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

The grapevine reference genome

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

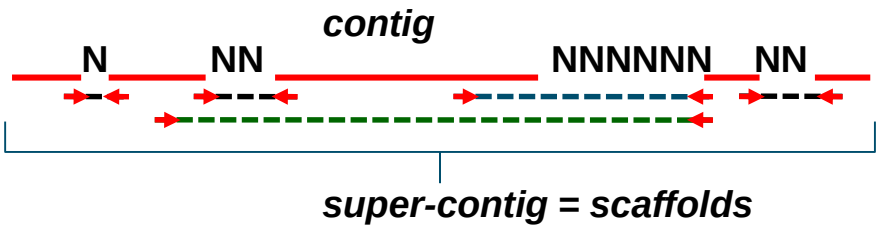
The Grapevine Reference Genome Assembly

- 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



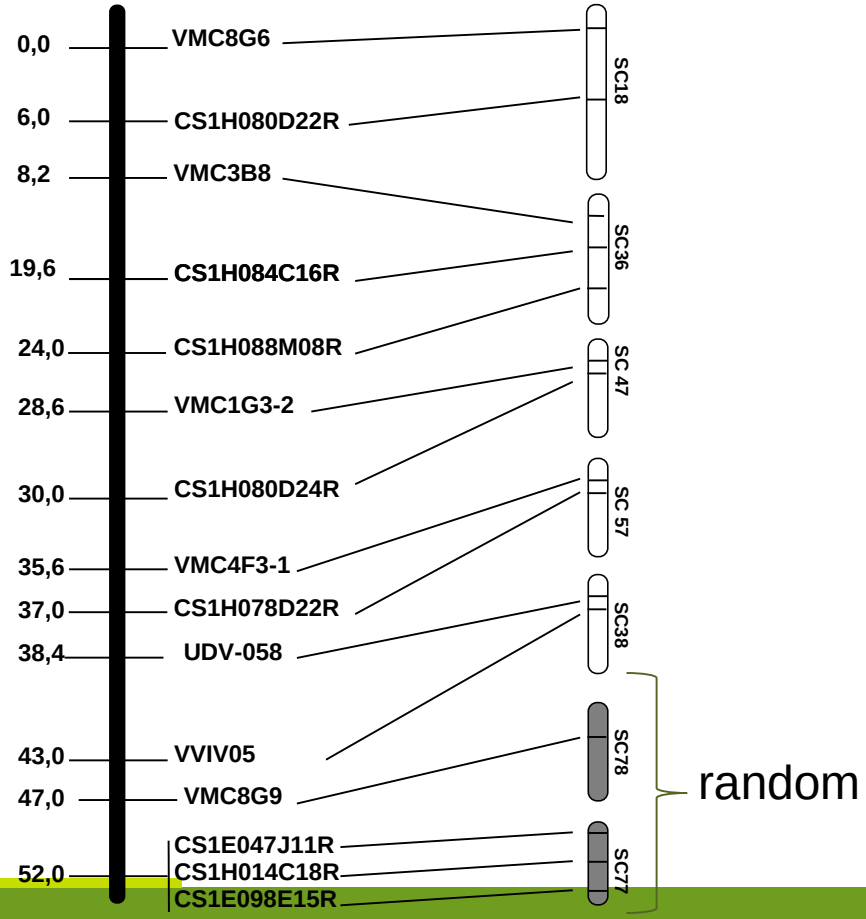
	8X	12X
Contigs Nb	19,578	17,196
Longest Ctg (kb)	557	774
Ctg N50 (kb)	65	100
Total size	467	481
Super-contig Nb	3,515	2,888
Longest sctg (kb)	12,675	13,101
Sctg N50 (kb)	2,065	3,300
Total size	487	499

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):

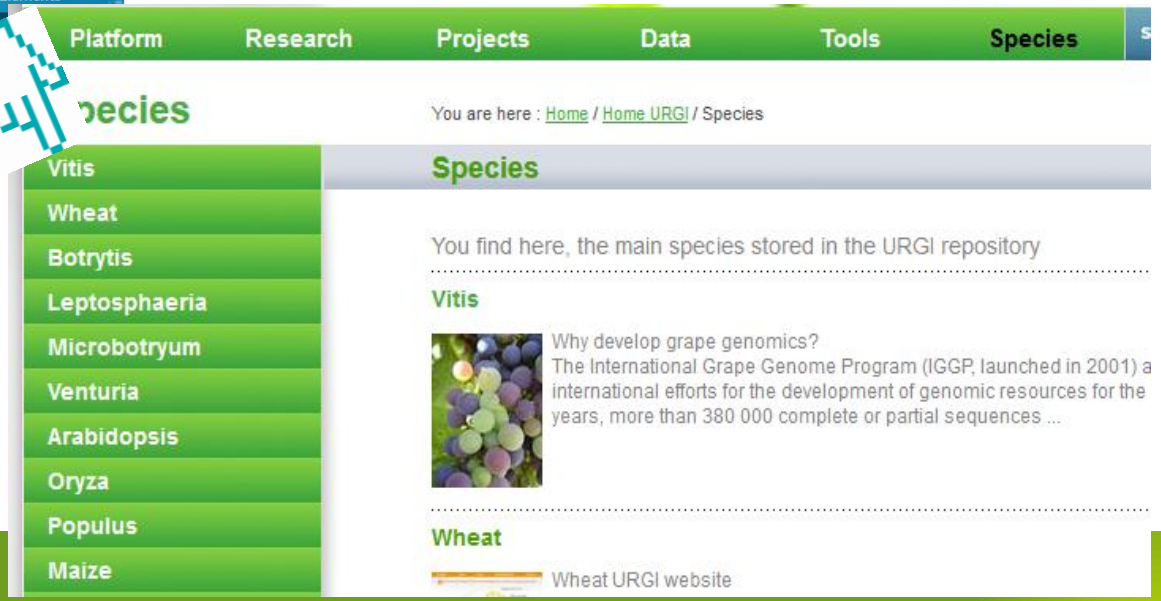
- 85% of the draft sequence is anchored on the genetic map

- 71% is anchored and oriented

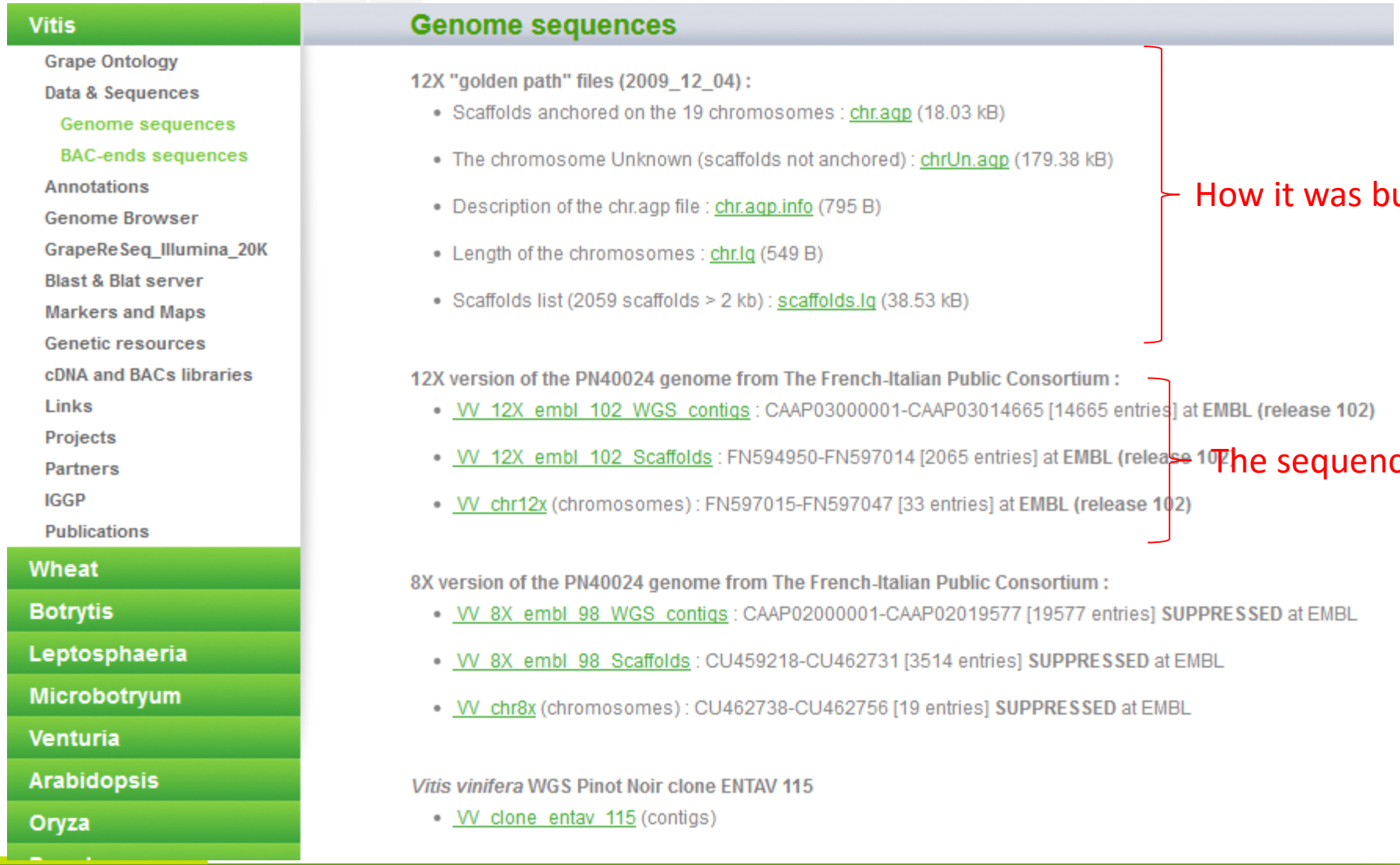
Cipriani et al (2011) *Genetics, Genomics and Breeding of Grapes.*

Where can you find the sequence?

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

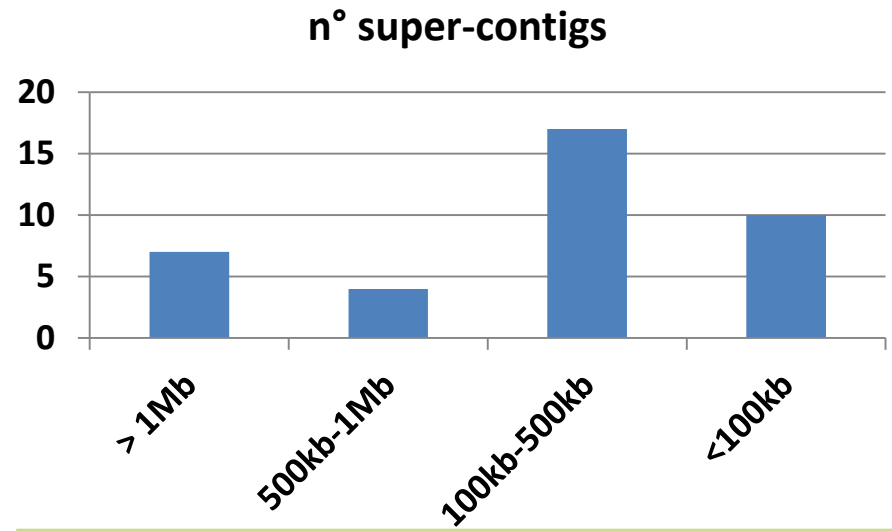


Sequence built of the 19 chromosomes



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

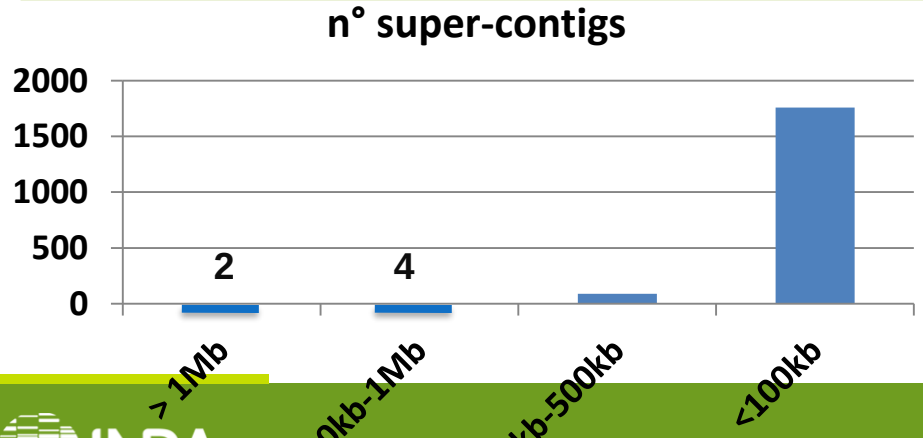
Anchored to a chromosome but not oriented : 38 sgtg; 16.8 Mb



4.33% of 'N'

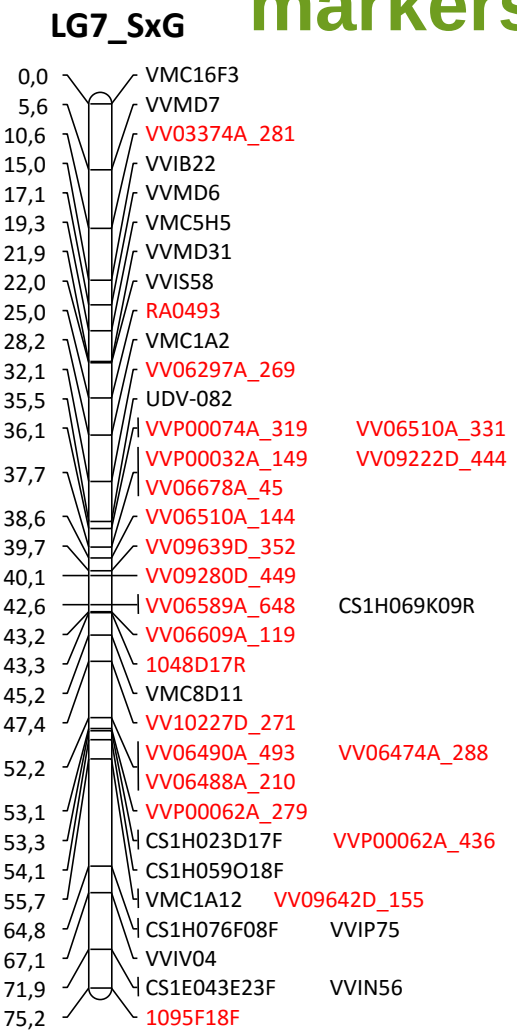
Anchored and oriented sequence : 2.33% of 'N'

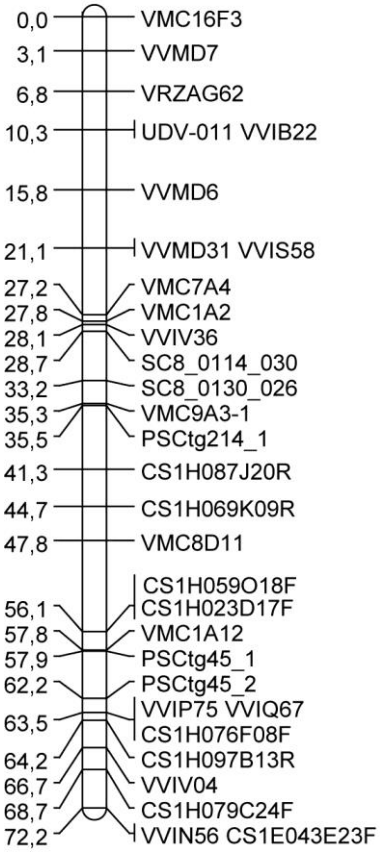
Not anchored to a chromosome : 44 of them >200kb; 15.6 Mb



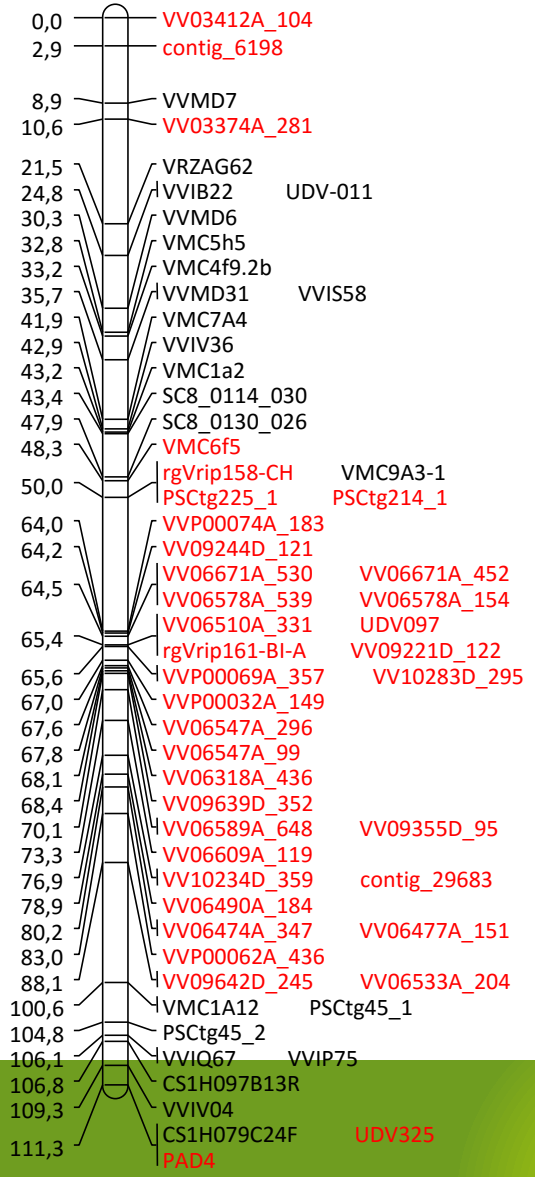
10.22% of 'N'
2064 gene models (7.8%)

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





Integrated SxG and ChxBI (Cipriani et al 2011)

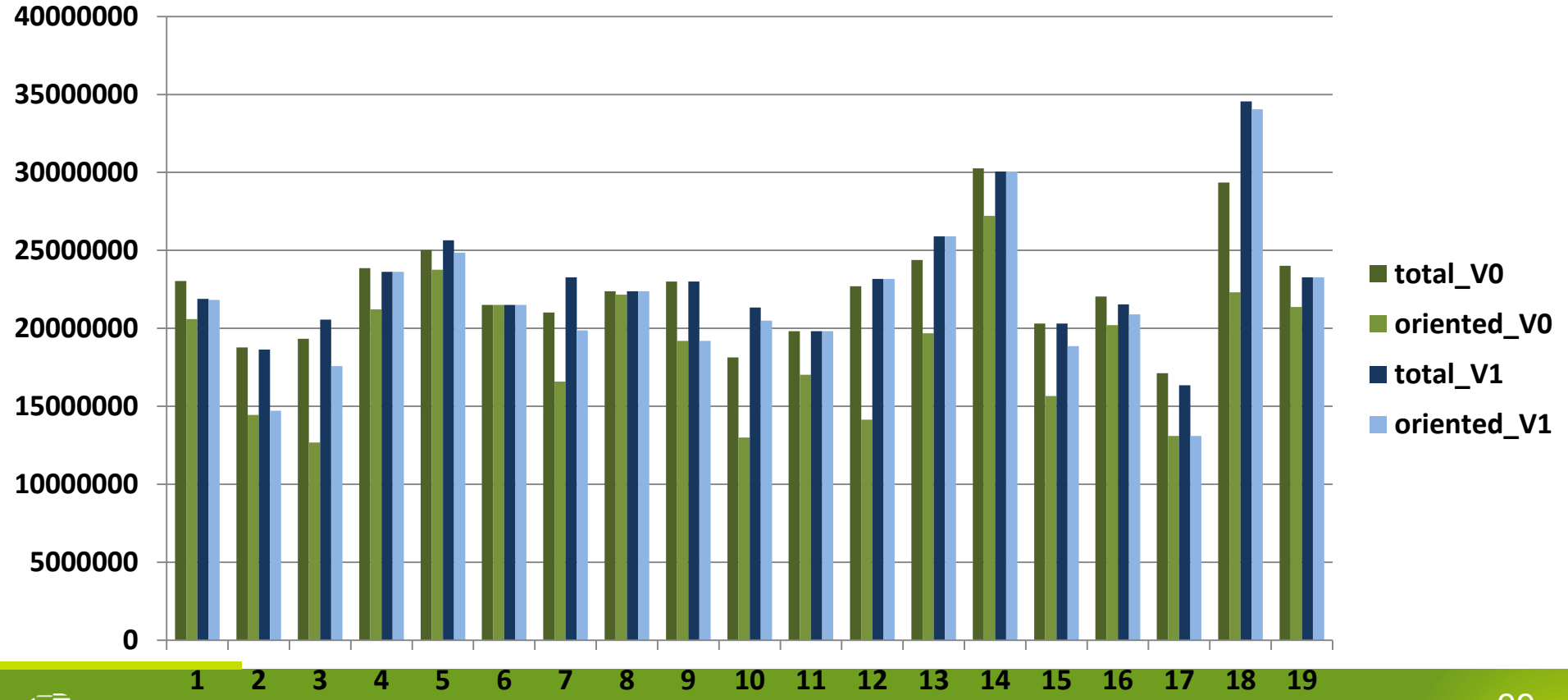


A. Canaguier

Anchoring the scaffolds on the new genetic map

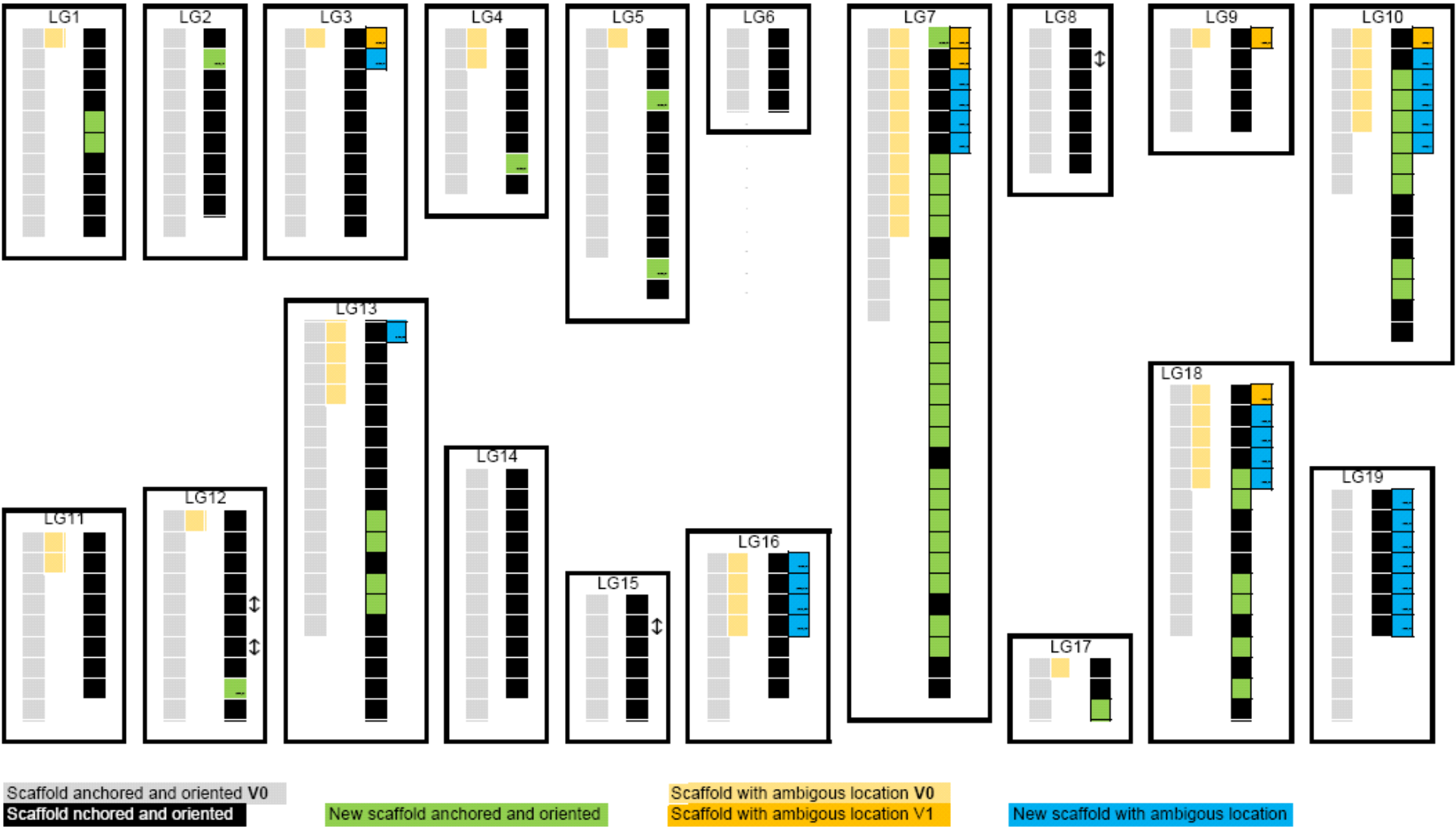
V0: 426 Mb anchored (85% genome) -> V1: 485 Mb anchored (97% genome)

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



A. Canaguier, G. DiGaspero, S. Scalabrin

A V1 version of the chromosome assembly



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

V0

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 Annotation

Automatic annotation using the Gaze software (Genoscope)

V1

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

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

Repeat annotation (IGA)

genomes.cribi.it/grape/

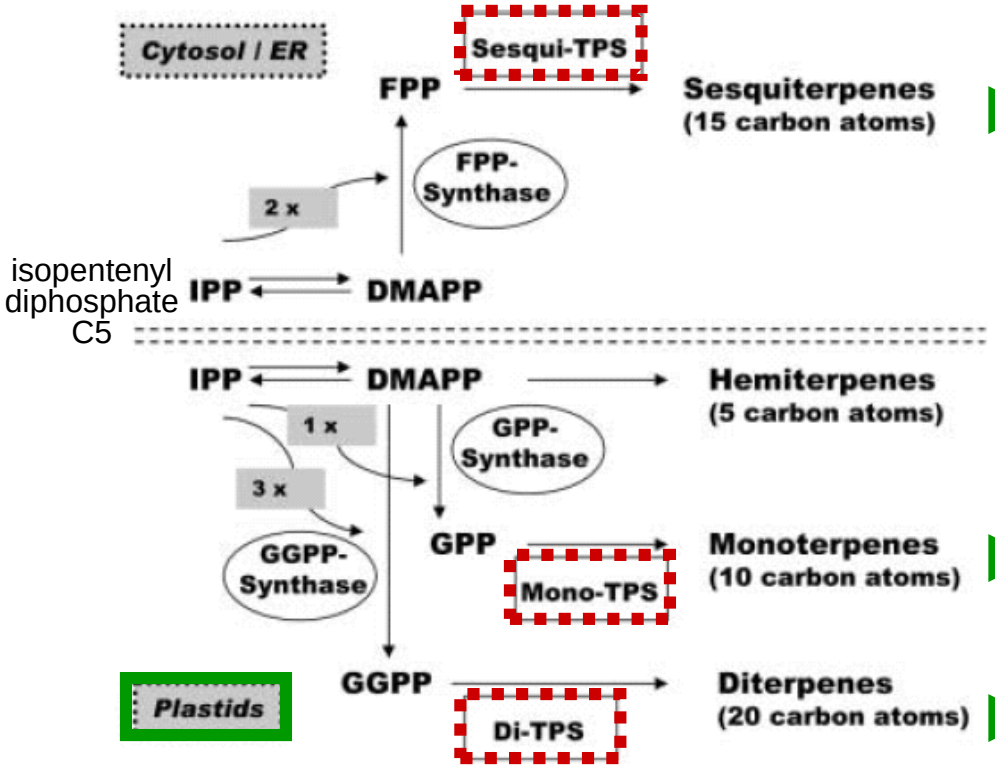
Automatic annotations

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

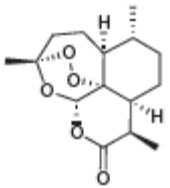
For gene families, the whole family must be expertised before any start

Curation of the annotation of gene families

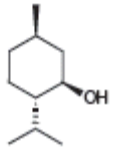
Terpene synthase gene family



plant-pathogen relationships



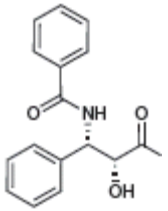
Artemisinin



Menthol

aroma
plant-pathogen relationships

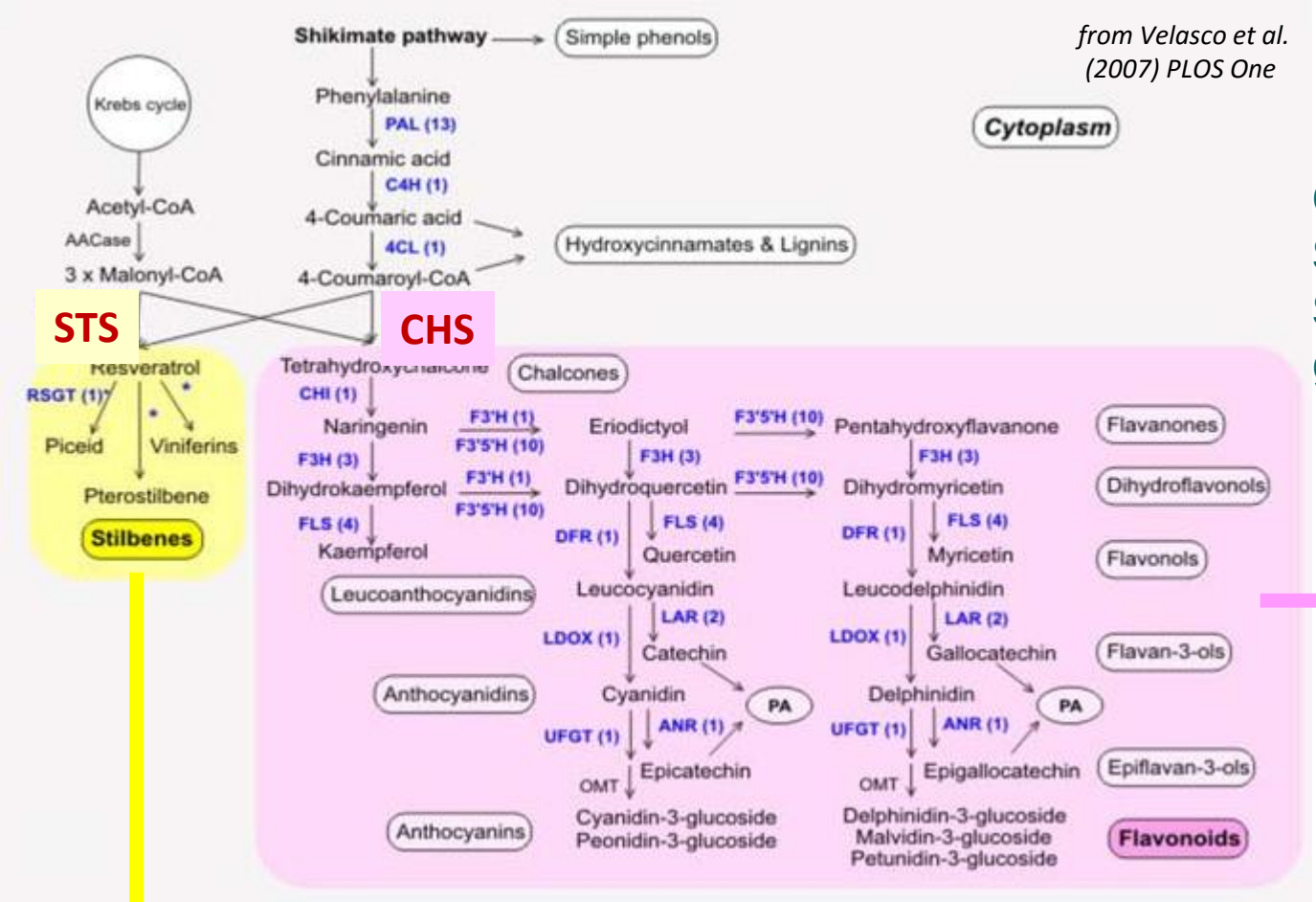
gibberellic acid
plant-pathogen relationships



Paclitaxel

From Sébastien Aubourg and coll

Curation of the annotation of gene families



from Velasco et al. (2007) PLOS One

Chalcone Synthase and Stilbene synthase gene familes

Colour and astringent features of wines

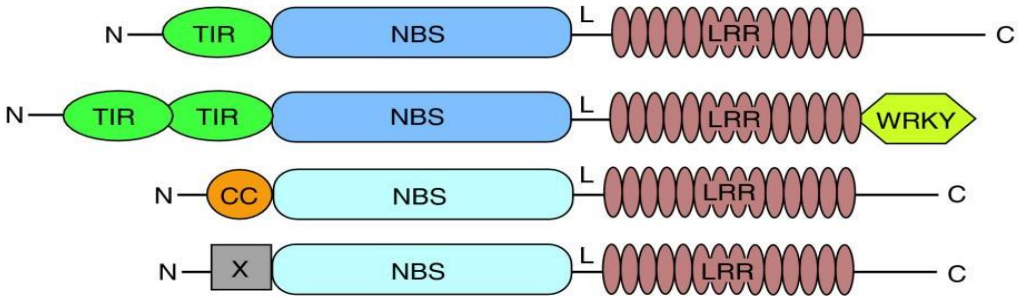
Resveratrol is involved in health benefits : 'French paradox'
Prevents cancer and cardiovascular disease.
Stilbenes are also known as fungicides.

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**

From GB/EMBL & flcDNA

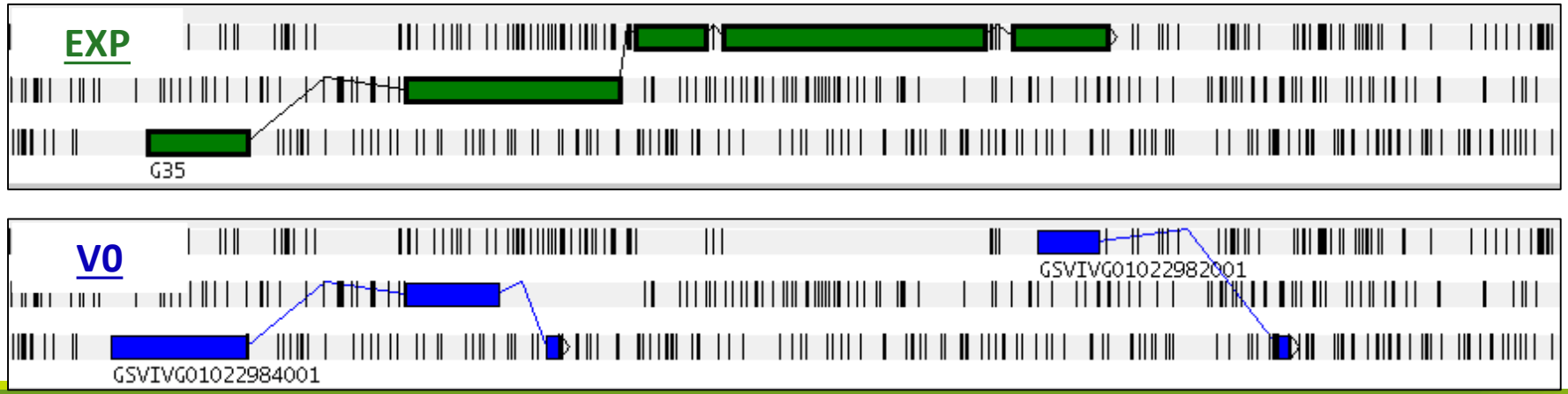
- ✓ **BLASTX results**
- ✓ **TPS knowledge**

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%

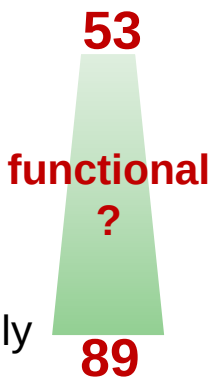


Manual curation of the TPS gene family

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

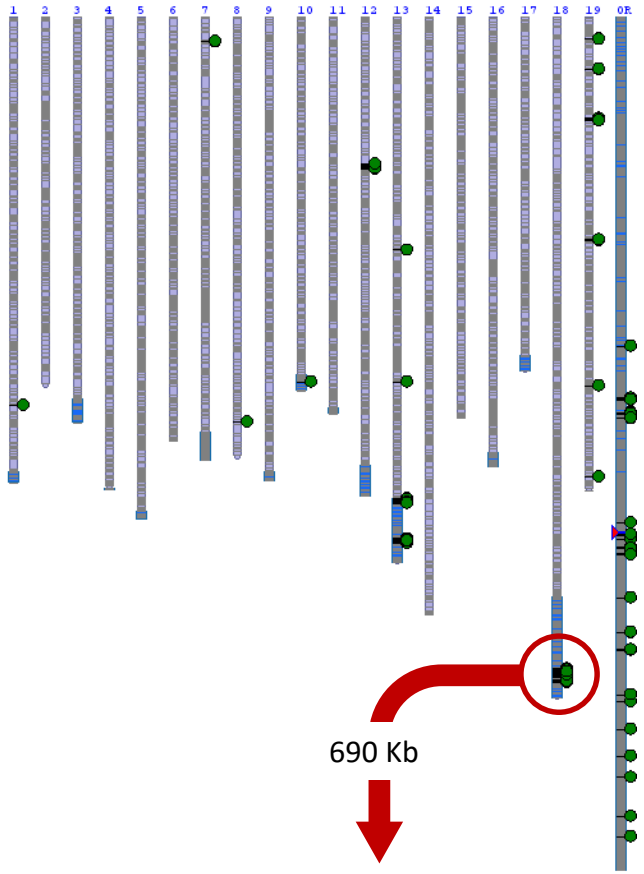
152 loci homologous to known TPS :

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

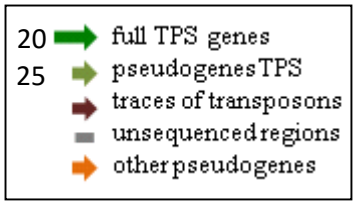


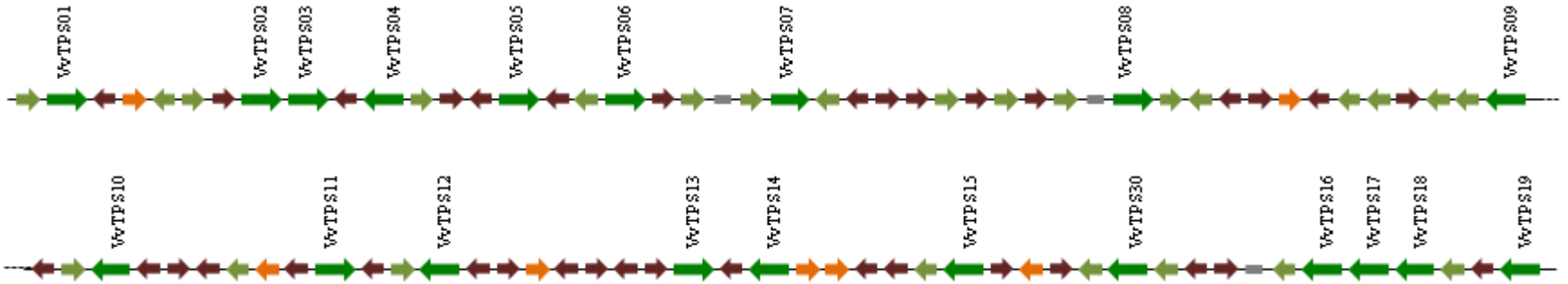
Martin et al 2010

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.





Manual curation of the CHS and STS gene families

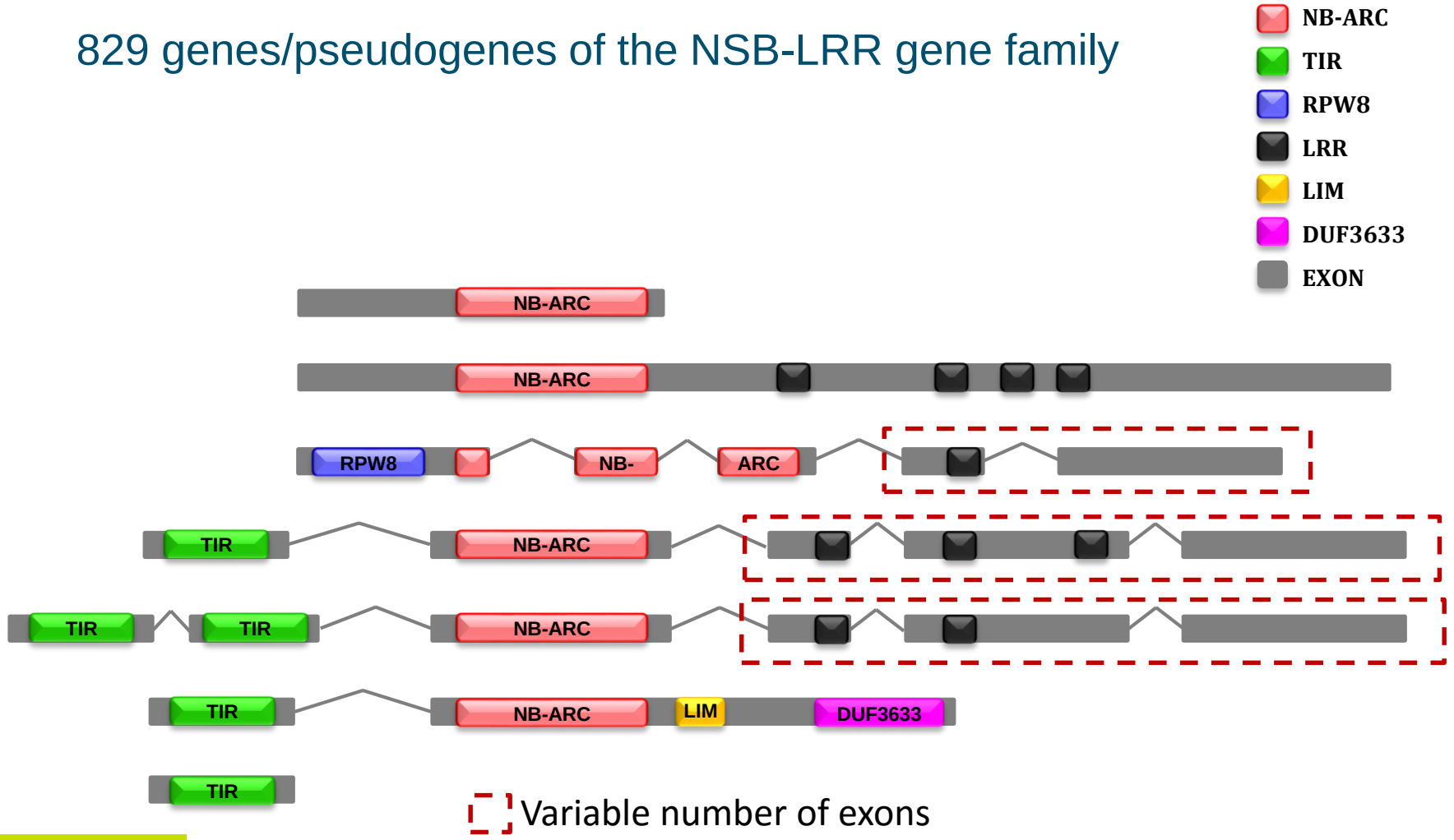
http://urgv.evry.inra.fr/projects/FLAGdb++

- 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



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

Apollo (login required)

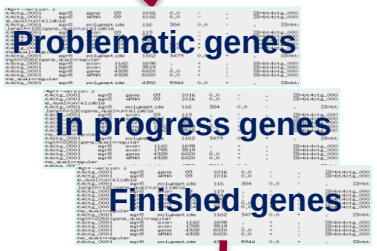
Write

Apollo WebStart

Read

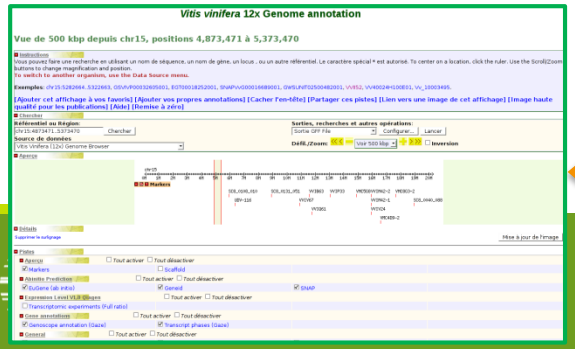
Chado

Export Edited genes (GFF3)



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

Gbrowse update



Read

Bio:Seq: Feature store

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

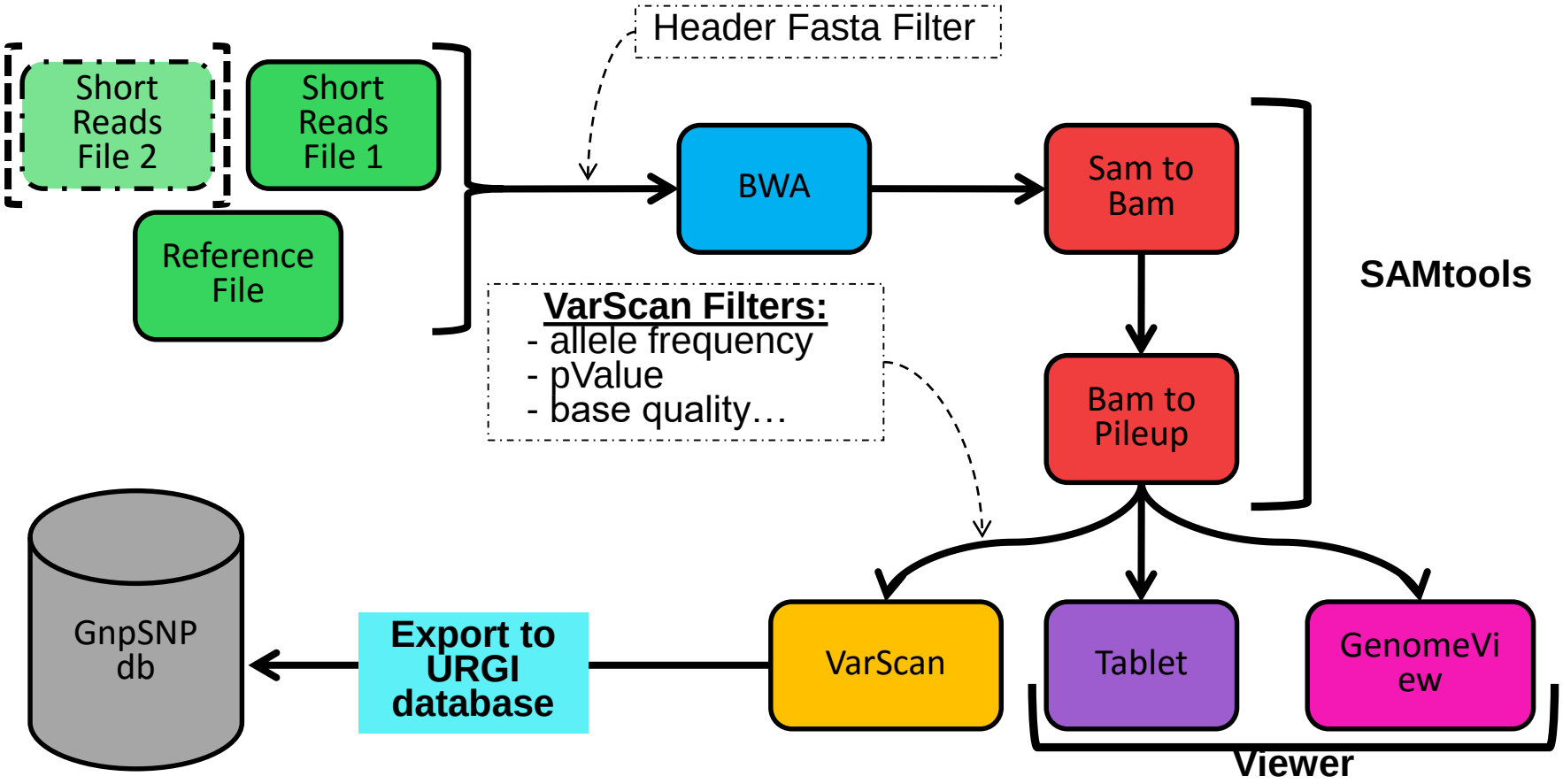
- Vigna italian project: 25 diverse *V. vinifera varieties*

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

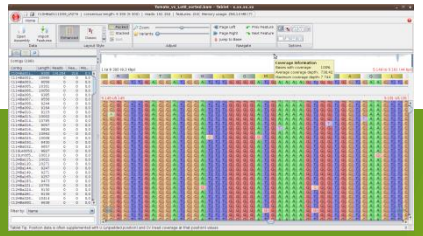
Read Length	Average	Standard dev	Min	Max
Read length (after trim)	94	6,38913	72	100
Bases after trimming	9 245 106 087	6 109 483 264	3 717 274 733	33 989 041 742
Genome coverage after trimming	19	12.6	8	70
Genome coverage with uniq reads	13	8.1	4	47
% of the genome covered	81%	5%	67%	93%
SNP detected	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

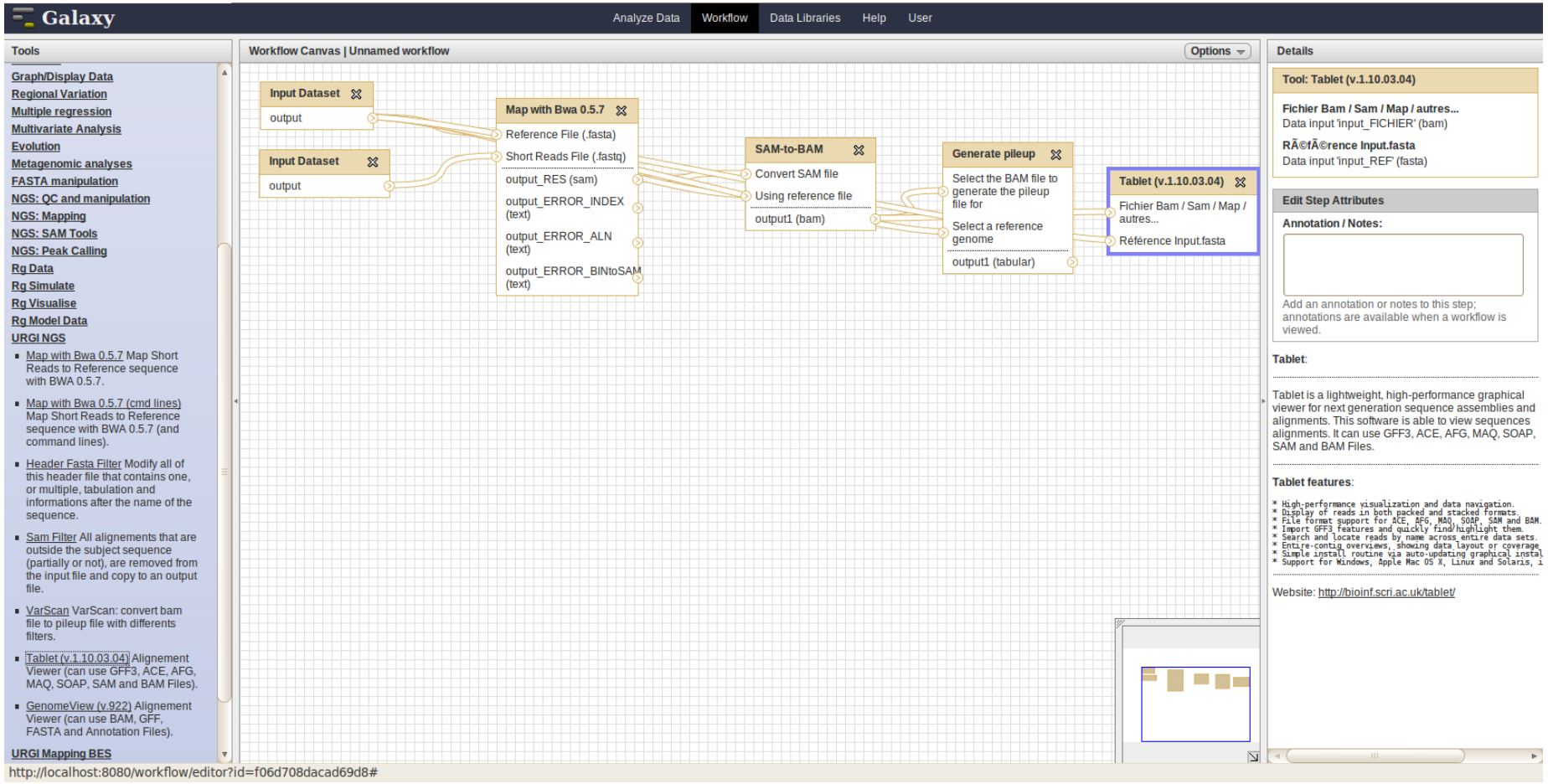


Calculations parallelized



MAPHiTS workflow in Galaxy

Galaxy is a web based genomic plateform (http://usegalaxy.org)



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 methodes that allow the assembly of complex genomes : improving the length of the reads, MATE paire reads, tuning the assemblers …
- Calculation power and time

Conclusions

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

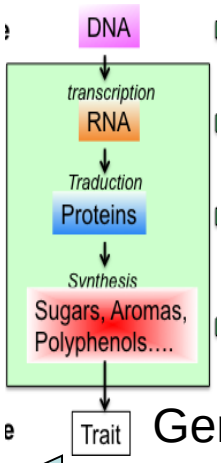
- How to predict a phenotype from the informations available on the genotype and its interactions with the environment?
- How will the existing varieties respond, and what are the new phenotypes we are looking for (ideotypes)?
- How do we develop and manage durable resistances in grape varieties?

→ Modeling at different scales

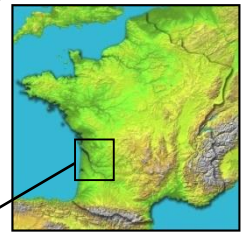
Cell Organ Plant Vineyard Region/landscape

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

Phenotyping/ environment characterisation at different scales



Genomics



Macro

Meso

Local

Micro

Region/landscape

Cell | Leaf | Canopy | Plot | Estate | Coop | AOC Region

1 cm | 10 cm | 1 m | 10 m | 100 m | 1 km | 10 km | 100 km | 1000 km

Genomics to help understand the underlying processes and caracterize 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

C. Caron

P. Hugueney
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