

Development of a workflow for SNP detection with Galaxy



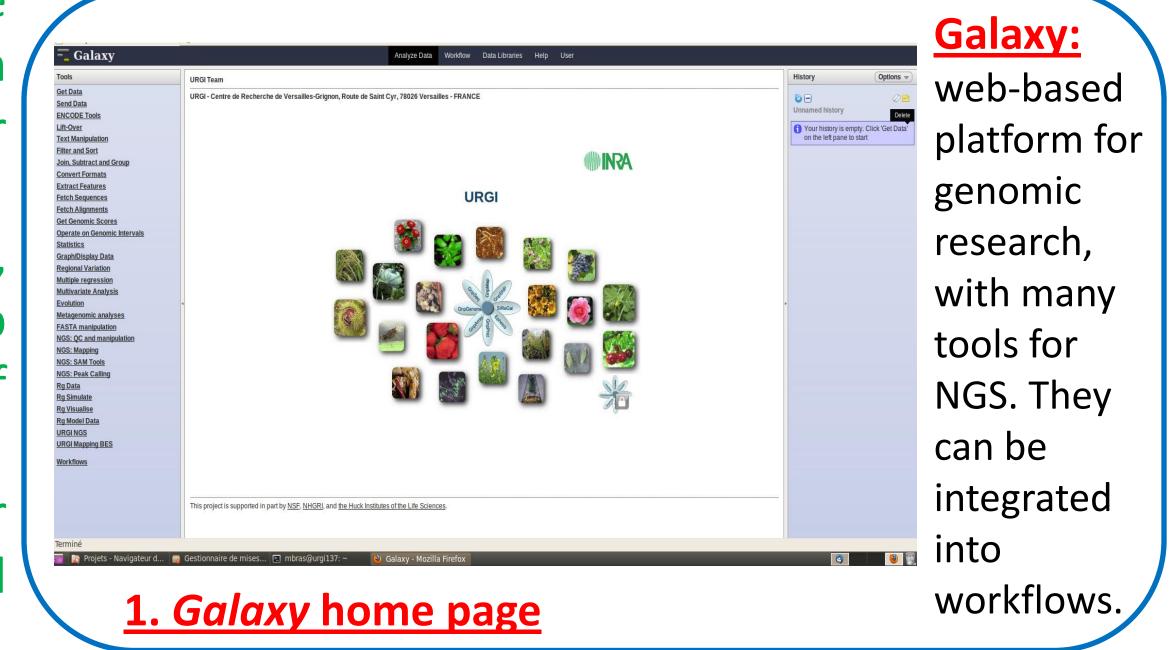
Marc Bras, Sandie Arnoux, Nacer Mohellibi, Nathalie Choisne, Olivier Inizan, Delphine Steinbach, Hadi Quesneville Unité de Recherche en Génomique-Info UR1164, INRA de Versailles-Grignon, Route de Saint Cyr, Versailles, 78026, France http://urgi.versailles.inra.fr urgi-contact@versailles.inra.fr

MAPHITS : Mapping Analysis Pipeline for High-Throughput Sequences

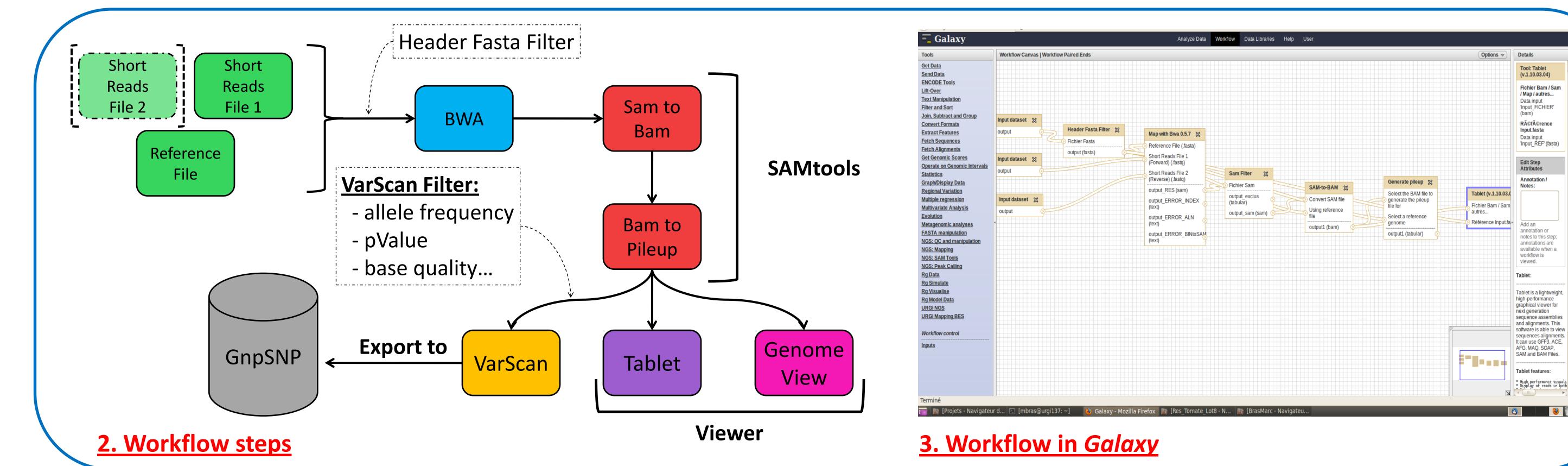
Introduction :

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.



Open Assembly	Import Features Data	Enhanced		Classic Layout Styl		 Packed Stacke Tag Va 	d	🔎 Zoor 🄐 Varia		Adjus	o	(je Right p to Basi	-		eature eature		Read Info Show Bas Read Nar	ses	RS	Off Center Cust or		
				Layout Styl][Aujus	эг. Т				140	avigate					ovenay	0			
Cantina (2.5						UNAN	HU, NW	N.F.F.	7.66TP	KIMBI		WWK.	MIN	Mara I		AH B	H REAL	a a leftal	14.6	A M		WW			MMK.
Contigs (2,5		Deede Fee	hu una a di an		եռնում	անևե	T = T = T	l de ra	ւրտ	э I.I.	Miles.	.441	labath (. K theo	m [ft	1614	1.4	1.6	uh d i	րթղ	shi sh	d.M.	ll Ibel	4 F I	ወቅፅብ
-	Length F 48367220	Reads Feat 2689	tures Misr 29	0.6 🔺	147413	74 to 14	766173 (25 Kb)	1	1.1	<u><u>v</u>.,</u>								1		14,75	59,599	to 14,	759,63	8 (40 bp
	23562801	2089	29	0.0	Y	L		F		F	M		Ν			N		L		L	Q		K		1
	20216605	Ō	0	0.0	A	ТТТ	G T	T	ГТ	ТТ	A T	G A	AT	TG	A A	A	TT	TA	C	ТТ	C A	AA	A	A A	ΤT
scaffold_4	23188140	0	0	0.0																					
	25802683	0	0	0.0																					
	26894541	0	0	0.0	14.759	599 U14,	759,599			I			1		1						I	14	.759.6	38 U1 4	,759,638
	15101417 18835763	0	0	0.0	A	T	,		Т	ТТ	AT	GA	AT	ТА	AA	A	TT	TA	C	ТТ	C A	AA	A	AA	TT
	12942059	0	0	0.0	A	ТТТ	GT	Т				-		TG	AA	A	ΤŤ	TA	C	T T	CA	AA	A	AA	ΤŤ
	21538349	ő	ŏ	0.0	A	ТТТ	GT	T	ГТ	ТТ	A T	GA	AT	ТА	A A	A	ТТ	TA	C	ТТ	C A	AA	A	A A	
	18885544	0	0	0.0	A	ТТТ	GT	T	ГТ	ТТ	A T	GA	AT	TG	A		TT	TA	C	ТТ	CA	AA	A	A A	TT
scaffold	14929429	0	0	0.0	A	ТТТ	GT	T	T	ТТ	A T	GA	AT	TG	AA	A	TT	TA	C	ТТ	CA	AA	A	AA	TT
	15658869	0	0	0.0	A		GI	+ -			AI	GA	A	A	AA	A		I A	C			AA	A	AA	
	17716633	0	0	0.0	A		C T	T -				C A	AT	TA	A A	A	TT	TA	C .	гт		AA	A	AA	TT
	15134944	0	0	0.0										A		A	TT	TA	Č.	ТТ	C A	AA	A	AA	TT
	14134809 14661173	0	0	0.0										Ğ	AA	A	τŤ	TA	C ·	тт	CA	AA	Â	A A	ΤŤ
	14966190	0	0	0.0																ТТ	C A	AA	A	A A	ТТ
	16008726	ŏ	Ő	0.0																					
scaffold	908939	ō	0	0.0																					
scaffold	663758	0	0	0.0																					
scaffold	606585	0	0	0.0																					
scaffold	728830	0	0	0.0																					
scaffold scaffold	589394 446543	0	0	0.0																					
scaffold	446543	0	0	0.0																					
scaffold	405091	0	0	0.0																					
scaffold	482830	õ	õ	0.0																					
scaffold	320685	0	0	0.0																					
scaffold	364605	0	0	0.0																					
scaffold	336490	0	0	0.0																					
scaffold	242579	0	0	0.0																					
scaffold scaffold	253944 269676	0	0	0.0																					
scaffold	264501	0	0	0.0																					
_			-		1																				
ilter by: 🚺	lame			•																					
																				_					
ablet Tip:																									0 <

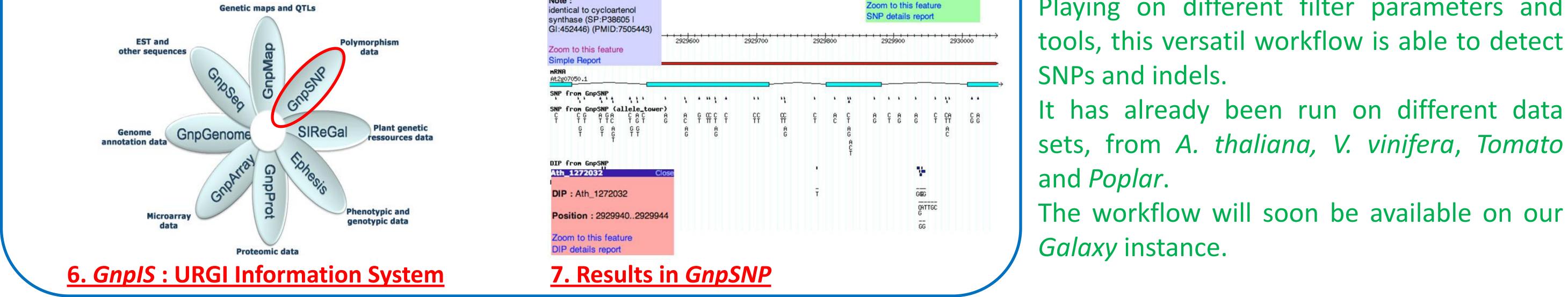
Results:

he alignment can be visualized with Tablet (or GenomeView^[6] to see nnotations). By changing the contrast, variants can be easily located (figure 4). */arScan* can predict SNPs nd indels. The output file contains information that an be used to filter them variant frequency, average pase quality, Pvalue...) figure 5).

	Chrom Position	Ref	Var	Readsl	Reads2	VarFre	q Strands	1	Strands2	Quall	Qual2 Pvalue		History	Options 🔻
	C10HBa0111D09_LR276	25	Т	С	3641	50	1,35%		1 53	39	2.515846373640937E-12			
	C10HBa0111D09_LR276	41	Α	Т	54	2	3,57%	2	1 49	39	0.49549549549543054		0 -	47 🖻
	C10HBa0111D09_LR276	139	Т	G	75	2	2,6%	2	1 54	35	0.49673202614372647		Unnamed history	· -
ools	C10HBa0111D09_LR276	142	Α	G	79	4	4,82%	2	1 54	34	0.12046848856664434		ofinamed history	
	C10HBa0111D09_LR276	155	Α	G	133	2	1,48%	2	1 54	44	0.4981412639408079		0. Drint CTDOUT	@ /) (
ulation.	C10HBa0111D09_LR276	162	Α	G	156	2	1,27%	2	1 55	39	0.4984126984129576		9: Print STDOUT	@ () S
lation	C10HBa0111D09_LR276	167	Т	G	181	4	2,16%	2	1 54	34	0.12297024877606934			
ort	C10HBa0111D09_LR276	174	С	Α	216	3	1,37%	2	1 55	38	0.2482837528605123		8: File after VarScan	۵0۵
ict and Group	C10HBa0111D09_LR276	213	Α	G	382	4	1,04%	2	1 54	38	0.12402786983967494		75,283 lines, format: ta	bular, database:
	C10HBa0111D09_LR276	221	Α	G	388	5	1,27%	2	1 55	44	0.061705346901980596		Info:	
mats	C10HBa0111D09_LR276	232	G	Т	366	16	4,19%	2	2 55	40	2.5996731837878082E-5		🗌 🔜 🕗	47 🛛
tures	C10HBa0111D09_LR276	233	Α	G	363	4	1,09%	2	2 55	39	0.061988753748865935			~ -
nces	C10HBa0111D09_LR276	233	Α	Т	363	9	2,42%	2	1 55	42	0.0037196533418433525		1 2	3 4 5
	C10HBa0111D09_LR276	236	Α	G	374	5	1,32%	2	2 54	39	0.06167601284258341		Chrom Posi	tion Ref Var Reads
nents	C10HBa0111D09_LR276	236	Α	Т	374	5	1,32%	2	2 54	38	0.06167601284258341		C10HBa0111D09_LR276 25	T C 3641
<u>c Scores</u>	C10HBa0111D09_LR276	240	G	Т	393	5	1,26%	2	1 55	38	0.06171532354636414		C10HBa0111D09_LR276 41	A T 54
Genomic Intervals	C10HBa0111D09_LR276	242	G	Т	378	4	1,05%	2	2 55	40	0.12401768384360315		C10HBa0111D09 LR276 139	T G 75
	C10HBa0111D09_LR276	243	G	Т	366	4	1,08%	2	2 55	48	0.12398580354487765		C10HBa0111D09_LR276 142	A G 79
Data	C10HBa0111D09_LR276	245	Α	G	355	8	2,2%	2	1 55	45	0.007513897101368674		C10HBa0111D09_LR276 155	A G 133
ay Data	C10HBa0111D09_LR276	249	Α	Т	338	4	1,17%	2	1 56	44	0.06195135480287766			
riation	C10HBa0111D09_LR276	264	G	Т	350	11	3,05%	2	2 55	38	9.039142638830804E-4			
ression	C10HBa0111D09_LR276	286	C	Α	351	10	2,77%	2	1 53	33	9.167905354483148E-4		-	
	C10HBa0111D09_LR276	325	Т	С	370	4	1,07%	2	1 56	39	0.12399665777194731			
Analysis	C10HBa0111D09_LR276	594	Α	Т	269	3	1,1%	2	1 55	43	0.24861878453047648			
•	C10HBa0111D09_LR276	596	G	T	275	4	1,43%	2	1 53	36	0.12365471395997031		•	
ic analyses	C10HBa0111D09_LR276	690	G	Т	293	9	2,98%	2	1 53	38	0.0036770927662112636			
	C10HBa0111D09_LR276	692	Α	Т	302	4	1,31%	2	1 54	34	0.12377351188801874			
ipulation	C10HBa0111D09_LR276	697	G	T	293	4	1,35%	2	1 53	40	0.12373631453265312			
<u>id manipulation</u>	C10HBa0111D09_LR276	698	A	T	290	4	1,36%	2	1 55	40	0.12372340890241723			
ing	C10HBa0111D09_LR276	701	G	T	283	6	2,08%	2	1 52	37	0.030441844418449254			
Tools	C10HBa0111D09_LR276	704	C	T	285	4	1,38%	2	1 55	35	0.12370130359432482			
	C10HBa0111D09_LR276	724	G		307	9		2	1 53	40	0.003687089954346247			
Calling	C10HBa0111D09_LR276	802	G	A	529	78	12,85%		1 55	56	4.70168251572379E-25			
	C10HBa0111D09_LR276	803	G	A	537	78	12,68%		1 55	56	4.877768202769912E-25			
	C10HBa0111D09_LR276	815	G	A	569	81	12,46%		1 55	57	5.793166760259983E-26			
	C10HBa0111D09_LR276	926	A	C T	459	5 7	1,08%		2 55	54	0.06182687541905745			
<u>e</u>	C10HBa0111D09_LR276	933 965	G G	T	521 444	7	1,33% 1,55%	2 2	2 55 2 56	53 51	0.015315456460637233 0.01526284451707899			
ata	C10HBa0111D09_LR276 C10HBa0111D09_LR276	965 1033	C		444 382	6	1,55%	2	2 55	34	0.03064750764429131			
	C10HBa0111D09_LR276	1033	c	Å	382	4	1,03%	2	2 55 1 55	34 38	0.12403038341704047			
ng BES	C10HBa0111D09_LR276	1035	c	Å	383	4	1,03%	2	1 55	39	0.12399933512438642			
ing DEO	C10HBa0111D09_LR276	1340	т	Å	735	4 13	1,07%	2	2 55	49	2,3163559232770034E-4			
	C10HBa0111D09_LR276	1444	ċ	Â	597	8	1,74%	2	1 54	50	0.003816343880444348			
	C10HBa0111D09_LR276	1568	т	Â	772	9	1,15%	2	2 55	41	0.0038167689589131094			
	C10HBa0111D09_LR276	1874	Ť	ĉ	681	9	1,3%	2	2 55	41	0.003805049641643428			
	C10HBa0111D09 LR276	1896	Ť	G	709	17	2,34%	2	2 55	37	1.3879502168617792E-5			
	C10HBa0111D09_LR276	1898	Ġ	Ť	682	23	3,26%	2	1 52	37	9.933821857197184E-8			
	C10HBa0111D09_LR276	1899	G	Å		17	2,47%		1 52	35	1.3808210956276015E-5			
	C10HBa0111D09_LR276	1899	G	Ť		13	1,9%		2 52	36	2.3050787455932537E-4			
	C10HBa0111D09_LR276	1907	Ť	ċ	719	9	1,24%		2 54	38	0.0038102973423658347			
	C10HBa0111D09 LR276	1908	Ġ	č	702	32	4,36%		1 53	37	3.296881523709554E-10	~		
	5101000111000_L12/0	1000	~	~	/ 02		-1,000	-	1 55					
					_									
ojets - Navigateur d 🗈	[mbras@urgi137: ~]	🛛 🕹 Gal	axy - M	ozilla Firef	ox 🔝	[Res_Tor	mate_Lot8	3 - N	📷 [BrasMarc -	Navigate	u		3	(

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.



At2g07050 Close	e				
Position : 29244442930669					
Name(curated) :				Ath_1272022	Clos
3-epoxysqualene mutase 3- epoxysqualenecycloartenol cyclase / (S)-2 cycloartenol				SNP : Ath_1272	
synthase (CAS1) / 2				Position : 2929	8342929834
Note : identical to cycloartenol synthase (SP:P38605 I GI:452446) (PMID:7505443)				Zoom to this fea SNP details rep	
Zoom to this feature Simple Report	2929600	2929700	2929800	2929900	2930000

Conclusion :

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

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/