



HAL
open science

The grapevine reference genome sequence

Sébastien Aubourg, Nadia Bentahar, Marc Bras, Aurelie A. Canaguier,
Nathalie Choisne, Gabriele Di Gaspero, Marie-Christine Le Paslier, Michele
Morgante, Simone Scalabrin, Anne-Francoise Adam-Blondon

► **To cite this version:**

Sébastien Aubourg, Nadia Bentahar, Marc Bras, Aurelie A. Canaguier, Nathalie Choisne, et al.. The grapevine reference genome sequence. IX International Symposium on Grapevine Physiology and Biotechnology, Apr 2013, La Serena, Chile. hal-02811094

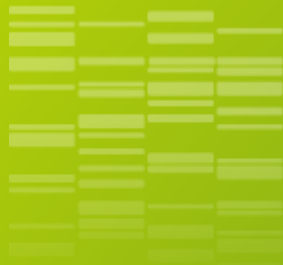
HAL Id: hal-02811094

<https://hal.inrae.fr/hal-02811094v1>

Submitted on 6 Jun 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



The grapevine reference genome

A-F Adam-Blondon
Unit of Research in Genomics Informatics



The Grapevine Reference Genome Assembly

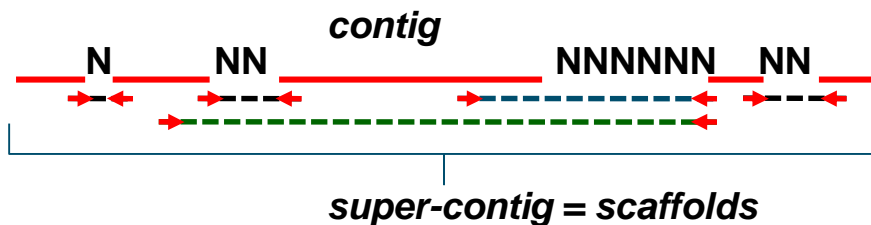


- Generation of scaffolds sequences
- Generation of the chromosome sequence

CURRENT STATUS: 12X.0 assembly

Library	Average insert length	Number of reads	Number of mate pairs	Coverage
BAC	100k	120091	57,606	0.12x
Fosmids	40k	251425	117,869	0.34x
Plasmids	10k	2848550	1,387,612	3.77x
Plasmids	3k	5518286	2,705,911	7.68x
Total		8738352	4,268,998	11.91x

Sanger shotgun sequencing approach



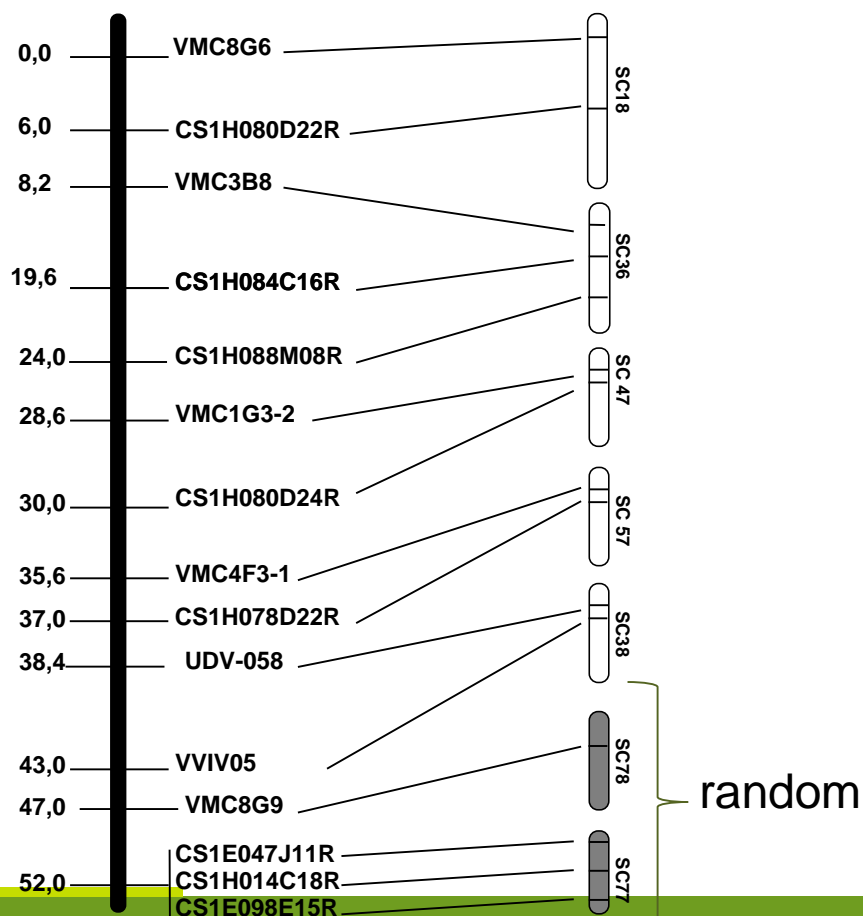
	8X	12X
Contigs Nb	19,578	17,196
Longest Ctg (kb)	557	774
Ctg N50 (kb)	65	100
Total size	467	481
Super-contig Nb	3,515	2,888
Longest sctg (kb)	12,675	13,101
Sctg N50 (kb)	2,065	3,300
Total size	487	499

Adam-Blondon et al (2011) *Genetics, Genomics and Breeding of Grapes*

Sequence built of the 19 chromosomes

Genetic map :
chromosome 12

Super-contigs of
sequence



12X.0 version (consensus map from 2
largers pop; **514** unique SSR markers):
- **85%** of the draft sequence is anchored on
the genetic map
- **71%** is anchored and oriented

Cipriani et al (2011) *Genetics, Genomics and Breeding of Grapes*.

Where can you find the sequence?

The screenshot shows the URGI website interface. At the top, there is a navigation bar with links for FEEDBACK, CONTACT, SITE MAP, and ABOUT US. The main header features the URGI logo and the text "PLANT AND FUNGI DATA INTEGRATION". Below this, a secondary navigation bar includes Platform, Research, Projects, Data, Tools, and Species. The Tools section is expanded, showing a circular menu with options like GnpSeq, GnpMap, GnpSnp, GnpGenome, SIReGal, GnpArray, Ephesis, and GnpProt. A central "GNPIS PORTAL" section is visible, along with a "SPECIES" section that lists various genomic tools and databases. A search bar is located on the right side of the page. The main content area on the left contains introductory text about URGI's research in genomics and bioinformatics. The right side of the page features a "WHAT'S NEW?" section with a recent update from March 2013 regarding Kegg Genes. At the bottom, there is an "EVENT & PUBLICATIONS" section with two entries from February 2013. A large blue cursor icon is overlaid on the "SPECIES" section, pointing towards the "Species" link in the navigation bar.

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

You are here : [Home](#) / [Home URGI](#) / [Species](#)

Species

You find here, the main species stored in the URGI repository

Vitis



Why develop grape genomics?

The International Grape Genome Program (IGGP, launched in 2001) a international efforts for the development of genomic resources for the years, more than 380 000 complete or partial sequences ...

Wheat

[Wheat URGI website](#)

Sequence built of the 19 chromosomes

Vitis	Genome sequences
<ul style="list-style-type: none">Grape OntologyData & Sequences<ul style="list-style-type: none">Genome sequencesBAC-ends sequencesAnnotationsGenome BrowserGrapeReSeq_Illumina_20KBlast & Blat serverMarkers and MapsGenetic resourcescDNA and BACs librariesLinksProjectsPartnersIGGPPublications	<ul style="list-style-type: none">12X "golden path" files (2009_12_04):<ul style="list-style-type: none">Scaffolds anchored on the 19 chromosomes : chr.agp (18.03 kB)The chromosome Unknown (scaffolds not anchored) : chrUn.agp (179.38 kB)Description of the chr.agp file : chr.agp.info (795 B)Length of the chromosomes : chr.lg (549 B)Scaffolds list (2059 scaffolds > 2 kb) : scaffolds.lg (38.53 kB)12X version of the PN40024 genome from The French-Italian Public Consortium :<ul style="list-style-type: none">WV 12X embl 102 WGS contigs : CAAP03000001-CAAP03014665 [14665 entries] at EMBL (release 102)WV 12X embl 102 Scaffolds : FN594950-FN597014 [2065 entries] at EMBL (release 102)WV chr12x (chromosomes) : FN597015-FN597047 [33 entries] at EMBL (release 102)8X version of the PN40024 genome from The French-Italian Public Consortium :<ul style="list-style-type: none">WV 8X embl 98 WGS contigs : CAAP02000001-CAAP02019577 [19577 entries] SUPPRESSED at EMBLWV 8X embl 98 Scaffolds : CU459218-CU462731 [3514 entries] SUPPRESSED at EMBLWV chr8x (chromosomes) : CU462738-CU462756 [19 entries] SUPPRESSED at EMBL<i>Vitis vinifera</i> WGS Pinot Noir clone ENTAV 115<ul style="list-style-type: none">WV clone entav 115 (contigs)

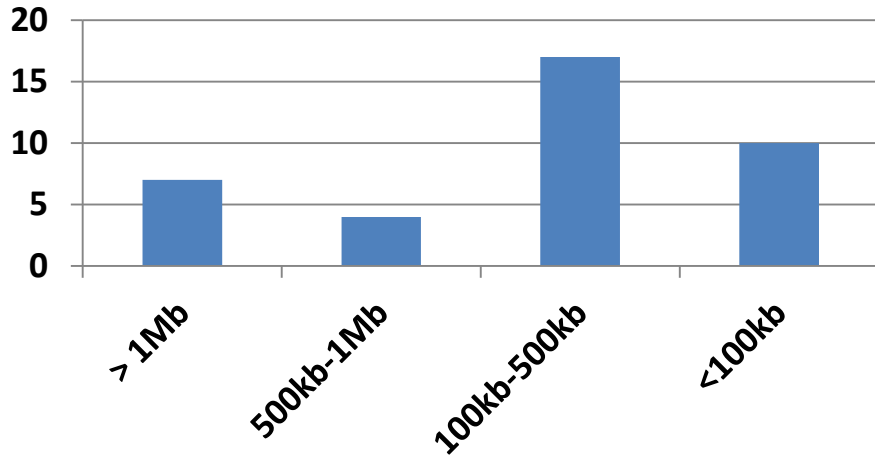
How it was built

The sequences

12X.0 genome assembly: characteristics of the « random » sequence

Anchored to a chromosome but not oriented : 38 sgtg; 16.8 Mb

n° super-contigs

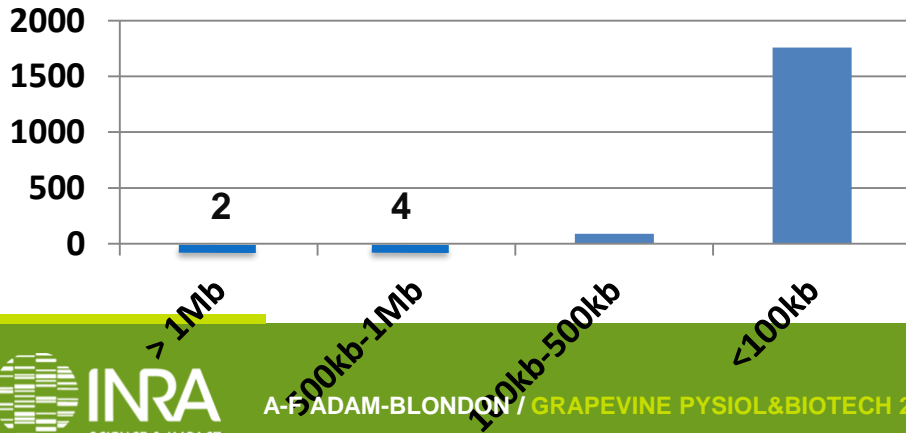


4.33% of 'N'

Anchored and oriented sequence : 2.33% of 'N'

Not anchored to a chromosome : 44 of them >200kb; 15.6 Mb

n° super-contigs

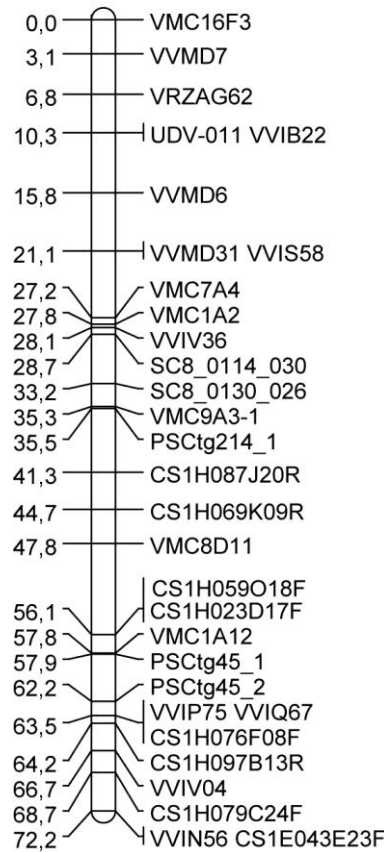
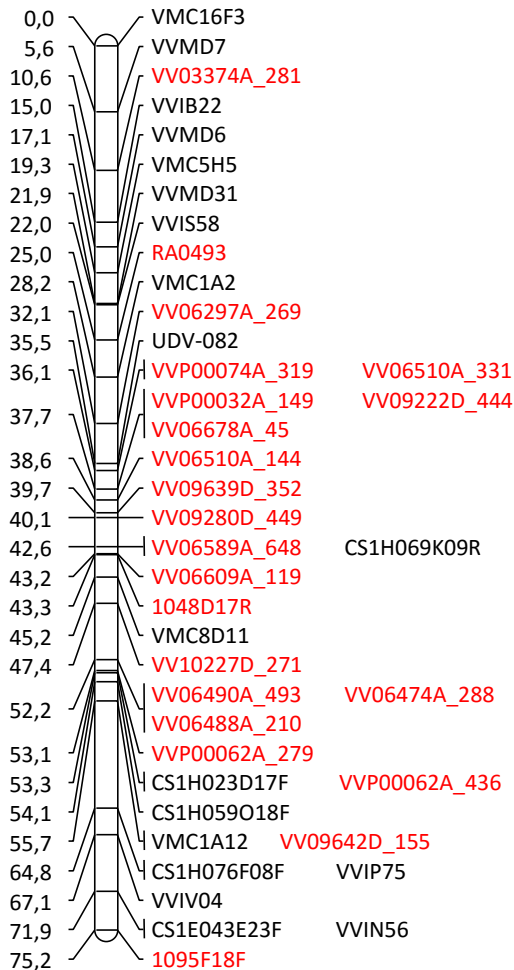


10.22% of 'N'

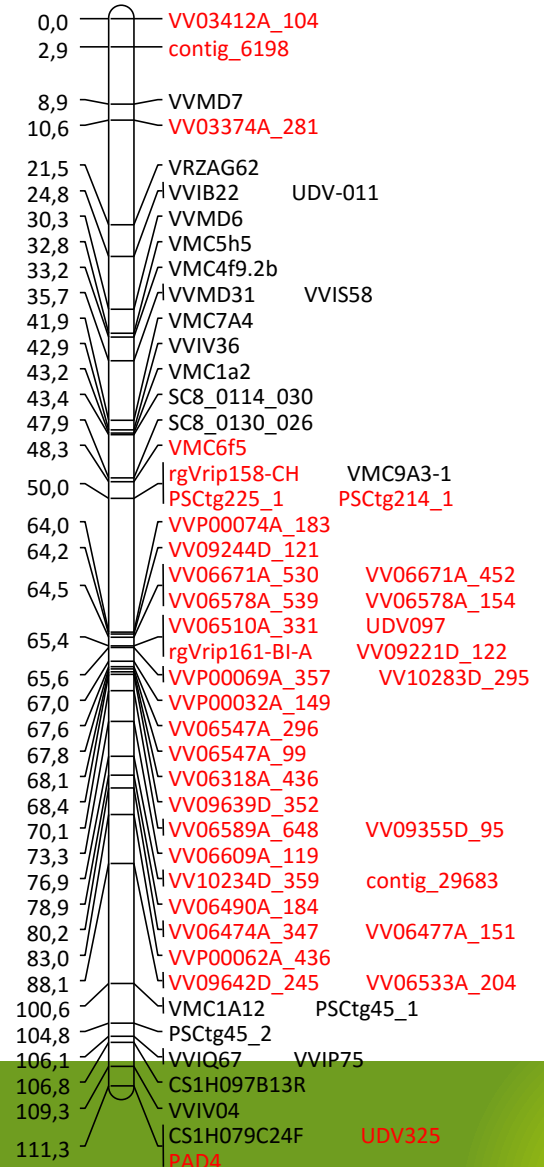
2064 gene models (7.8%)

Improvement of the genetic map with SNP markers developed from the scaffolds

LG7_SxG



Integrated SxG and ChxBI
(Cipriani et al 2011)

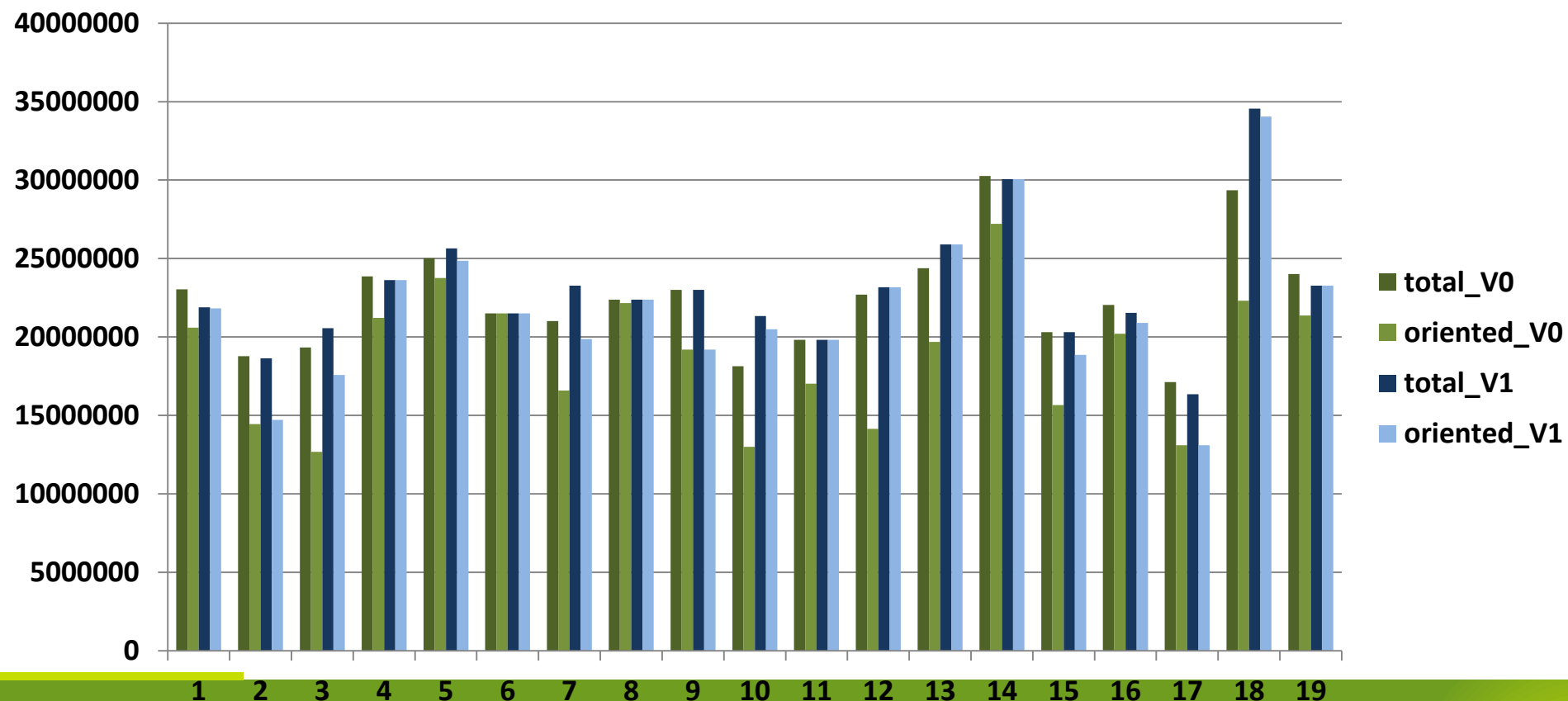


A. Canaguier

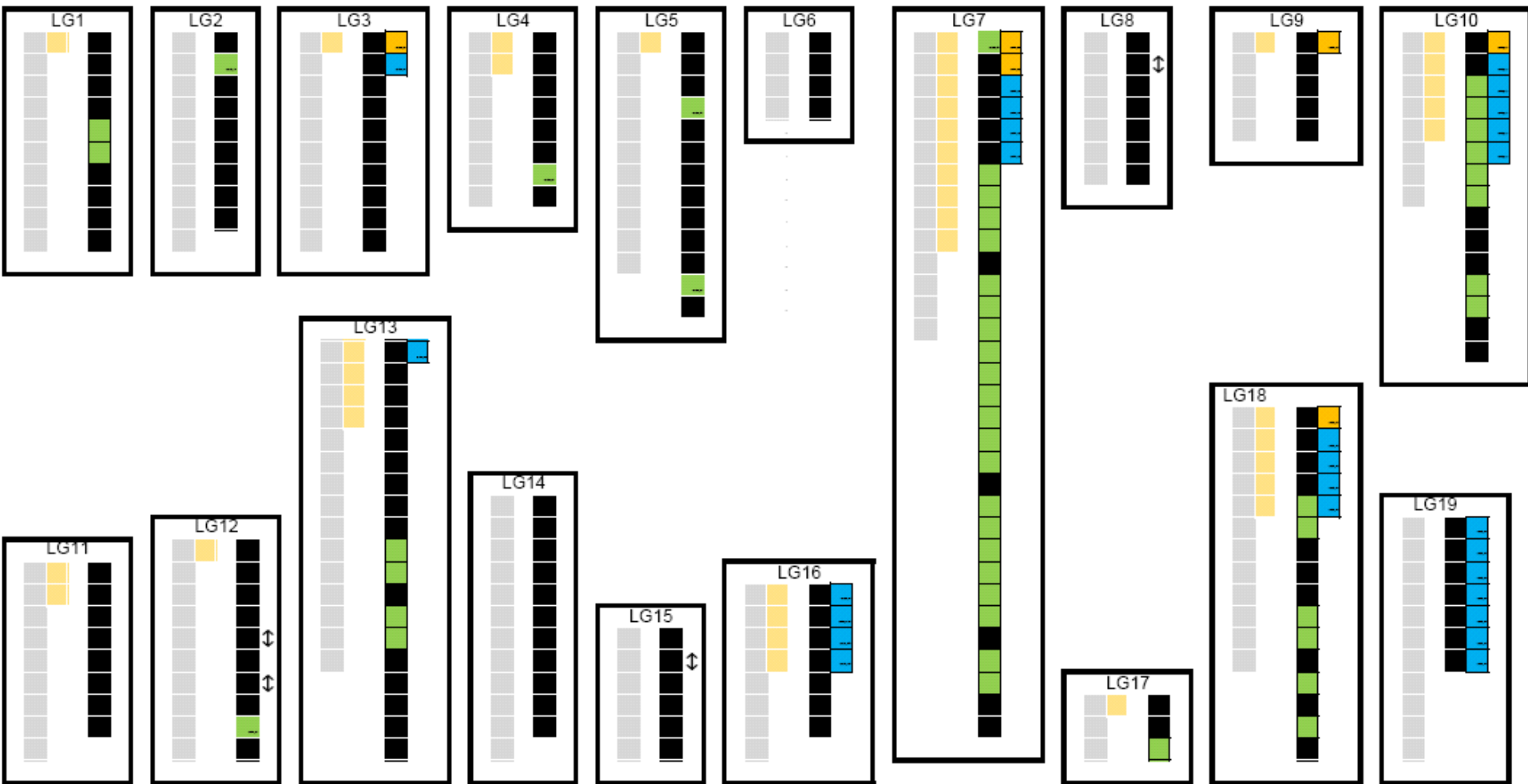
Anchoring the scaffolds on the new genetic map

V0: 426 Mb anchored (85% genome) -> V1: 485 Mb anchored (97% genome)

V0: 356 Mb oriented (71% genome) -> V1: 395 Mb oriented (79% genome)



A V1 version of the chromosome assembly



Scaffold anchored and oriented V0
Scaffold anchored and oriented

New scaffold anchored and oriented

Scaffold with ambiguous location V0
Scaffold with ambiguous location V1

New scaffold with ambiguous location

1 square = 1 scaffold that can be 200kb to >3Mb long

A 12X.1 chromosome assembly soon available

- The scaffold sequence does not change: same gaps, same errors
- The number, the order and the orientation of the scaffolds along the chromosomes changes

In the future: integrating changes in the scaffolds sequences based on re-sequencing data?

The Grapevine Reference Genome Annotation



- Automatic annotation
- Expert-based manual curation of the annotation

Automatic annotations

Input :

- 339,008 Vitis ESTs (NCBI)
- flcDNAs (99,828 reads from 5 libraries; URGV + Genoscope)
- Deep EST sequencing using Illumina-Solexa (175M reads from 4 libraries; IGA) and Roche-454

V0

V0 Annotation

Automatic annotation using the Gaze software (Genoscope)



Automatic annotation combining the gaze output and a Jigsaw output : **29 971 gene models**

V1

Used for the whole genome Nimbelgen transcriptome array (Univ. Verona)

Repeat annotation (IGA)

genomes.cribi.it/grape/

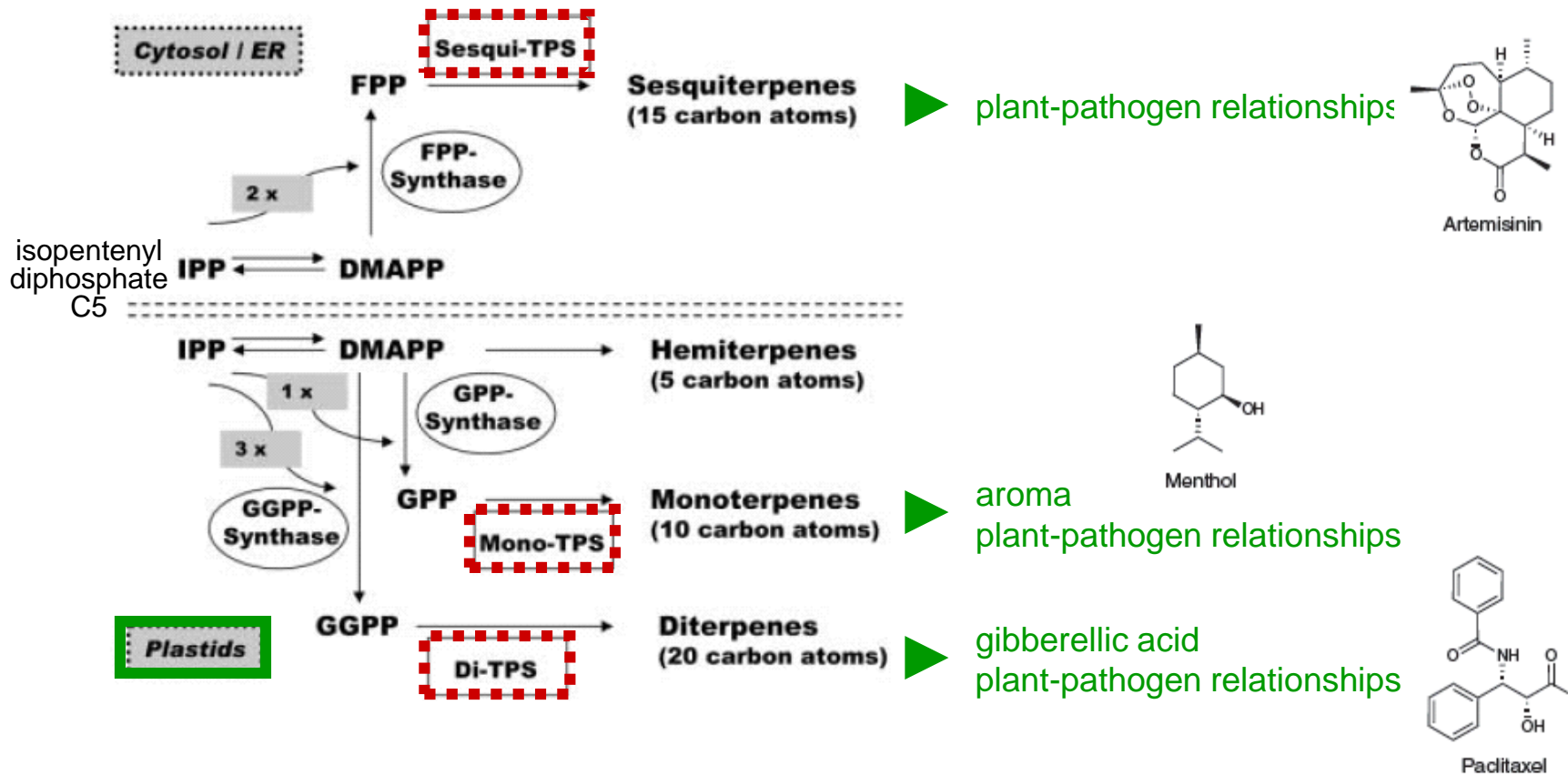
Automatic annotations

A gene model proposed by an automatic annotation must be expertized before any functional analysis based on it

For gene families, the whole family must be expertised before any start

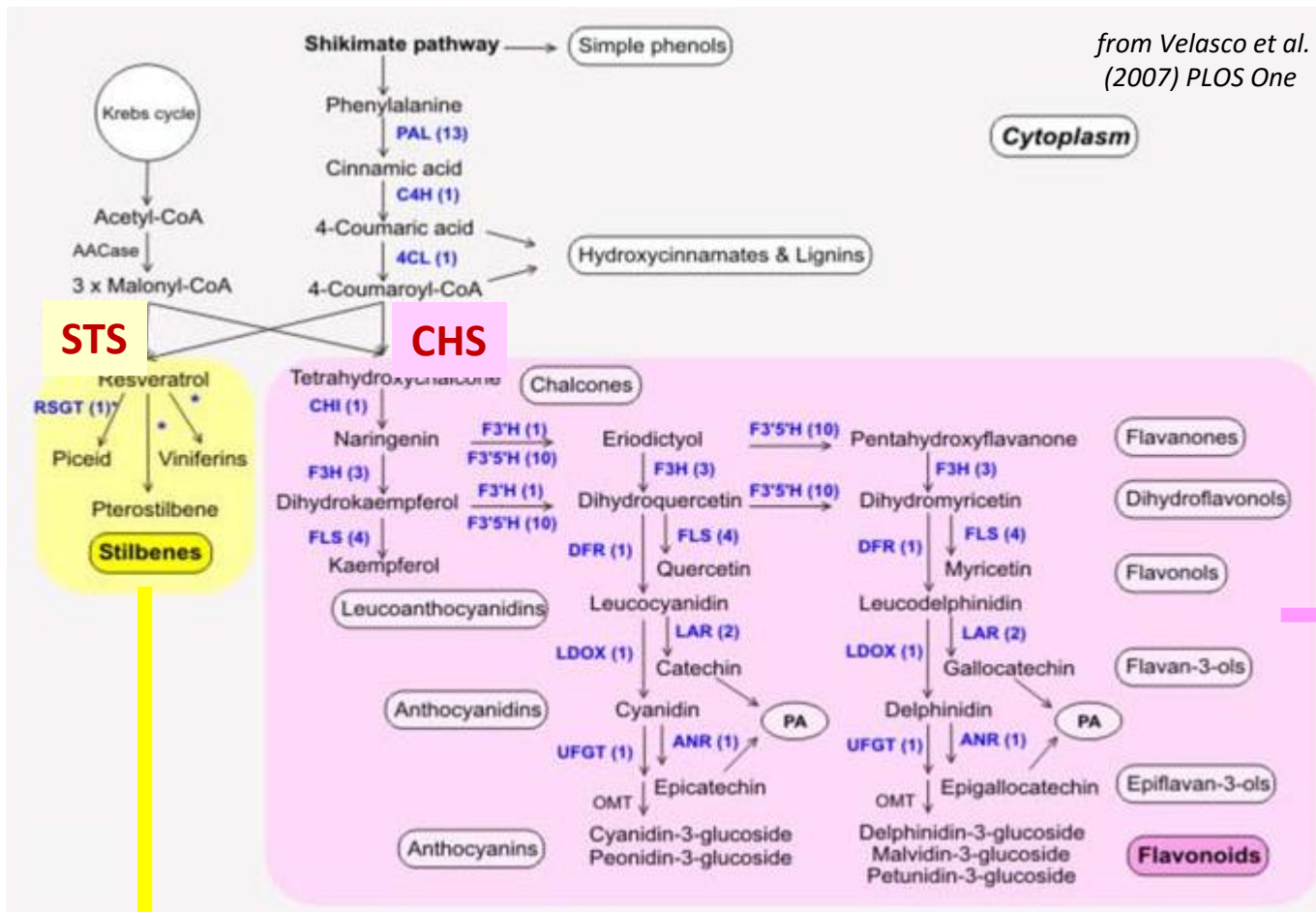
Curation of the annotation of gene families

Terpene synthase gene family



From Sébastien Aubourg and coll

Curation of the annotation of gene families



Chalcone Synthase and Stilbene synthase gene families

Colour and astringent features of wines

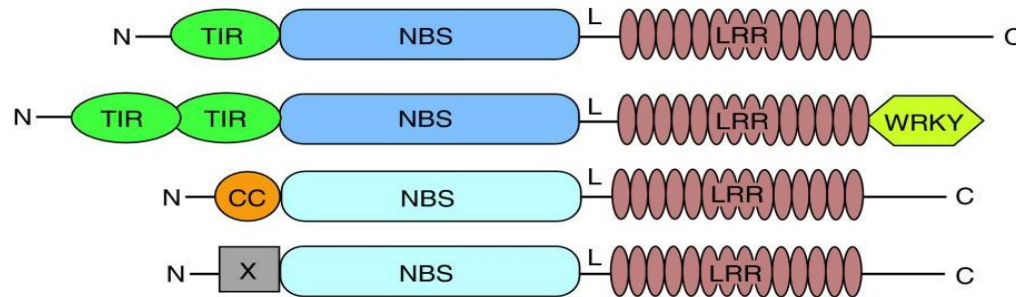
Resveratrol is involved in health benefits : 'French paradox'
 Prevents cancer and cardiovascular disease.
 Stilbenes are also known as fungicides.

From Sébastien Aubourg and coll

Curation of the annotation of gene families

NBS-LRR gene family

Resistance to diseases



NBS : Nucleotide-Binding Site

LRR : Leucine Rich Repeat

TIR : Toll/Interleukin-1 Receptor *

CC : coiled-coil *

* signal transduction domain

From Sébastien Aubourg and coll

Expert-based annotation process

Known TPS genes/proteins from Swiss-Prot (Arabidopsis, Pinus, Citrus...)

TBLASTN

Vitis 12x genome
PN40024

152 loci TPS-like

FLAGdb++



Artemis

- ✓ GAZE predictions
 - ✓ EUGENE predictions
 - ✓ PFAM motifs
 - ✓ EST
 - ✓ cDNA
 - ✓ BLASTX results
 - ✓ TPS knowledge
- From GB/EMBL & fIcDNA

Clean and exhaustive structural annotation of the PN40024 TPS gene family

- Detect and correct erroneous annotations and missing genes
- Discriminate between pseudogenes, partial and complete genes

Expert annotation is a necessity before starting any functional analysis on a gene family

Ann.vers	Missed genes/pseudogenes			Bad models for coding genes		
	TPS	STS&CHS	NBS-LRR	TPS	STS&CHS	NBS LRR
V0	35%	58%	44%	77%	77%	65%
V1	na	na	34%	na	na	68%
Eugene	8%	2%	6%	19%	36%	53%

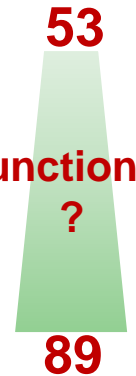


Manual curation of the TPS gene family

<http://urgv.evry.inra.fr/projects/FLAGdb++>

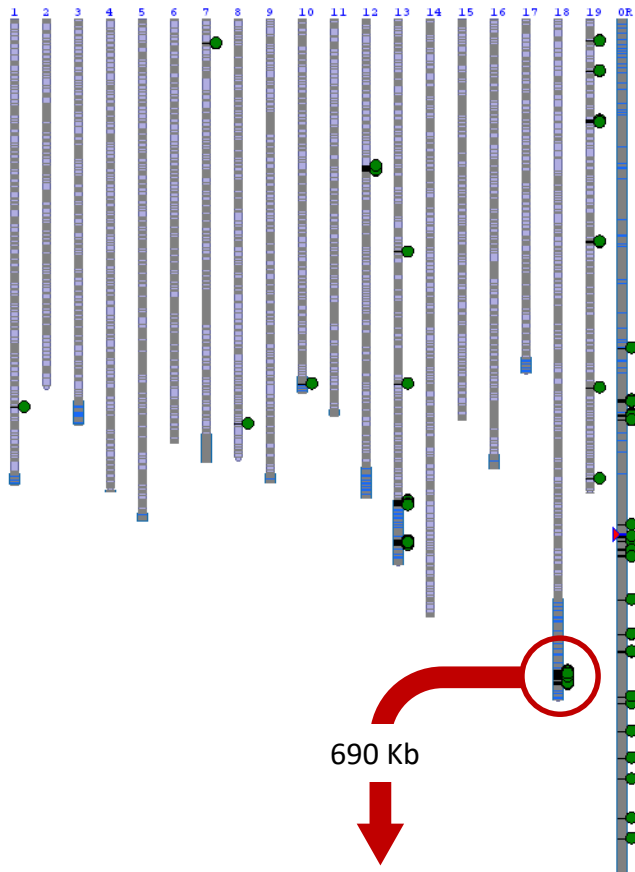
152 loci homologous to known TPS :

- **53 full** and perfect TPS genes
- **16 complete** TPS genes with only one punctual problem of sequence (sequencing error ?) **functional ?**
- **20 partial** TPS genes disrupted by unsequenced gap in the 12x assembly **89**
- **63 pseudogenes** with frameshifts, stop codons and/or large deletion(s)



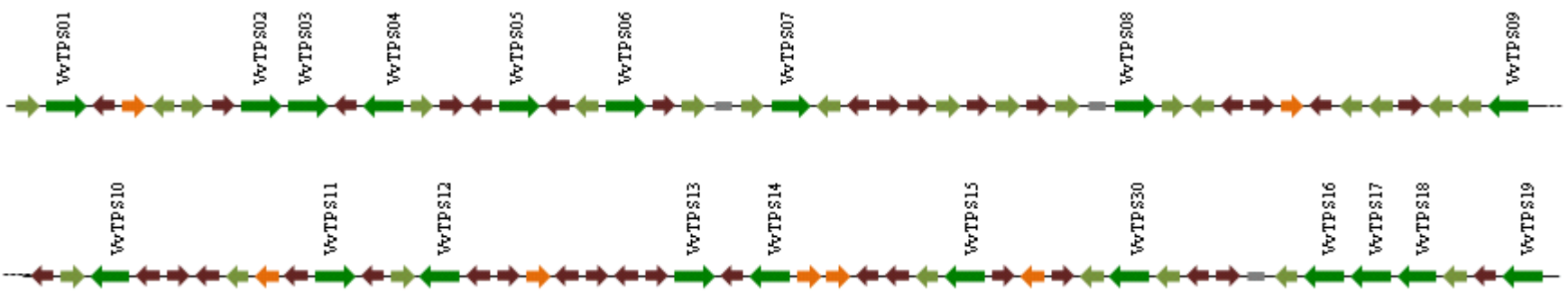
Martin et al 2010

Topological organisation of the TPS gene family



129 out of the 152 TPS loci (**85%**) are organized in tandem repeats in 13 clusters containing from 2 to 45 genes.

- 20 → full TPS genes
- 25 → pseudogenes TPS
- traces of transposons
- unsequenced regions
- other pseudogenes



Manual curation of the CHS and STS gene families

<http://urgv.evry.inra.fr/projects/FLAGdb++>

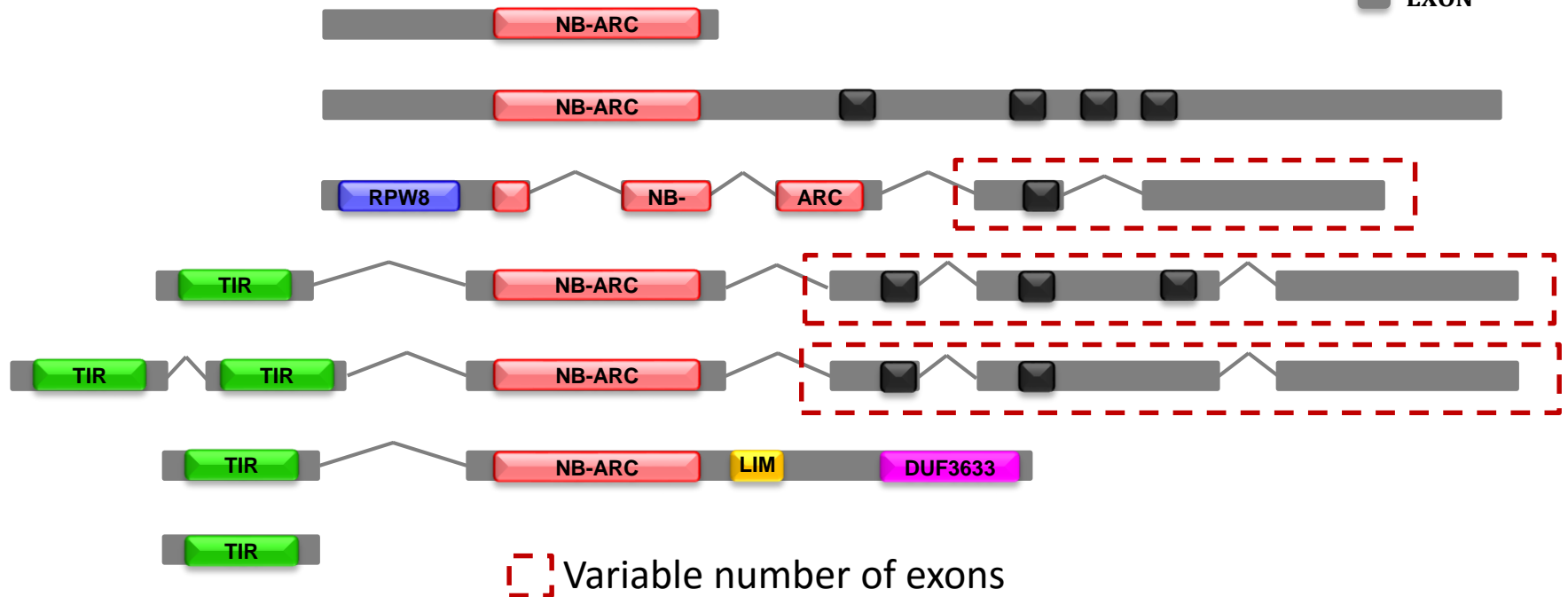
- 62 genes/pseudogenes of the CHS and stilbene synthase gene family (Parage et al 2012):
 - 3 CHS genes highly expressed and 11 CHS-like
 - 48 STS genes distributed on 2 clusters (chromosomes 10 and 16). At least 73% of them are expressed. Only 2 grape STS have been experimentally characterized. All the STS share between 91% and 99.7% of identity.

Manual curation of the NBS LRR gene family

<http://urgv.evry.inra.fr/projects/FLAGdb++>

829 genes/pseudogenes of the NSB-LRR gene family

- NB-ARC
- TIR
- RPW8
- LRR
- LIM
- DUF3633
- EXON



How can we gather / collect the results of manual curation to implement a V2 annotation?

Apollo (login required)



Apollo WebStart

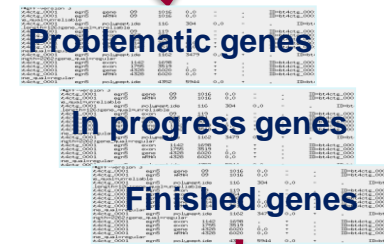
Write



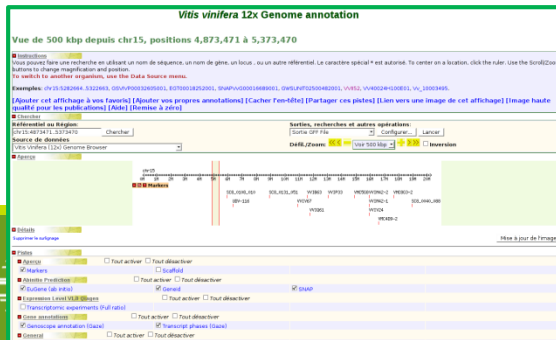
Read

Export Edited genes (GFF3)

Never used!
Too complex?
Need of time and dedication for coordination



Gbrowse update



Read



New complete release

Perspectives

- Automatic annotation of the 12X.1 assembly (COST action coord M. Bouzayen, M. Pezzotti)
- Include in the final version of the annotation the gene families that have been manually expertised

Sequence polymorphism in the Vitaceae family



Re-sequencing of panels of diverse genotypes

- Development of a high density and high throughput genotyping chip
- Assembly of new genomes

Re-sequencing a set of diverse genotypes in the Vitaceae family to develop a 20K genotyping chip

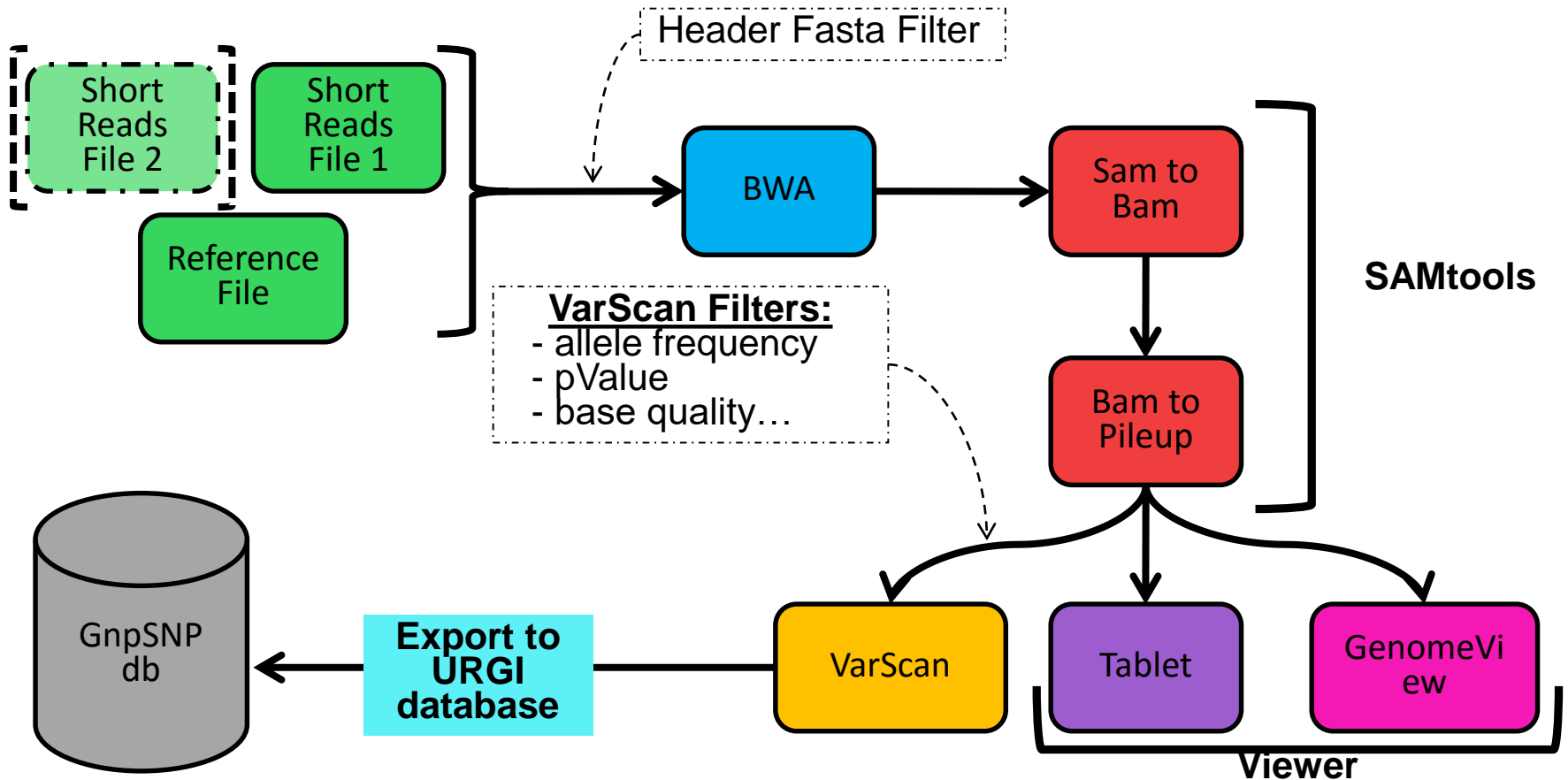
- EU KBBE GrapeReSeq project and of a french ANR project Muscares:
 - 20 diverse *Vitis vinifera*
 - 4 *Vitis sylvestris*
 - 3 *V. cinerea*
 - 3 *V. berlandieri*
 - 3 *V. aestivalis*
 - 3 *V. labrusca*
 - 6 *M. rotundifolia*
- Vigna italian project: 25 diverse *V. vinifera* varieties

Re-sequencing a set of diverse genotypes in the Vitaceae family

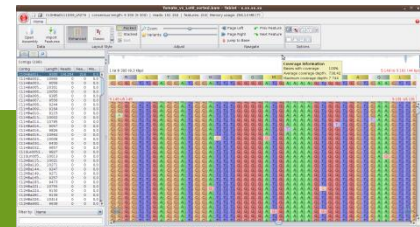
Read Length	Average	Standard dev	Min	Max
Read length (after trim)	94	6,38913	72	100
Bases after trimming	9 245 106 087	6 109 483 264	3 717 274 733	33 989 041 742
Genome coverage after trimming	19	12.6	8	70
Genome coverage with uniq reads	13	8.1	4	47
% of the genome covered	81%	5%	67%	93%
SNP detected	4 285 569	1 798 766	734 738	8 661 975
SNP after filters	926 880	336 957	162 806	1 545 045

- GA II and/or High Seq sequencing
- Filtering for position in known structural variants, in repeats, with too few or too many reads

Pipeline for SNP detection: MAPHITS Mapping Analysis Pipeline for High-Throughput Sequencing



Calculations parallelized



MAPHiTS workflow in Galaxy

Galaxy is a web based genomic platform (<http://usegalaxy.org>)

The screenshot displays the Galaxy web interface for a workflow editor. The main canvas shows a workflow with the following steps:

- Input Dataset** (two instances)
- Map with Bwa 0.5.7**: Takes Reference File (.fasta) and Short Reads File (.fastq) as input. 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 include output1 (bam).
- Generate pileup**: Takes output1 (bam) and Select a reference genome as input. Outputs include output1 (tabular).
- Tablet (v.1.10.03.04)**: Takes output1 (tabular) and Réference Input.fasta as input.

The right sidebar provides details for the **Tablet (v.1.10.03.04)** tool:

- Tool:** Tablet (v.1.10.03.04)
- Fichier Bam / Sam / Map / autres...**: Data input 'input_FICHER' (bam)
- Référence Input.fasta**: Data input 'input_REF' (fasta)
- Edit Step Attributes**: Annotation / Notes field.
- Tablet:** Tablet is a lightweight, high-performance graphical viewer for next generation sequence assemblies and 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 install.
 - Support for Windows, Apple Mac OS X, Linux and Solaris, i
- Website:** <http://bioinf.scri.ac.uk/tablet/>

Problem : management of big files, big calculations

18K GrapeReSeq Infinium genotyping chip

Details in poster n°157

http://urgi.versailles.inra.fr/Species/Vitis/GrapeReSeq_Illumina_20K

Generating genome assemblies for new genotypes

Sequencing of the Börner genome: Reinhardt Töpfer coming talk

Sequencing of the Norton genome (W. Qiu)

Sequencing of the Muscadine genome (D. Ware lab)

Difficulties to tackle:

- Development of methods that allow the assembly of complex genomes : improving the length of the reads, MATE paired reads, tuning the assemblers ...
- Calculation power and time

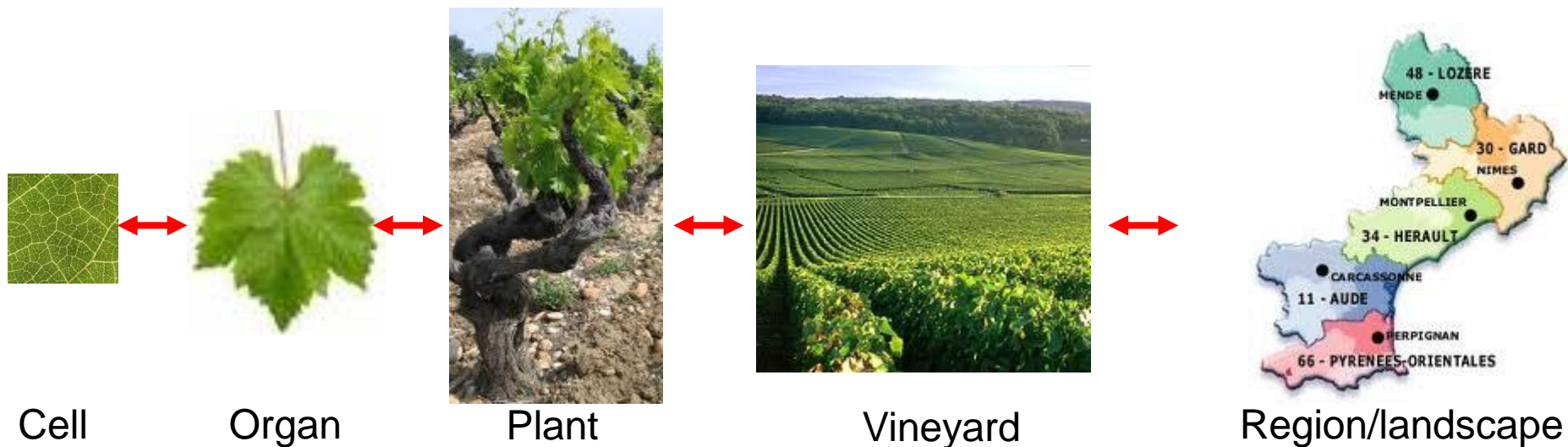


Conclusions

How are we using these resources to tackle the big questions of viticulture?

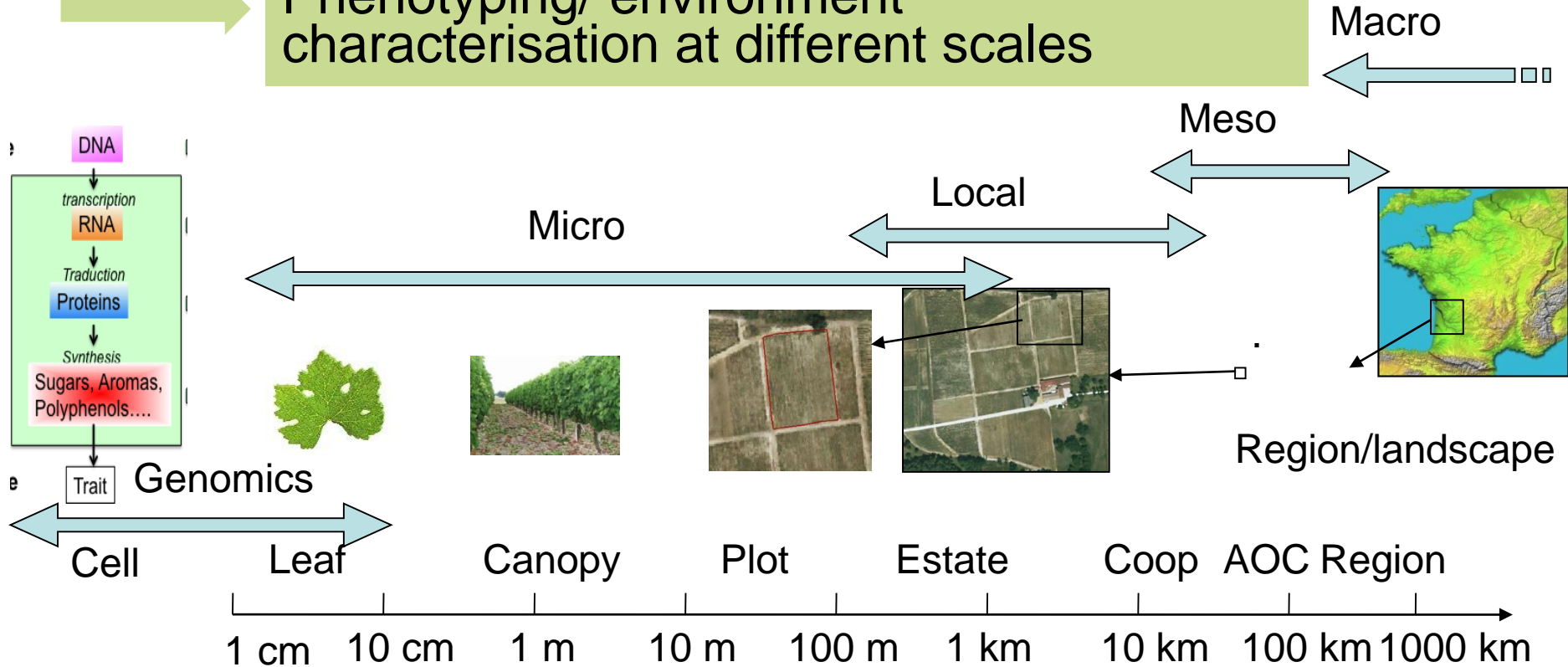
- How to predict a phenotype from the informations available on the genotype and its interactions with the environment?
- How will the existing varieties respond, and what are the new phenotypes we are looking for (ideotypes)?
- How do we develop and manage durable resistances in grape varieties?

Modeling at different scales



How are we using these resources to tackle the big questions of viticulture?

Phenotyping/ environment characterisation at different scales



Genomics to help understand the underlying processes and characterize the biotic environment

Two main efforts to be done for the future

Informatics, Bioinformatics and statistics

Phenotyping

Both need long term efforts and resources in some aspects

www.vitaceae.org

Acknowledgements



Predictive genomics team

S. Aubourg
F Samson
C Guichard
JP Tamby
N. Bentahar

Grapevine and Tree Genomic team

A. Canaguier

mis:ole C. Caron



P. Huguency
P. Mestre



N. Choisne
N. Mohelibi
M. Bras



M. Lepaslier
D. Brunel



J-M Boursiquot
V. Laucou



French-Italian Public Consortium for Vitis



Joerg Bohlmann
Diane Martin



M. Morgante
G. Di Gaspero
S. Scalabrin



L. Hausmann
R. Töpfer



J-M Martinez-Zapater



Silvia Vezzulli

