	↑ ↓	Short reads: /home/...

Features CDS

Name	St...	StopInt...	Sp...

Details on selected items:

III. Integration of tools in Galaxy

Return to homepage

Tools

Management of Galaxy

History

URGI public site

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

The screenshot displays the URGI public site interface, which is a Galaxy web portal. At the top, there is a navigation bar with the 'Galaxy' logo and menu items: 'Analyze Data', 'Workflow', 'Data Libraries', 'Help', and 'User'. Below this, a breadcrumb trail shows 'You are here : Home' with sub-links for 'Data', 'Tools', 'Species', and 'Contact'. A search bar is located on the right side of the top navigation.

The main content area features a 'Welcome to URGI' message with the INRA logo. A central graphic consists of a flower-like shape with petals labeled 'GnpMap', 'GnpGroup', 'SiReGal', 'Ephias', 'GnpPhor', and 'GnpGenome'. Surrounding this central graphic are numerous small square images of various plants and crops, including tomatoes, green beans, roots, grapes, a rose, strawberries, and a corn cob.

On the left side, there is a 'Tools' sidebar with a list of categories and sub-items: 'Get Data', 'Send Data', 'ENCODE Tools', 'Lift-Over', 'Text Manipulation', 'Filter and Sort', 'Join, Subtract and Group', 'Convert Formats', 'Extract Features', 'Fetch Sequences', 'Fetch Alignments', 'Get Genomic Scores', 'Operate on Genomic Intervals', 'Statistics', 'Graph/Display Data', 'Regional Variation', 'Multiple regression', 'Multivariate Analysis', 'Evolution', 'Metagenomic analyses', 'FASTA manipulation', 'NGS: QC and manipulation', 'NGS: Mapping', 'NGS: SAM Tools', 'NGS: Peak Calling', 'Rg Data', 'Rg Simulate', 'Rg Visualise', 'Rg Model Data', 'MyTools', and 'Workflows'. A 'URGI Génomique-info' menu is also present with options like 'Home', 'About us', 'Projects', 'Resources', 'Research', 'Register', and 'Site map'.

On the right side, there is a 'History' panel with an 'Options' dropdown, a 'Your history is empty. Click 'Get Data' on the left pane to start' message, and a 'Get Data' button.

Below the main content area, there are two sections: 'What's new?' and 'Events'. The 'What's new?' section includes a post from March 25th 2010 about 'GnpMap' and another from March 12th 2010 about 'new funny private databanks'. The 'Events' section includes a post from April 14-15th 2010 about a 'Second training session : Talend Open Studio' and another from March 16-17th about a 'Second training session : Extrem programming'.

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

a. Tools → Get Data → Upload File

Tools

- Get Data
 - Upload File from your computer

File Format:
 Auto-detect

Which format? See help below

File:
 Parcourir...

URL/Text:

Here you may specify a list of URLs (one per line) or paste the contents of a file.

Convert spaces to tabs:
 Yes
 Use this option if you are entering intervals by hand.

Genome:
 Click to Search or Select Build

Execute

Auto-detect
 The system will attempt to detect Axt, Fasta, Fastqsolexa, Gff, Gff3, Html, Lav, Maf, Tabular, Wiggle, Bed and Interval (Bed with headers) formats. If your file is not detected properly as one of the known formats, it most likely means that it has some format problems (e.g., different number of columns on different rows). You can still coerce the system to set your data to the format you think it should be. You can also upload compressed files, which will automatically be decompressed.

Ab1
 A binary sequence file in 'ab1' format with a '.ab1' file extension. You must manually select this 'File Format' when uploading the file.

Axt
 blastz pairwise alignment format. Each alignment block in an axt file contains three lines: a summary line and 2 sequence lines. Blocks are separated from one another by blank lines. The summary line contains chromosomal position and size information about the alignment. It consists of 9 required fields.

Bam
 A binary file compressed in the BGZF format with a '.bam' file extension.

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

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

The screenshot shows the Galaxy web interface. At the top, there is a navigation bar with 'Analyze Data', 'Workflow', 'Data Libraries', 'Help', and 'User'. On the left, there is a 'Tools' sidebar with various categories like 'Get Data', 'Text Manipulation', etc. On the right, there is a 'History' sidebar showing a list of jobs.

A red box highlights a notification message in the center of the page:

i Your upload has been queued. History entries that are still uploading will be blue, and turn green upon completion.
Please do not use your browser's "stop" or "reload" buttons until the upload is complete, or it may be interrupted.
 You may safely continue to use Galaxy while the upload is in progress. Using "stop" and "reload" on pages other than Galaxy is also safe.

A red box also highlights the 'History' sidebar, which shows a job titled '2: Genome_Tomate.fasta' with a blue background, indicating it is still uploading.

A diagram in the center of the page, enclosed in a dotted red box, illustrates the status of jobs in a workflow:

Waiting	→	5: (BWA) Output SAM
Ongoing	→	2: Genome_Tomate.fasta
Finished → Error	→	5: (BWA) Output SAM
Finished → OK	→	3: Genome_Tomate.fasta

The diagram shows that jobs are represented by colored boxes: grey for 'Waiting', blue for 'Ongoing', red for 'Finished → Error', and green for 'Finished → OK'. Arrows point from the text labels to the corresponding job boxes. A red arrow points from the 'Finished → Error' row to the 'History' sidebar, indicating that the error state corresponds to the job shown in the history.

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

a. Choose a tool from the list

The screenshot shows the Galaxy web interface with the following elements:

- Left Panel (Tools):** A list of tool categories. The 'NGS: QC and manipulation' category is highlighted with a red box, and 'Map with Bowtie for Illumina' is selected within it.
- Central Panel (Tool Configuration):** Titled 'Map with Bowtie for Illumina'. It contains several configuration options:
 - 'Will you select a reference genome from your history or use a built-in index?': 'Use a built-in index' (selected).
 - 'Select a reference genome': A dropdown menu.
 - 'Is this library mate-paired?': 'Single-end' (selected).
 - 'FASTQ file': A dropdown menu.
 - 'Bowtie settings to use': 'Commonly used' (selected).
 - 'Suppress the header in the output SAM file': A checked checkbox.
 - 'Execute' button: An orange arrow points to this button.
- Right Panel (History):** Shows 'Unnamed history' and a message: 'Your history is empty. Click 'Get Data' on the left pane to start'.
- Annotations:**
 - A green bracket and arrow labeled 'b. Set options and parameters' points to the configuration options.
 - An orange arrow labeled 'c. Execute' points to the 'Execute' button.
 - A 'Help' link is visible next to the 'Execute' button.

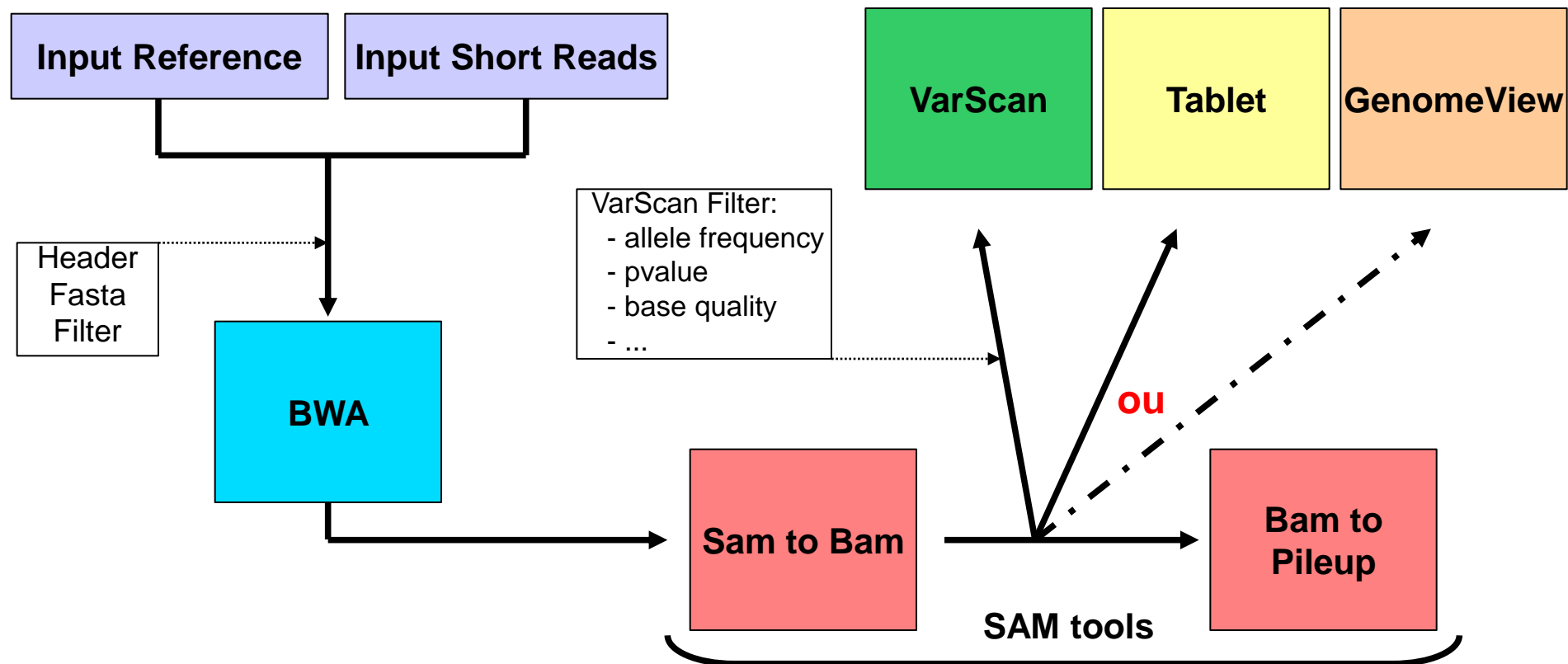
III.2. How to use a tool on Galaxy ?

The screenshot displays the Galaxy web interface. At the top, there is a navigation bar with the 'Galaxy' logo and menu items: 'Analyze Data', 'Workflow', 'Data Libraries', 'Help', and 'User'. On the left side, there is a 'Tools' sidebar with various categories like 'FASTA manipulation', 'NGS: QC and manipulation', 'NGS: Mapping', 'NGS: SAM Tools', 'NGS: Peak Calling', 'Rg Data', 'Rg Simulate', 'Rg Visualise', 'Rg Model Data', 'MyTools', and 'Workflows'. The main content area features a green notification box with a checkmark icon, stating: 'The following job has been successfully added to the queue: 15: REF.fasta Modif'. Below this, it provides instructions: 'You can check the status of queued jobs and view the resulting data by refreshing the History pane. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered.' On the right side, there is a 'History' pane with an 'Options' dropdown. It shows a list of jobs under 'Unnamed history', including '15: REF.fasta Modif' and '14: Genome_Tomate.fasta'. A red arrow labeled 'Output' points from the notification box to the '15: REF.fasta Modif' entry in the history pane. A blue arrow labeled 'Input' points from the '14: Genome_Tomate.fasta' entry in the history pane to the notification box.

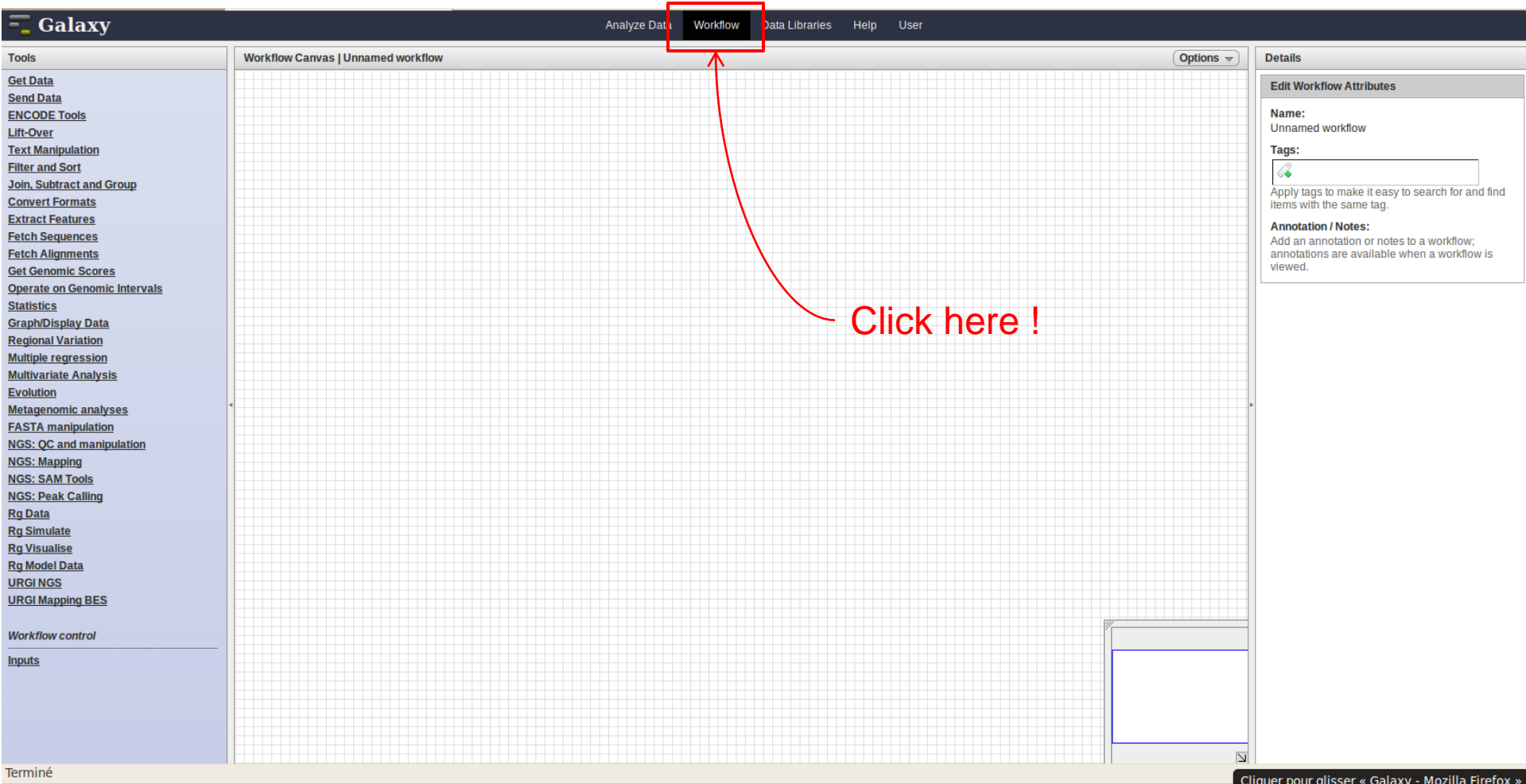
III.3. How to integrate a tool in Galaxy ?

We can easily integrate a new tool !!!

III.4. Our workflow for SNPs detection.



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



The screenshot shows the Galaxy web interface. The top navigation bar includes 'Analyze Data', 'Workflow', 'Data Libraries', 'Help', and 'User'. The 'Workflow' menu is highlighted with a red box, and a red arrow points to it from the text 'Click here!'. The main area is a 'Workflow Canvas | Unnamed workflow' with a grid background. On the left, there is a 'Tools' sidebar with various tool categories like 'Get Data', 'Text Manipulation', 'Statistics', etc. On the right, there is a 'Details' sidebar with 'Edit Workflow Attributes' section, including fields for 'Name' (Unnamed workflow), 'Tags', and 'Annotation / Notes'. The bottom status bar shows 'Terminé' and a mouseover tooltip: 'Cliquez pour glisser « Galaxy - Mozilla Firefox »'.

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

The screenshot displays the Galaxy workflow editor interface. On the left, the 'Tools' panel is visible, with the 'Inputs' category selected and the 'Input dataset' tool highlighted. The central 'Workflow Canvas' shows an 'Unnamed workflow' with two 'Input Dataset' tool steps. A blue arrow points from the 'Input dataset' tool in the 'Inputs' panel to the top 'Input Dataset' step in the canvas. The 'Details' panel on the right shows the configuration for the selected 'Input dataset' step, including a name field and an annotation field.

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

Parameters, Help, ...

Bwa

URGINGS

- Map with Bwa 0.5.7 Map Short Reads to Reference sequence with BWA 0.5.7.
- Header Fasta Filter Modify all of this header file that contains one, or multiple, tabulation and informations after the name of the sequence.
- Sam Filter All alignments that are outside the subject sequence (partially or not), are removed from

Details

Tool: Map with Bwa 0.5.7

Type of Short Reads
Single-end

Reference File (.fasta)
Data input 'input_REF' (fasta)

Short Reads File (.fastq)
Data input 'input_SR' (fastq)

Algorithm for Reference indexing
is

Use default parameters for Bwa
Yes

Edit Step Attributes

Annotation / Notes:

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

Link to Galaxy's Website: <http://bio-bwa.sourceforge.net/bwa.shtml>

AUTHOR:

Heng Li at the Sanger Institute wrote the key source codes and integrated the following codes for BWT construction: bwtsv (<http://l.cs.hku.hk/~ckwong3/bwtsv/>), implemented by Chi-Kwong Wong at the University of Hong Kong and IS (<http://yuta.256.googlepages.com/sais>) originally proposed by Nong Ge (<http://www.cs.sysu.edu.cn/nong/>) at the Sun Yat-Sen University and implemented by Yuta Mori.

LICENSE AND CITATION:

The full BWA package is distributed under GPLv3 as it uses source codes from BWT-SW which is covered by GPL - Sorting_hash_table_BWT and JS

http://localhost:8080/workflow/editor?id=f06d708dacad69d8#

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

Link the two tools

Tool: Map with Bwa 0.5.7

Type of Short Reads:

Reference File (.fasta): Data input 'input_REF' (fasta)

Short Reads File (.fastq): Data input 'input_SR' (fastq)

Algorithm for Reference indexation:

Use default parameters for Bwa:

Link to Galaxy's Website: <http://bio-bwa.sourceforge.net/bwa.shtml>

AUTHOR:
Heng Li at the Sanger Institute wrote the key source codes and integrated the following codes for BWT construction: bwtsv (<http://li.cs.hku.hk/~ckwong3/bwtsv/>), implemented by Chi-Kwong Wong at the University of Hong Kong and IS (<http://yuta.256.googlepages.com/sais>) originally proposed by Nong Ge (<http://www.cs.sysu.edu.cn/nong/>) at the Sun Yat-Sen University and implemented by Yuta Mori.

LICENSE AND CITATION:
The full BWA package is distributed under GPLv3 as it uses source codes from BWT-SW which is covered by GPL. Sorting, hash table, BWT and IS

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

The screenshot displays the Galaxy workflow editor interface. The main canvas shows a workflow with two 'Input Dataset' tools connected to a 'Map with Bwa 0.5.7' tool. The 'Map with Bwa 0.5.7' tool has several outputs: 'output_RES (sam)', 'output_ERROR_INDEX (text)', 'output_ERROR_ALN (text)', and 'output_ERROR_BINtoSAM (text)'. The 'output_RES (sam)' output is connected to a 'SAM-to-BAM' tool. The 'SAM-to-BAM' tool has two inputs: 'Convert SAM file' (connected to 'output_RES (sam)') and 'Using reference file' (connected to 'output_ERROR_INDEX (text)'). The 'SAM-to-BAM' tool has one output: 'output1 (bam)'. A purple box highlights the 'SAM-to-BAM' tool in the workflow canvas, and a purple arrow points from a text label 'Sam to Bam' to it. The left sidebar shows a list of tools, with 'SAM-to-BAM' highlighted. The right sidebar shows the details for the 'SAM-to-BAM' tool, including the tool name, reference list source, and step attributes.

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

The screenshot shows the Galaxy workflow editor interface. The workflow canvas contains the following steps:

- Input Dataset** (two instances) - outputs connect to the **Map with Bwa 0.5.7** tool.
- Map with Bwa 0.5.7** - outputs include `output_RES (sam)`, `output_ERROR_INDEX (text)`, `output_ERROR_ALN (text)`, and `output_ERROR_BINtoSAM (text)`.
- SAM-to-BAM** - takes `output_RES (sam)` and `Using reference file` as input, outputs `output1 (bam)`.
- Generate pileup** - takes `output1 (bam)` and `Select a reference genome` as input.

In the left sidebar, under **NGS: SAM Tools**, the tool **Generate pileup from BAM dataset** is highlighted with a red box. An orange arrow points from this tool to the **Generate pileup** tool in the workflow canvas. The text **Bam to Pileup** is written in orange next to the arrow.

The **Details** panel for the **Generate pileup** tool shows the following options:

- Tool:** Generate pileup
- Will you select a reference genome from your history or use a built-in index?** (Use one from the history)
- Select the BAM file to generate the pileup file for** (Data input 'input1' (bam))
- Select a reference genome** (Data input 'ownFile' (fasta))
- Whether or not to print the mapping quality as the last column** (Do not print the mapping quality as the last column)
- Whether or not to print only output pileup lines containing indels** (Print all lines)
- Where to cap mapping quality** (60)
- Call consensus according to MAQ model?** (No)

The **Edit Step Attributes** section includes an **Annotation / Notes** field and a **What it does** section describing the tool's function.

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

The screenshot shows a Galaxy workflow editor with the following steps:

- Input Dataset** (two instances) feeds into **Map with Bwa 0.5.7**.
- Map with Bwa 0.5.7** outputs: output_RES (sam), output_ERROR_INDEX (text), output_ERROR_ALN (text), output_ERROR_BINtoSAM (text).
- These outputs feed into **SAM-to-BAM**.
- SAM-to-BAM** outputs: Convert SAM file, Using reference file, output1 (bam).
- These outputs feed into **Generate pileup**.
- Generate pileup** outputs: Select the BAM file to generate the pileup file for, Select a reference genome, output1 (tabular).
- These outputs feed into **Tablet (v.1.10.03.04)**.
- Tablet (v.1.10.03.04)** outputs: Fichier Bam / Sam / Map / autres..., Référence Input.fasta.

The **Tablet** step is highlighted with a blue box and an arrow pointing to it from the word "Tablet" written in the center of the canvas. A small thumbnail of the Tablet interface is shown in the bottom right corner of the workflow canvas.

Details Panel (Right):

- Tool:** Tablet (v.1.10.03.04)
- Data input:** 'input_FICHER' (bam), 'input_REF' (fasta)
- Edit Step Attributes:** Annotation / Notes: (empty text area)
- Tablet:** Tablet is a lightweight, high-performance graphical viewer for next generation sequence assemblies and alignments. This software is able to view sequences alignments. It can use GFF3, ACE, AFG, MAQ, SOAP, SAM and BAM Files.
- Tablet features:**
 - High-performance visualization and data navigation.
 - Display of reads in both packed and stacked formats.
 - File format support for ACE, AFG, MAQ, SOAP, SAM and BAM.
 - Import GFF3 features and quickly find/highlight them.
 - Search and locate reads by name across entire data sets.
 - Entire-contig overviews, showing data layout or coverage.
 - Simple install routine via auto-updating graphical installer.
 - Support for Windows, Apple Mac OS X, Linux and Solaris.
- Website:** <http://bioinf.scri.ac.uk/tablet/>

Tools Panel (Left):

- Graph/Display Data
- Regional Variation
- Multiple regression
- Multivariate Analysis
- Evolution
- Metagenomic analyses
- FASTA manipulation
- NGS: QC and manipulation
- NGS: Mapping
- NGS: SAM Tools
- NGS: Peak Calling
- Rg Data
- Rg Simulate
- Rg Visualise
- Rg Model Data
- URGI NGS
 - Map with Bwa 0.5.7 Map Short Reads to Reference sequence with BWA 0.5.7.
 - Map with Bwa 0.5.7 (cmd lines) Map Short Reads to Reference sequence with BWA 0.5.7 (and command lines).
 - Header Fasta Filter Modify all of this header file that contains one, or multiple, tabulation and informations after the name of the sequence.
 - Sam Filter All alignments that are outside the subject sequence (partially or not), are removed from the input file and copy to an output file.
 - VarScan VarScan: convert bam file to pileup file with differents filters.
 - Tablet (v.1.10.03.04) Alignment Viewer (can use GFF3, ACE, AFG, MAQ, SOAP, SAM and BAM Files).**
 - Genomeview (v.922) Alignment Viewer (can use BAM, GFF, FASTA and Annotation Files).

URL: <http://localhost:8080/workflow/editor?id=f06d708dacad69d8#>

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

Workflow Canvas | Unnamed workflow

Tools

- Graph/Display Data
- Regional Variation
- Multiple regression
- Multivariate Analysis
- Evolution
- Metagenomic analyses
- FASTA manipulation
- NGS: QC and manipulation
- NGS: Mapping
- NGS: SAM Tools
- NGS: Peak Calling
- Rg Data
- Rg Simulate
- Rg Visualise
- Rg Model Data
- URGI NGS
 - Map with Bwa 0.5.7 Map Short Reads to Reference sequence with BWA 0.5.7.
 - Map with Bwa 0.5.7 (cmd lines) Map Short Reads to Reference sequence with BWA 0.5.7 (and command lines).
 - Header Fasta Filter Modify all of this header file that contains one, or multiple, tabulation and informations after the name of the sequence.
 - Sam Filter All alignments that are outside the subject sequence (partially or not), are removed from the input file and copy to an output file.
 - VarScan VarScan: convert bam file to pileup file with differents filters.
 - Tablet (v.1.10.03.04) Alignment Viewer (can use GFF3, ACE, AFG, MAQ, SOAP, SAM and BAM Files).**
 - Genomeview (v.922) Alignment Viewer (can use BAM, GFF, FASTA and Annotation Files).

URGI Mapping BES

<http://localhost:8080/workflow/editor?id=f06d708dacad69d8#>

Details

Tool: Tablet (v.1.10.03.04)

Fichier Bam / Sam / Map / autres...
Data input 'Input_FICHIER' (bam)
RÃ©fÃ©rence Input.fasta
Data input 'Input_REF' (fasta)

Edit Step Attributes

Annotation / Notes:

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

Tablet:

Tablet is a lightweight, high-performance graphical viewer for next generation sequence assemblies and alignments. This software is able to view sequences alignments. It can use GFF3, ACE, AFG, MAQ, SOAP, SAM and BAM Files.

Tablet features:

- High-performance visualization and data navigation.
- Display of reads in both packed and stacked formats.
- File format support for ACE, AFG, MAQ, SOAP, SAM and BAM.
- Import GFF3 features and quickly find/highlight them.
- Search and locate reads by name across entire data sets.
- Entire-contig overviews, showing data layout or coverage.
- Simple install routine via auto-updating graphical installer.
- Support for Windows, Apple Mac OS X, Linux and Solaris.

Website: <http://bioinf.scri.ac.uk/tablet/>

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

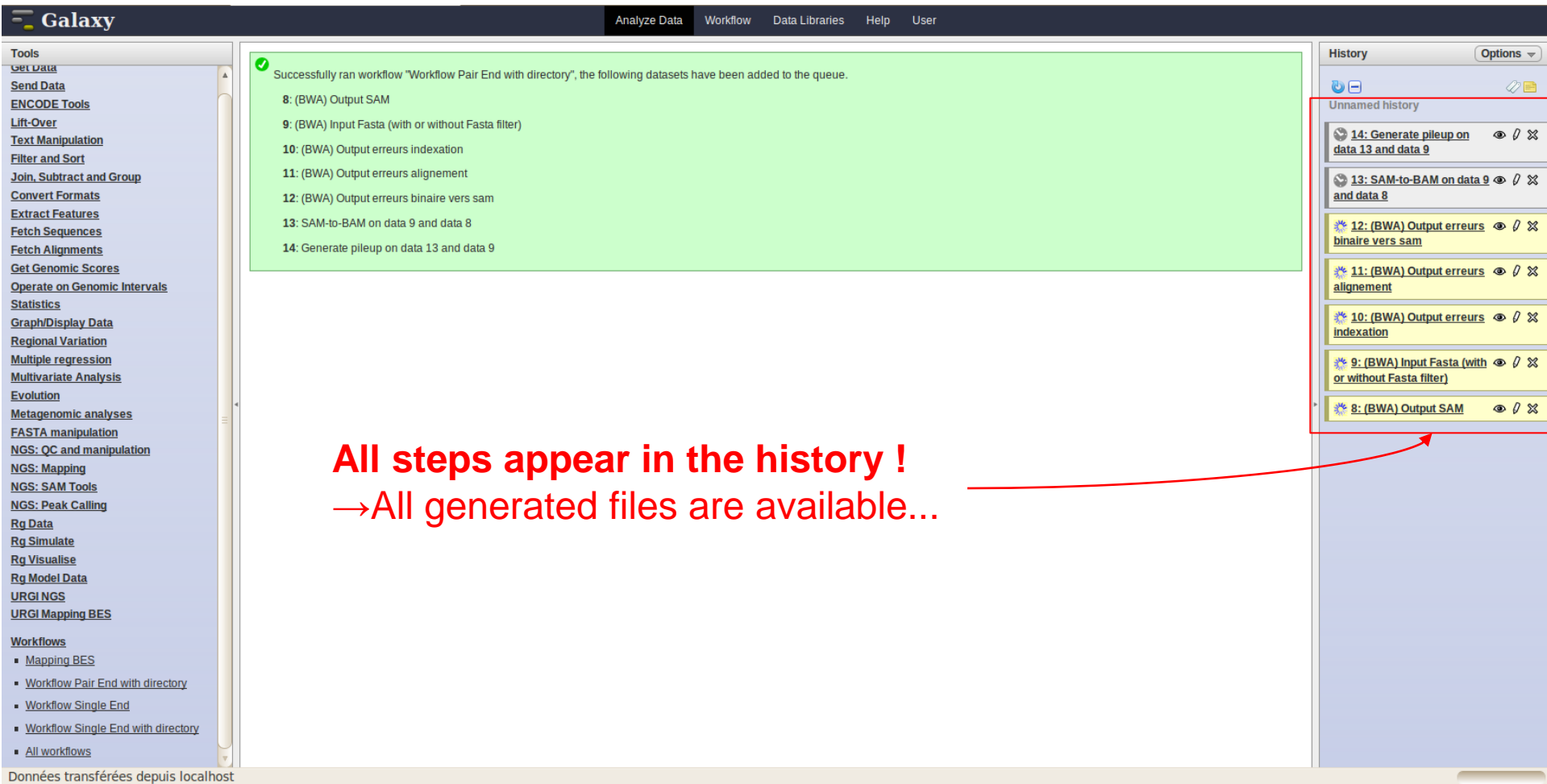
a. Choose the workflow from the list

b. Set options and parameters

c. Execute

http://localhost:8080/workflow/run?id=529fd61ab1c6cc36

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



The screenshot shows the Galaxy web interface. At the top, there's a navigation bar with 'Galaxy' and tabs for 'Analyze Data', 'Workflow', 'Data Libraries', 'Help', and 'User'. On the left, there's a 'Tools' sidebar with various categories like 'Get Data', 'Send Data', 'ENCODE Tools', etc. The main content area displays a green notification box: 'Successfully ran workflow "Workflow Pair End with directory", the following datasets have been added to the queue.' Below this, a list of datasets is shown: 8: (BWA) Output SAM, 9: (BWA) Input Fasta (with or without Fasta filter), 10: (BWA) Output erreurs indexation, 11: (BWA) Output erreurs alignement, 12: (BWA) Output erreurs binaire vers sam, 13: SAM-to-BAM on data 9 and data 8, and 14: Generate pileup on data 13 and data 9. On the right, the 'History' panel shows a list of these steps, with a red box highlighting the entire history list. A red arrow points from the text 'All steps appear in the history!' to this box. At the bottom left, it says 'Données transférées depuis localhost'.

All steps appear in the history !
 →All generated files are available...

III.7. Results in Tablet

scaffold_4 | consensus length: 23 188 140 (23 188 140) | reads: 1 088 | features: 91 | Memory usage: 575,97 MB (8)

Home Advanced

Open Assembly Import Features Data

Enhanced Direction Classic Layout Style

Packed Stacked Tag Variants

Zoom: Variants: Adjust

Page Left Page Right Jump to Base

Prev Feature Next Feature Navigate

Read Info Show Bases Read Names

RS Off RS Centre RS Custom Overlays

Contigs (2 518):

Contig	Length	Reads	Feat...	Mis...
scaffold_1	4836...	0	30	0,0
scaffold_2	2356...	0	134	0,0
scaffold_3	2021...	0	49	0,0
scaffold_4	2318...	1088	91	3,4
scaffold_5	2580...	0	0	0,0
scaffold_6	2689...	0	0	0,0
scaffold_7	1510...	0	0	0,0
scaffold_8	1883...	0	0	0,0
scaffold_9	1294...	0	0	0,0
scaffold_10	2153...	0	0	0,0
scaffold_11	1888...	0	0	0,0
scaffold_12	1492...	0	0	0,0
scaffold_13	1565...	0	0	0,0
scaffold_14	1771...	0	0	0,0
scaffold_15	1513...	0	0	0,0
scaffold_16	1413...	0	0	0,0
scaffold_17	1466...	0	0	0,0
scaffold_18	1496...	0	0	0,0
scaffold_19	1600...	0	0	0,0
scaffold_20	9089...	0	0	0,0
scaffold_21	6637...	0	0	0,0
scaffold_22	6065...	0	0	0,0
scaffold_23	7288...	0	0	0,0
scaffold_24	5893...	0	0	0,0
scaffold_25	4465...	0	0	0,0
scaffold_26	4095...	0	0	0,0
scaffold_27	4050...	0	0	0,0
scaffold_28	4828...	0	0	0,0
scaffold_30	3206...	0	0	0,0
scaffold_31	3646...	0	0	0,0
scaffold_32	3364...	0	0	0,0
scaffold_34	2425...	0	0	0,0
scaffold_35	2539...	0	0	0,0
scaffold_36	2696...	0	0	0,0
scaffold_37	2645...	0	0	0,0
scaffold_39	2237...	0	0	0,0
scaffold_40	2591...	0	0	0,0
scaffold_41	2471...	0	0	0,0
scaffold_43	2036...	0	0	0,0
scaffold_44	2377...	0	0	0,0
scaffold_45	1668...	0	0	0,0

1 to 25000 (25 Kb) 17 714 to 17 797 (84 bp)

17 714 U17 714 17 797 U17 797

Filter by: Name

Tablet Tip: Green (rather than blue) navigation arrows mean you have reached the edge of the current BAM data block

III.7. Results in Tablet

scaffold_4 | consensus length: 23 188 140 (23 188 140) | reads: 1 088 | features: 91 | Memory usage: 631,31 MB (7)

Home Advanced

Open Assembly Import Features Data

Enhanced Direction Classic Layout Style

Packed Stacked Tag Variants

Zoom: Variants: Adjust

Page Left Page Right Jump to Base Navigate

Prev Feature Next Feature

Read Info Show Bases Read Names RS Off RS Centre RS Custom Overlays

Contigs (2 518):

Contig	Length	Reads	Feat...	Mis...
scaffold_1	4836...	0	30	0,0
scaffold_2	2356...	0	134	0,0
scaffold_3	2021...	0	49	0,0
scaffold_4	2318...	1088	91	3,4
scaffold_5	2580...	0	0	0,0
scaffold_6	2689...	0	0	0,0
scaffold_7	1510...	0	0	0,0
scaffold_8	1883...	0	0	0,0
scaffold_9	1294...	0	0	0,0
scaffold_10	2153...	0	0	0,0
scaffold_11	1888...	0	0	0,0
scaffold_12	1492...	0	0	0,0
scaffold_13	1565...	0	0	0,0
scaffold_14	1771...	0	0	0,0
scaffold_15	1513...	0	0	0,0
scaffold_16	1413...	0	0	0,0
scaffold_17	1466...	0	0	0,0
scaffold_18	1496...	0	0	0,0
scaffold_19	1600...	0	0	0,0
scaffold_20	9089...	0	0	0,0
scaffold_21	6637...	0	0	0,0
scaffold_22	6065...	0	0	0,0
scaffold_23	7288...	0	0	0,0
scaffold_24	5893...	0	0	0,0
scaffold_25	4465...	0	0	0,0
scaffold_26	4095...	0	0	0,0
scaffold_27	4050...	0	0	0,0
scaffold_28	4828...	0	0	0,0
scaffold_30	3206...	0	0	0,0
scaffold_31	3646...	0	0	0,0
scaffold_32	3364...	0	0	0,0
scaffold_34	2425...	0	0	0,0
scaffold_35	2539...	0	0	0,0
scaffold_36	2696...	0	0	0,0
scaffold_37	2645...	0	0	0,0
scaffold_39	2237...	0	0	0,0
scaffold_40	2591...	0	0	0,0
scaffold_41	2471...	0	0	0,0
scaffold_43	2036...	0	0	0,0
scaffold_44	2377...	0	0	0,0
scaffold_45	1668...	0	0	0,0

Filter by: Name

1 to 25000 (25 kb)

17 714 to 17 797 (84 bp)

17 714 U17 714

17 797 U17 797

Tablet Tip: Green (rather than blue) navigation arrows mean you have reached the edge of the current BAM data block

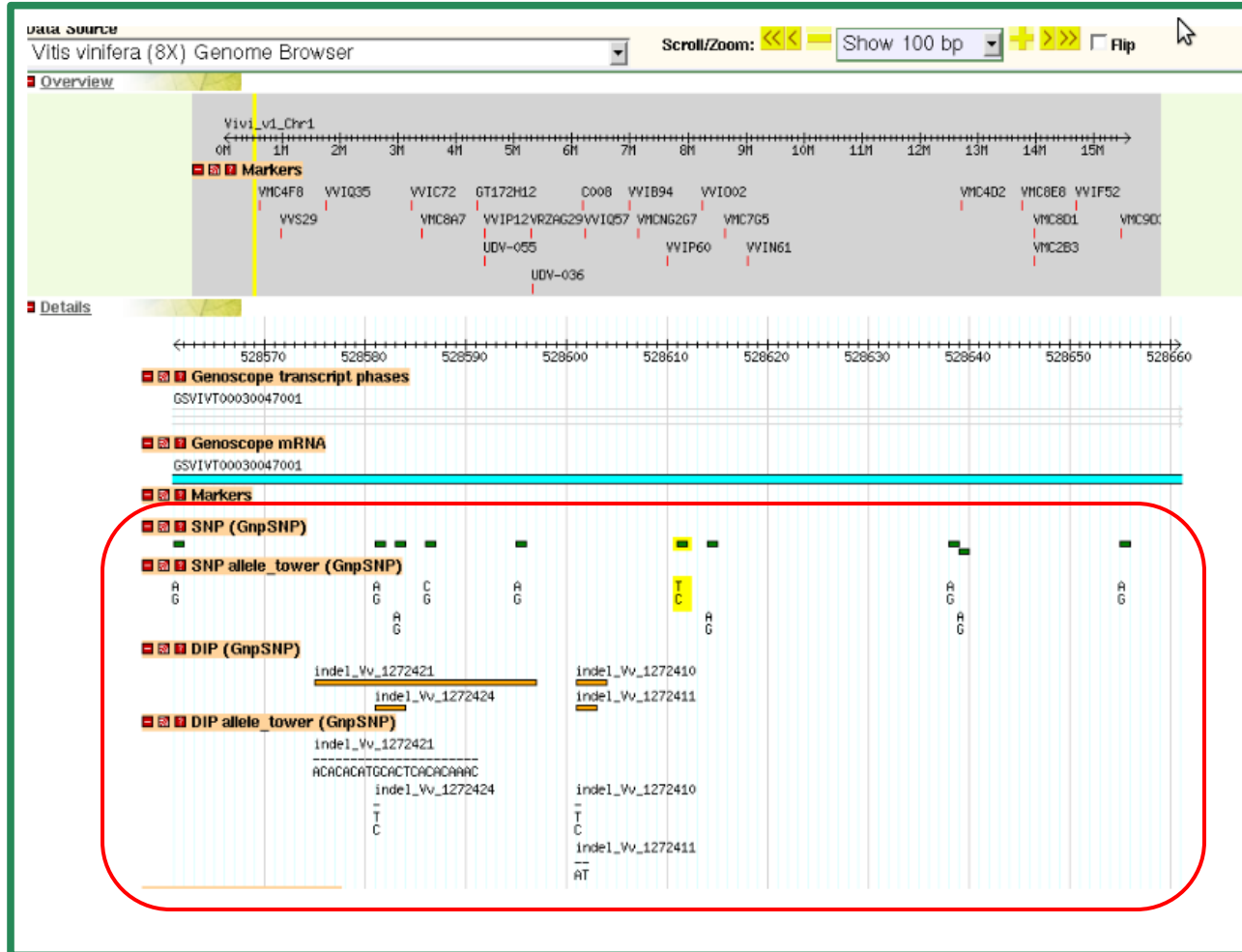
IV. Validation Process

- 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).

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.

V. Outlook



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 ...).

GnpIS - Genetic & Genomic Information System

Quick search

You can find the indexed databases list [here](#).

Examples: VVI*, VVIF52, gene, transposable_element, arabidopsis, AY109603, Xcfe107-3B

Search:

Advanced search

BioMart

[Galaxy](#)

Link to Galaxy

Specific modules

Genetic maps and QTLs

EST and other sequences

Polymorphism data

Plant genetic resources data

Phenotypic and genotypic data

Proteomic data

Microarray data

Genome annotation data

GnpSeq

GnpMap

GnpSNP

SReGal

Ephesis

GnpProt

GnpArray

GnpGenome

Acknowledgement

URGI Nathalie CHOISNE
URGI Sebastien REBOUX
URGI Olivier INIZAN
URGI Nacer MOHELLIBI
URGI Delphine STEINBACH
URGI Hadi QUESNEVILLE

URGV

Anne-Francoise ADAM

URGV

Patricia FAIVRE RAMPANT

EPGV

Dominique BRUNEL

EPGV

Marie-Christine LE PASLIER

EPGV

Stéphane SCHLUB

GAFL

Stéphane MUNOS

GAFL

Jean-Paul BOUCHET

