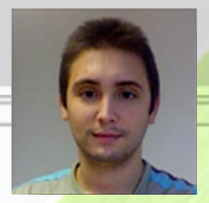
Development of a workflow for SNP detection with Galaxy




Marc Bras, Sandie Arnoux, Nacer Mohellibi, Nathalie Choisne, Olivier Inizan, Delphine Steinbach, Hadi Quesneville

Unité de Recherche en Génomique-Info UR1164, INRA de Versailles-Grignon, Route de Saint Cyr, Versailles, 78026, France

http://urgi.versailles.inra.fr urgi-contact@versailles.inra.fr

MAPHiTS : Mapping Analysis Pipeline for High-Throughput Sequences

Introduction :

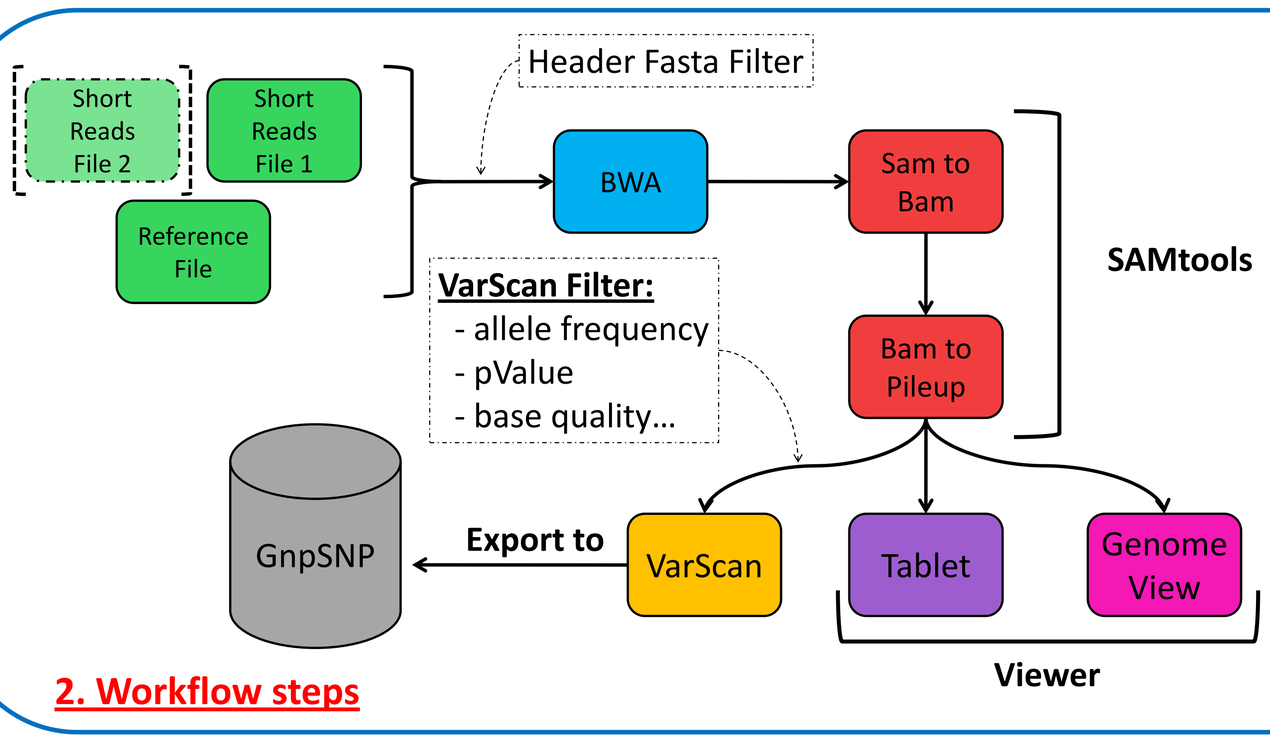
A Single-Nucleotide Polymorphism (SNP) is a DNA sequence variation. It can be used as a marker to characterize genetic variations between lineages. They can be used to detect complex traits such as those involved in diseases or agronomical performance.

The URGI platform developed a pipeline for SNPs detection from short reads, integrated in the Galaxy[1] workflow manager. *Galaxy* allows, through a web page, to chain different tools graphically. In addition, a large number of workflows can be built and shared.

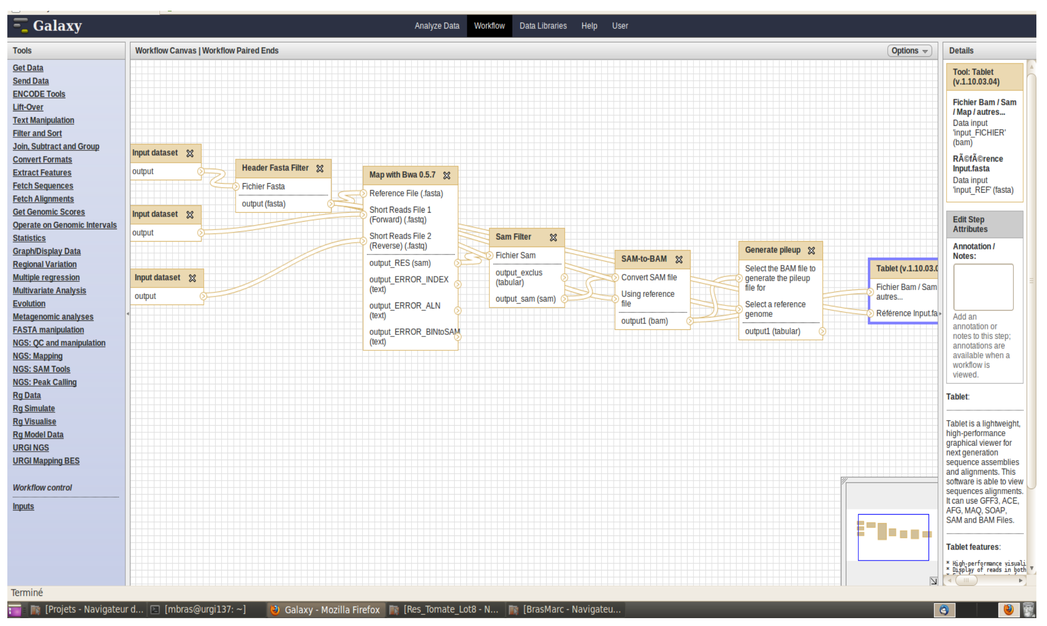
From a reference genome and a set of short reads (single-end or pair-ends), our workflow use *BWA*[2], *SAMtools*[3], *VarScan*[4] and *Tablet*[5] to predict SNPs and indels with number of filters, such as genome coverage, allele frequency, pValue.

Galaxy: web-based platform for genomic research, with many tools for NGS. They can be integrated into workflows.

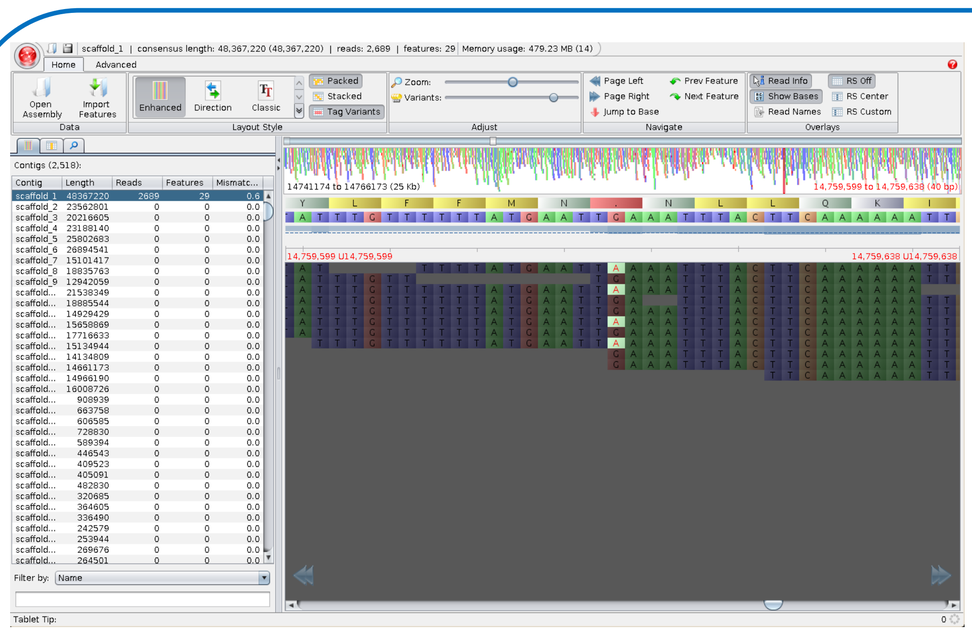
1. *Galaxy* home page



2. Workflow steps



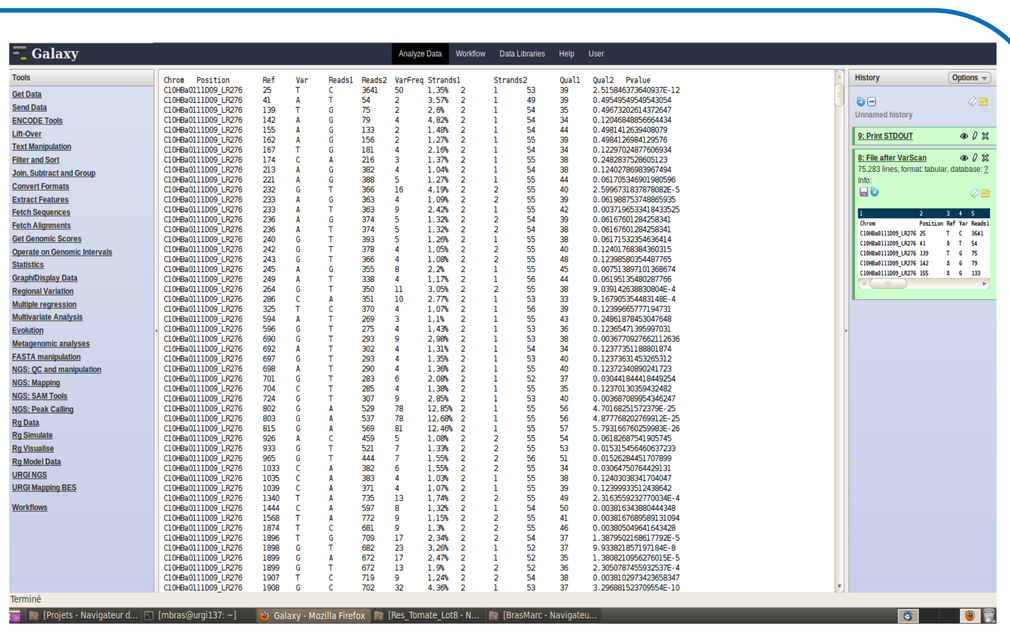
3. Workflow in *Galaxy*



4. Alignment results in *Tablet*

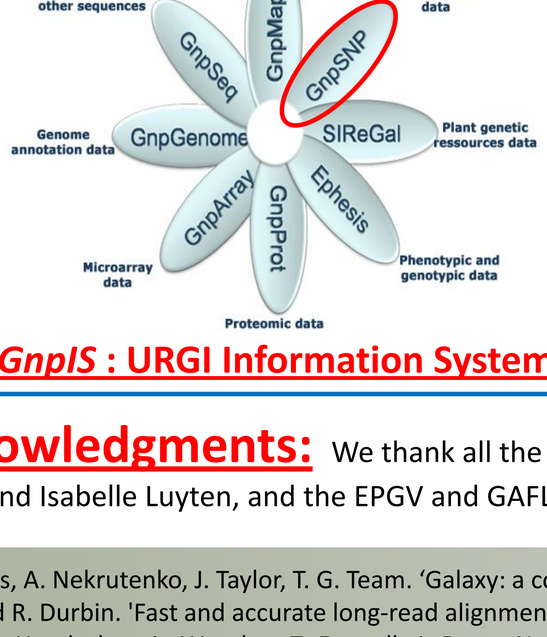
Results:

The alignment can be visualized with *Tablet* (or *GenomeView*[6] to see annotations). By changing the contrast, variants can be easily located (figure 4). *VarScan* can predict SNPs and indels. The output file contains information that can be used to filter them (variant frequency, average base quality, Pvalue…) (figure 5).

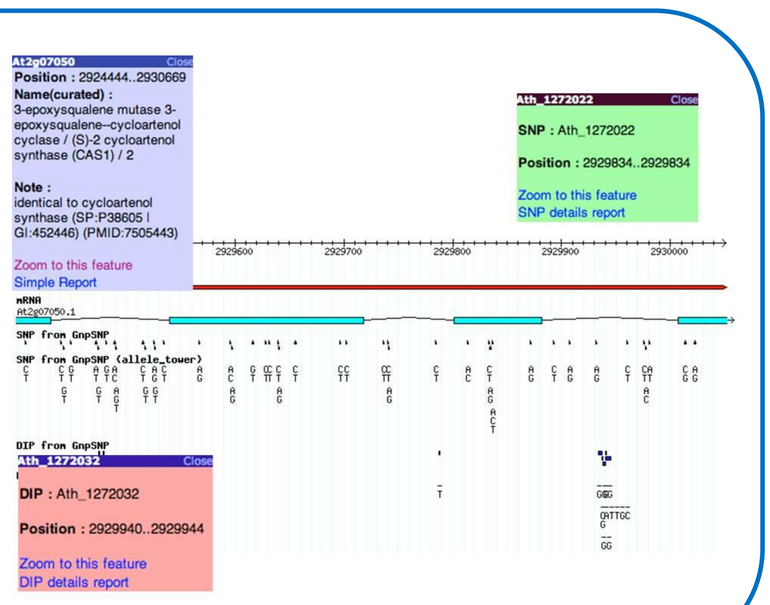


5. *VarScan* results in *Galaxy*

Insertion in database: the tool from *Galaxy* to *GnpSNP* (URGI database) export SNPs and indels to be visualized on Gbrowse.



6. *GnpIS* : URGI Information System



7. Results in *GnpSNP*

Conclusion :

Playing on different filter parameters and tools, this versatil workflow is able to detect SNPs and indels.

It has already been run on different data sets, from *A. thaliana, V. vinifera*, *Tomato* and *Poplar*.

The workflow will soon be available on our *Galaxy* instance.

Acknowledgments: We thank all the members of the URGI for their fruitful remarks, the members of the development team, the system and database administrators Sébastien Reboux and Isabelle Luyten, and the EPGV and GAFL teams. This work is supported by the ANR project GrapeReseq.

[1] J. Goecks, A. Nekrutenko, J. Taylor, T. G. Team. 'Galaxy: a comprehensive approach for supporting accessible, reproductible, and transparent computational research in the life sciences'. Genome Biology 11, R86+ (2010)

[2] H. Li and R. Durbin. 'Fast and accurate long-read alignment with Burrows-Wheeler transform'. Bioinformatics (2010). [PMID: 20080505]

[3] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin and 1000 Genome Project Data Processing Subgroup. 'The Sequence alignment/map (SAM) format and SAMtools'. Bioinformatics, 25, 2078-9 (2009). [PMID: 19505943]

[4] Koboldt DC, Chen K, Wylie T, Larson DE, McLellan MD, Mardis ER, Weinstock GM, Wilson RK, & Ding L (2009). 'VarScan: variant detection in massively parallel sequencing of individual and pooled samples'. Bioinformatics (Oxford, England), 25 (17), 2283-5 [PMID: 19542151]

[5] I. Milne, M. Bayer, L. Cardle, P. Shaw, G. Stephen, F. Wright and D. Marshall (2010). 'Tablet—next generation sequence assembly visualization'.Bioinformatics 2010 26(3):401-402.

[6] http://genomeview.sourceforge.net/