Development of a Galaxy workflow for SNP detection in Grapevine and Poplar whole genome Illumina resequencing data

Nathalie Choisne[1], Marc Bras[1], Nacer Mohellibi[1], Sandie Arnoux[1], Hadi Quesneville[1], Juliette Goarin[2], Jean-Michel Boursiquot[2], Vanina Guérin[3], Marie-Christine Le Paslier[4], Aurélie Bérard[4], Stéphane Schlub[4], Dominique Brunel[4], Rémi Bounon[5], Frédérique Bitton[5], Patricia Faivre-Rampant[5], Anne-Françoise Adam-Blondon[5]

[1] INRA-URGI (Unité de Recherche en Génomique-Info), UR1164, Versailles, France
[2] INRA-DIAPC (Diversité et Adaptation des Plantes Cultivées), UMR1097, Montpellier, France
[3] INRA-AGPF (Amélioration, Génétique et Physiologie Forestières), UR0588, Orléans, France
[4] INRA-EPGV (Etude du Polymorphisme des Génomes Végétaux), US1279 CEA-IG/CNG, Evry, France
[5] INRA-URGV (Unité de Recherche en Génomique Végétale) UMR1165, Evry, France

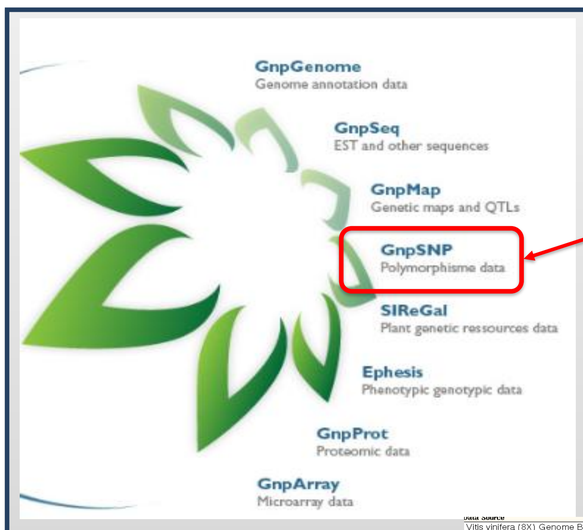
Introduction

Large SNPs discovery projects are undergoing in poplar and grapevine using the Illumina sequencing technology. In grapevine, 30 genotypes from different *Vitis* species are currently being resequenced and the reads obtained will be aligned along the grapevine reference genome sequence. In poplar, resequencing on *P. nigra* is divided in two steps. The first one is the deep resequencing of a few individuals (i) to construct a reference genome of the species and (ii) to identify SNPs. The second one consists of the resequencing of several genotypes at low coverage (2x) to maximize SNP discovery. Libraries and paired-ends sequencing (2x75bp and 2x100bp) on GAIIx were performed by EPGV group and CNG (Centre National de Génotypage) *Biological resources* and *Sequencing* platforms.

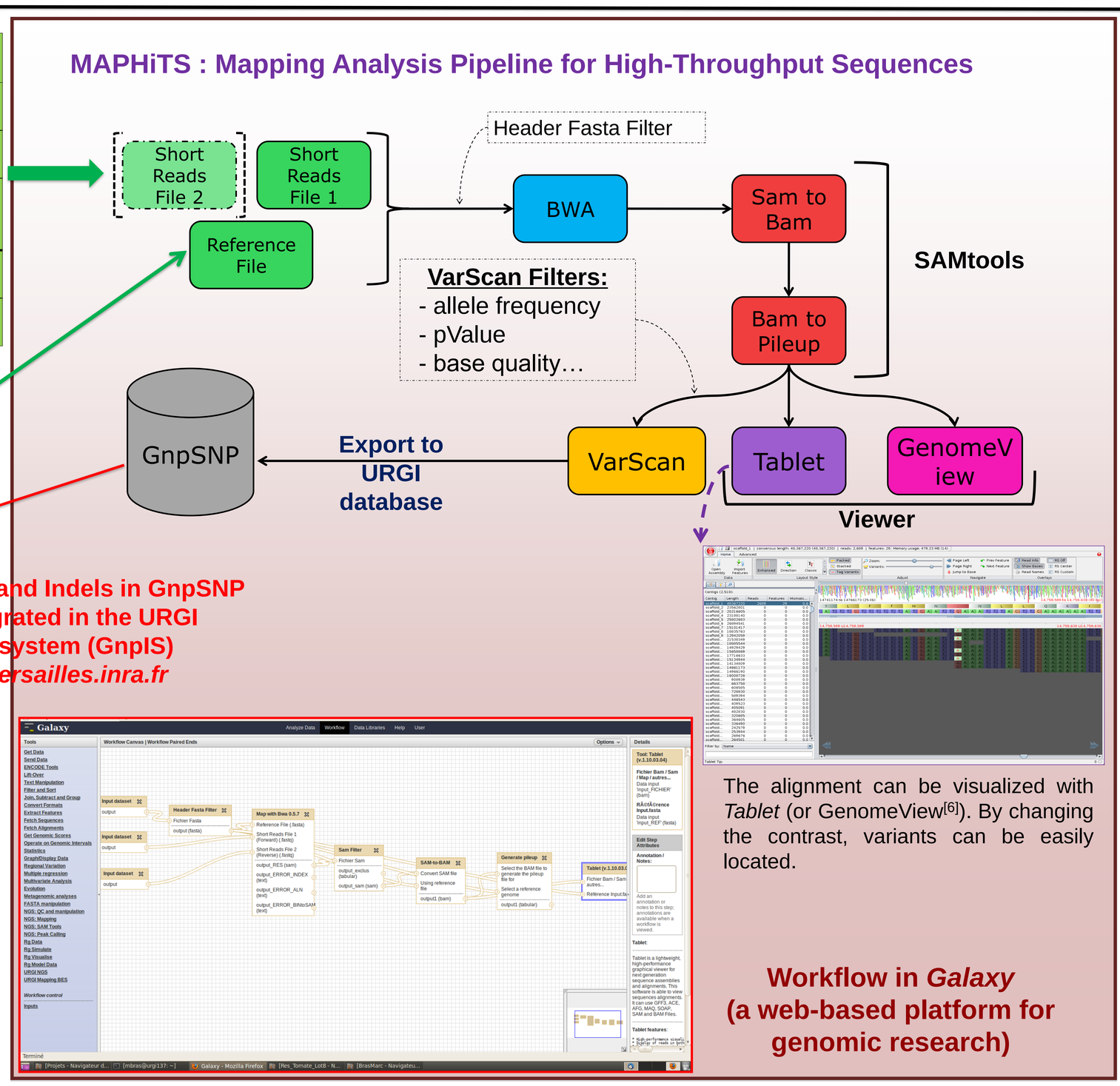
Sequencing data are being analysed using MAPHiTS (Mapping Analysis Pipeline for High-Throughput Sequences), a pipeline for SNPs detection developed by the URGI platform using the *Galaxy* workflow manager [1]. MAPHiTS is currently running with the following public tools *BWA* [2,3], *SAMtools* [3], *Tablet* [4] and *VarScan* [5]. MAPHiTS workflow is able to deliver all SNPs and small indels found in the data set and to filter them according to various parameters such as the genome coverage, the allele frequency and pValue.

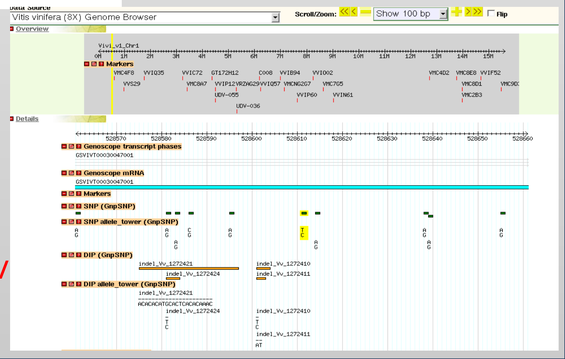
Species	Genotype	DNA	GAII Cycle number	Paired-end	Sizing (in pb)	Sort-reads number
Grape	*V. vinifera* PN40024	Total DNA	76	PE	400	34 533 968 x 2
	V. vinifera PN40024	Total DNA	114	PE	400	34 576 039 x 2
	V. vinifera Sultanine	Total DNA	76	PE	400	35 491 387 x 2
Poplar	*P. trichocarpa*	Total DNA	76	PE	600	21 625 677 x 2
	P. nigra	Nuclear DNA	76	PE	600	30 983 861 x 2

Reference genomes: *Vitis vinifera* PN40024 (12X version)
Populus trichocarpa V2



Insertion of SNPs and Indels in GnpSNP database integrated in the URGI Information system (GnpIS)
http://urgi.versailles.inra.fr





Visualization of SNPs and Indels on Gbrowse (GnpGenome)
http://urgi.versailles.inra.fr/index.php/urgi/Data/Genome/Run-GnpGenome

The alignment can be visualized with *Tablet* (or GenomeView[6]). By changing the contrast, variants can be easily located.

Workflow in *Galaxy* (a web-based platform for genomic research)

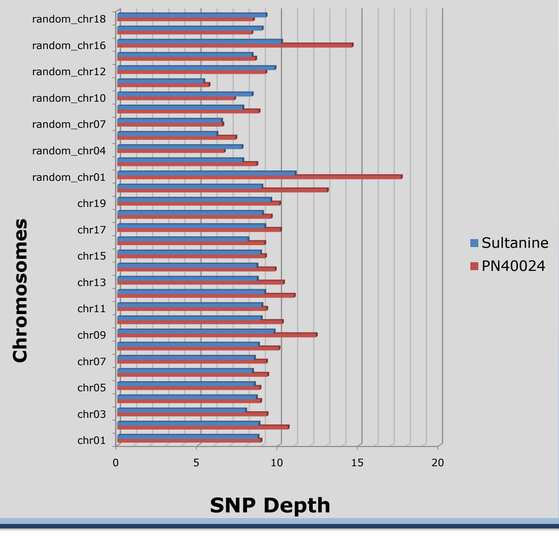
Preliminary results

Species	Genotype	Nb GAII Cycle	Nb short reads (SR)	Nb SH mapped	% SR mapped	Nb SNP	SNP depth min	SNP depth max	SNP depth average
Grape	PN40024	76	69 067 936	64 214 870	93	35 734	5,66	17,61	9,59
	PN40024	114*	69 152 078	59 547 300	86	53 434	7,78	20,07	13,31
	Sultanine	76	70 982 774	56 261 710	79	874 313	5,35	11,03	8,57
Poplar	*P. trichocarpa*	76	43 251 354	37 840 958	87	195 500	6,23	8,66	7,10
	P. nigra	76	61 967 722	41 421 646	67	1 753 673	6,79	8,91	7,73

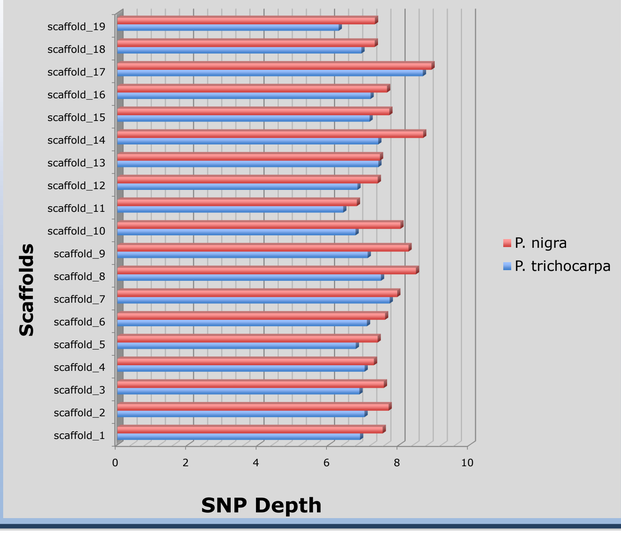
* For the grape PN40024 genotype, a GAII run with 114 cycles has been tested

SNP depth : [(Nb SR x Nucleotides) / chr length]

SNP depth for Grape *PN40024* and *Sultanine*



SNP depth for *P. trichocarpa* and *P. nigra*



MAPHiTS parameters used :
BWA- % identity: 96%
VarScan - Min Coverage: 10
VarScan - Min Reads variant: 4
VarScan - Min Variant Allele Frequency: 30%
VarScan - Min Base Quality: 30
VarScan - Min pValue: 1,00E-03

References

[1] J. Goecks et al (2010). 'Galaxy: a comprehensive approach for supporting accessible, reproductible, and transparent computational research in the life sciences'. Genome Biology 11, R86+
[2] H. Li and R. Durbin (2010). 'Fast and accurate long-read alignment with Burrows-Wheeler transform'. Bioinformatics. [PMID: 20080505]
[3] H. Li et al. (2009) 1000 Genome Project Data Processing Subgroup. 'The Sequence alignment/map (SAM) format and SAMtools'. Bioinformatics, 25, 2078-9. [PMID: 19505943]
[4] I. Milne et al. (2010). 'Tablet—next generation sequence assembly visualization'.Bioinformatics 26(3):401-402.
[5] Koboldt DC et al. (2009). 'VarScan: variant detection in massively parallel sequencing of individual and pooled samples'. Bioinformatics (Oxford, England), 25 (17), 2283-5 [PMID: 19542151]
[6] http://genomeview.sourceforge.net/

Acknowledgments : We thank IGA for helpful discussion and INRA – AIP Bioressources, the PLANT - KBBE2008 and Evoltree projects for the financial support.