PAG XXV, January 14-18, 2017 - San Diego, CA, USA

Pan Genome and Structural Variation Analysis using different Approaches for *de novo* Assembly and Haplotype Reconstruction

Davide Scaglione
dscaglione@igatechnology.com

image: camera di commercio di Bolzano

IGA
ISTITUTO DI GENOMICA APPLICATA

INTERNATIONAL PLANT & ANIMAL GENOME CONFERENCE

IGATech™

Outline

Mapping-based SVs identification

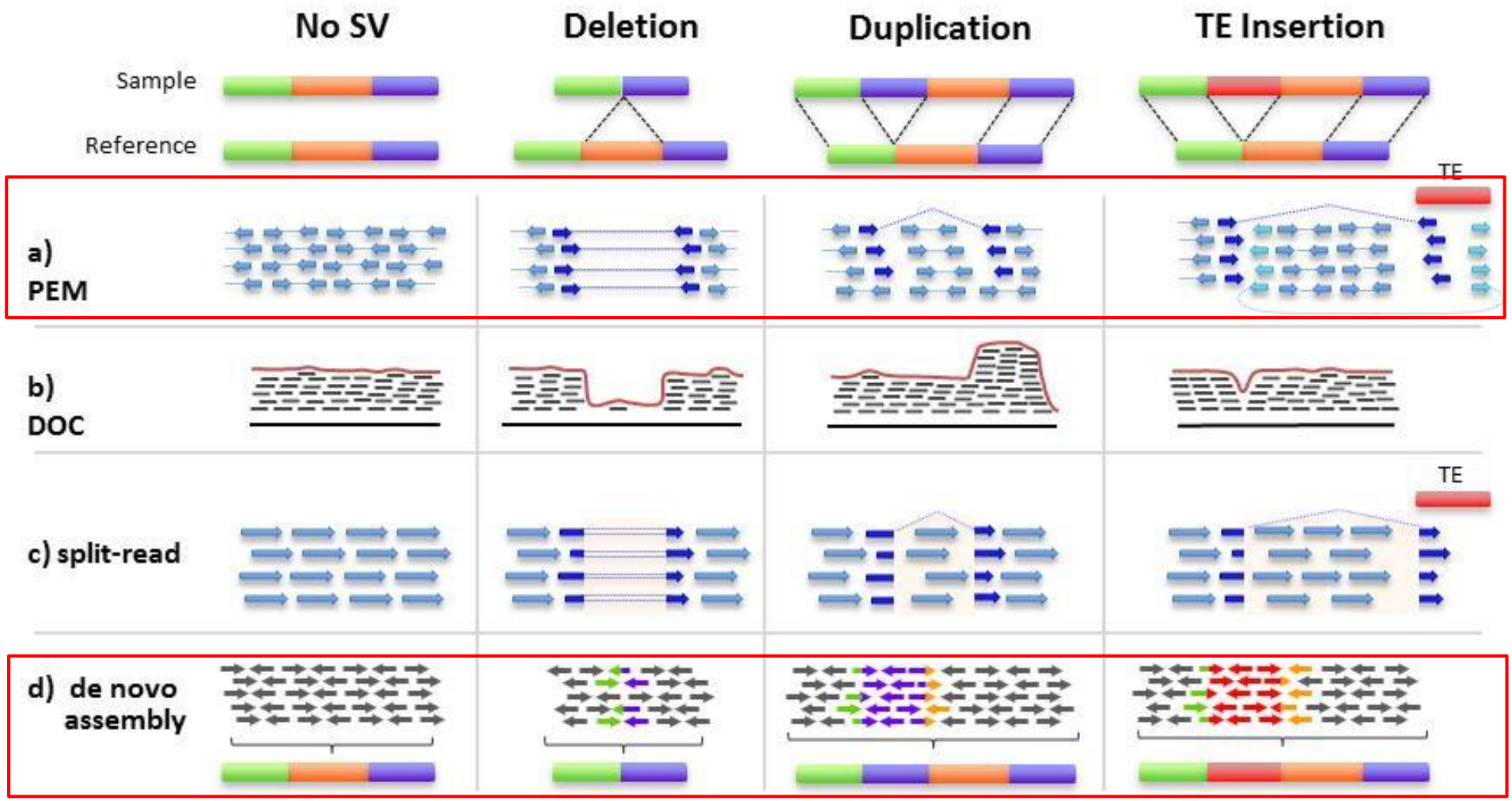
De novo assembly of 6 grape varieties

Reconstruction/improvement of reference chromosomes with Hi-C data

Impact of SVs in gene-space and RNA-Seq analyses

Mapping-based SVs identification

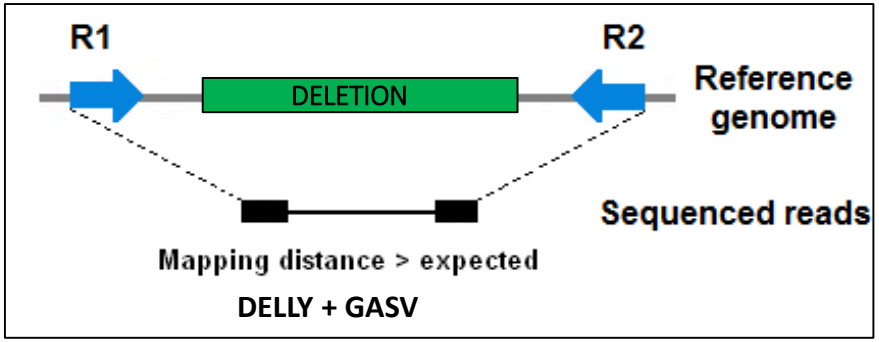
Detecting Structural Variation Using NGS Data



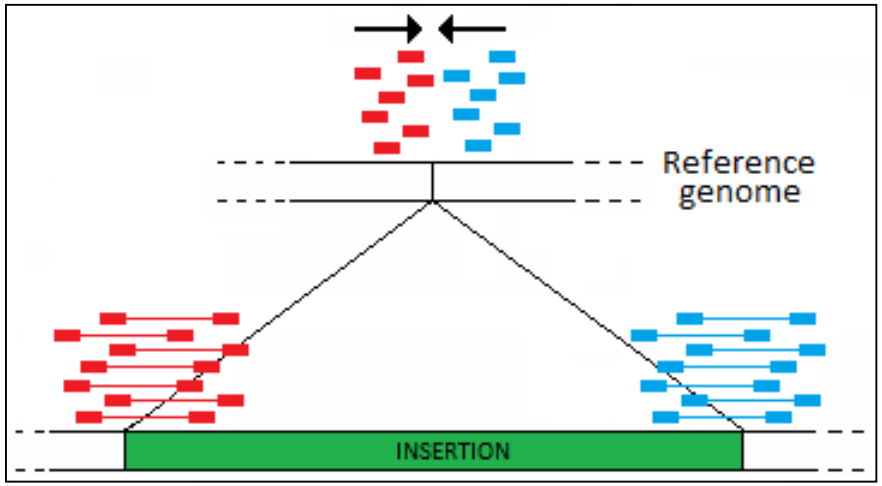
Marroni et al. Curr. Opin. Pl. Biol. 2014

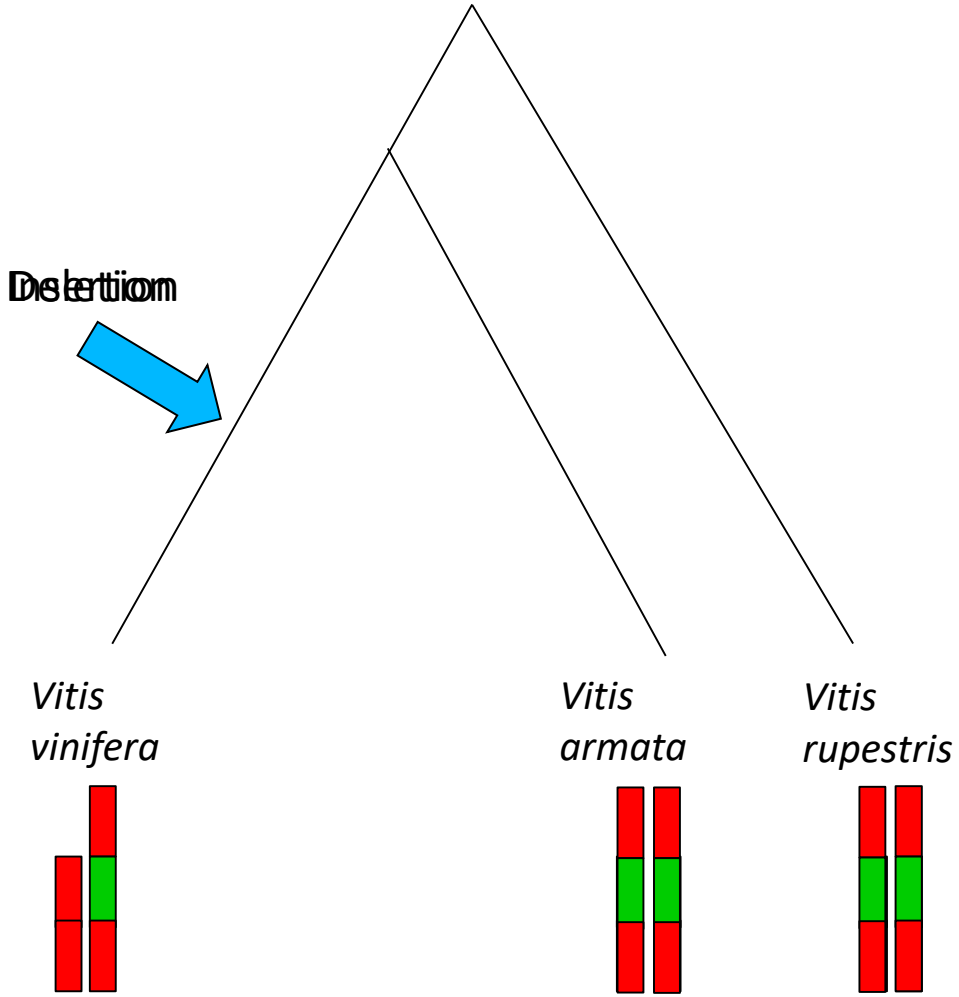
SV Detection and SV Origin

Deletion detection



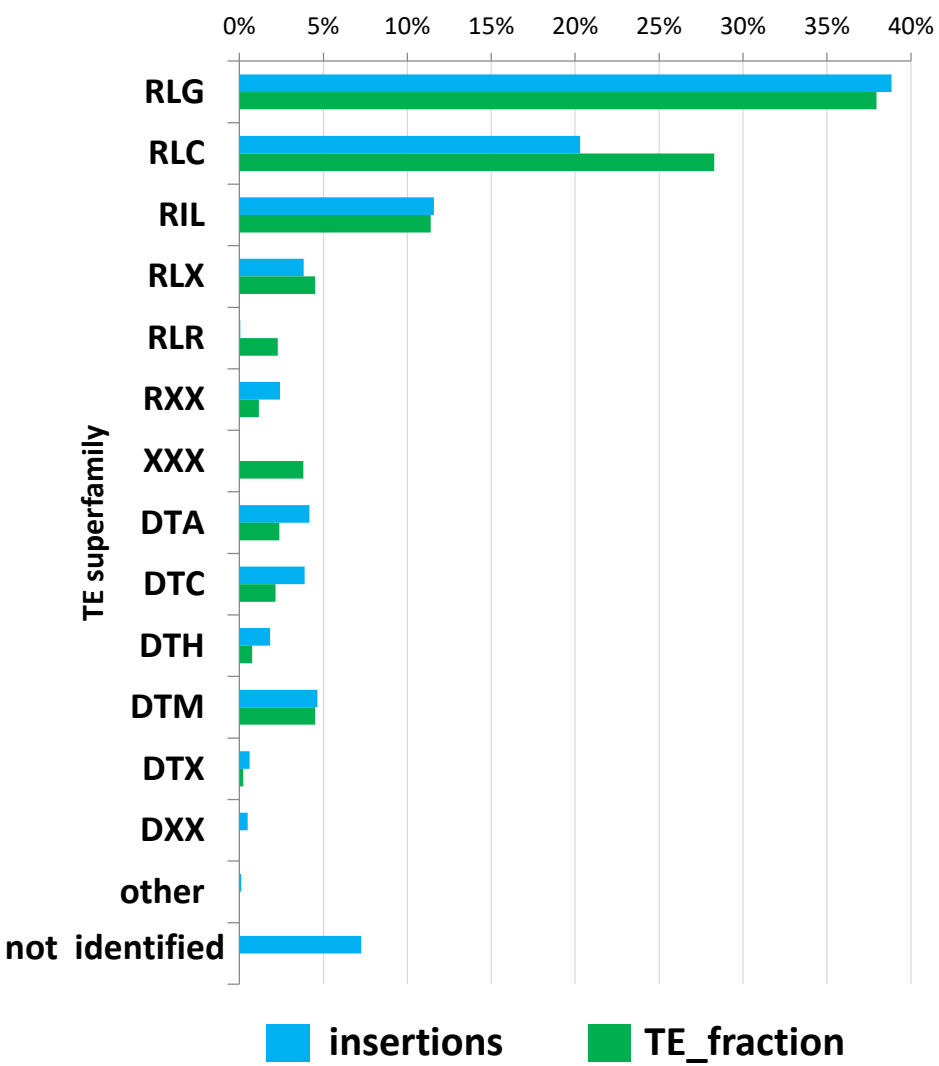
Insertion detection



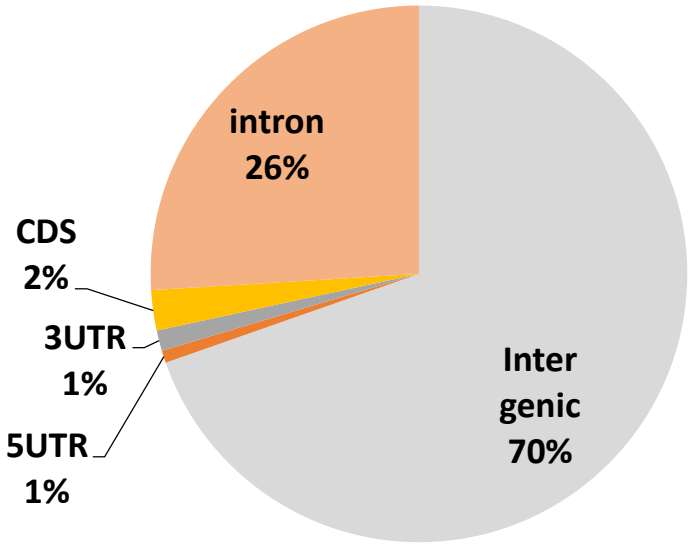


SV detected: 1-25 Kbp

SV Variation: Insertions

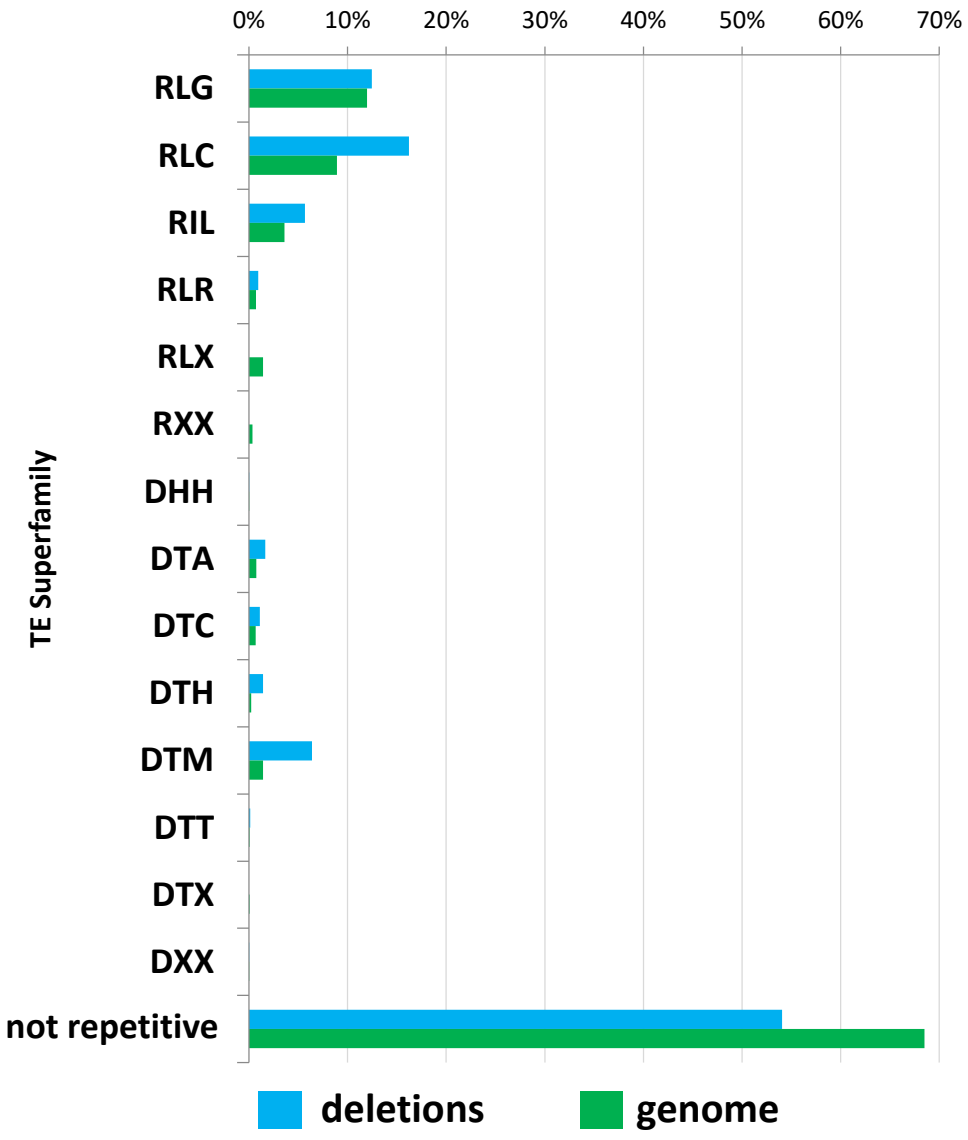


Total insertions 54500
Total length 316.84 Mbp
N. of genes involved 2351

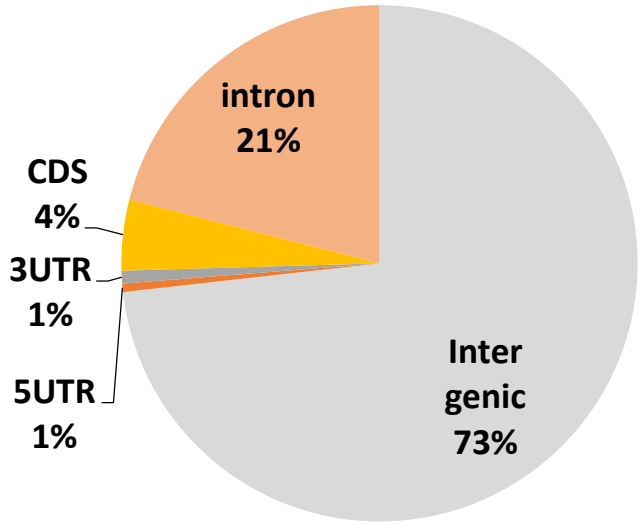


Transposon insertions

SV Variation: Deletions



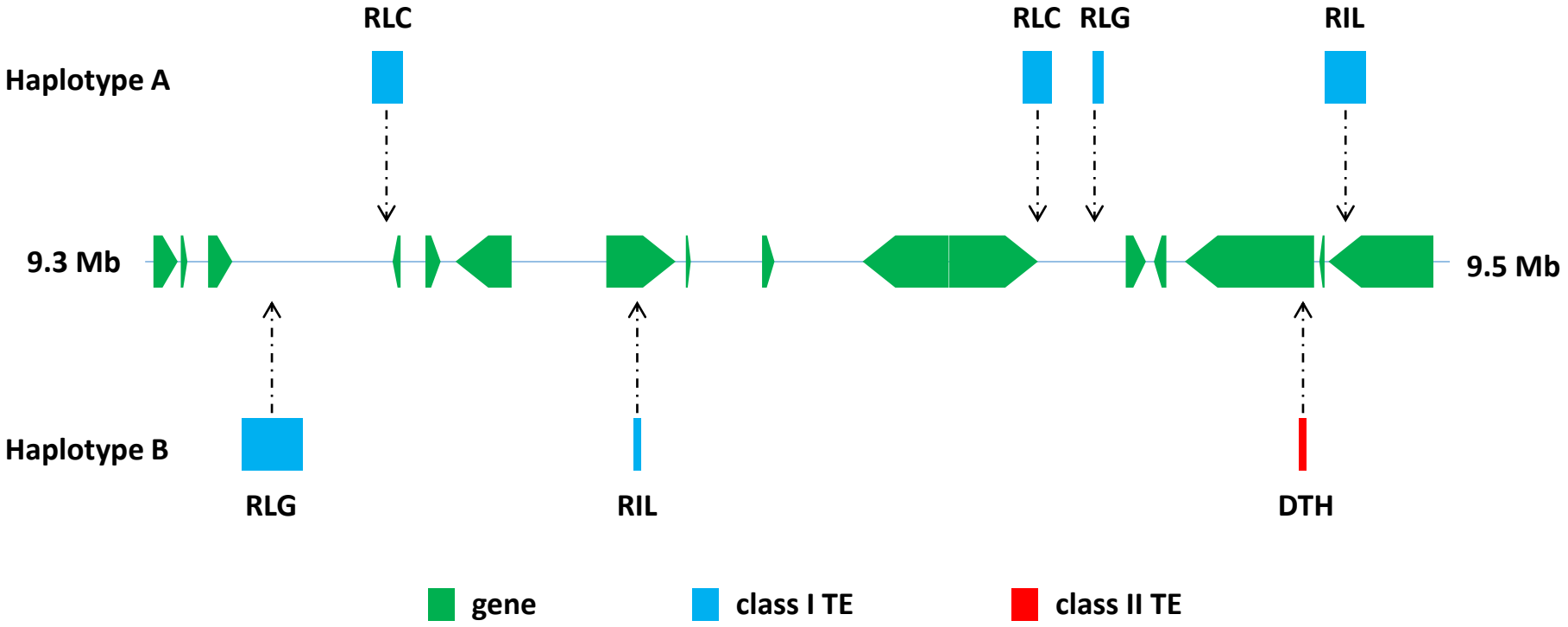
Total deletions 7856
Total length 42.89 Mbp
N. of genes involved 436



Real deletions (DSB+NHEJ)
Transposon excisions

Reconstructing haplotype SV through genetics

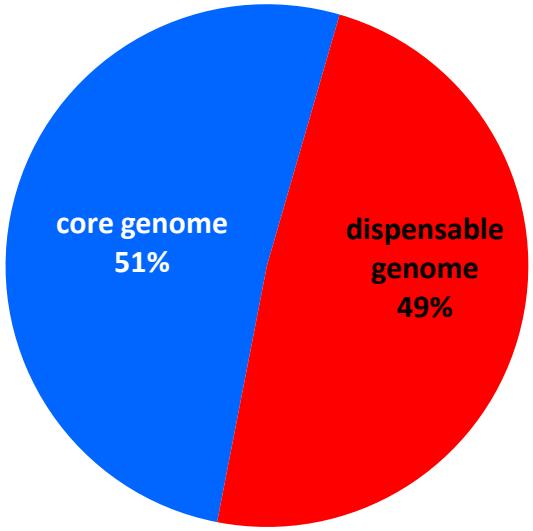
Comparing the two homologous chromosomes :
Sequencing of 4 highly homozygous individuals from self-pollinated Rkatsiteli allowed to cover 93% of chromosomes lenght with haplotype-specific reads



The pan-genome in grapevine: short SVs

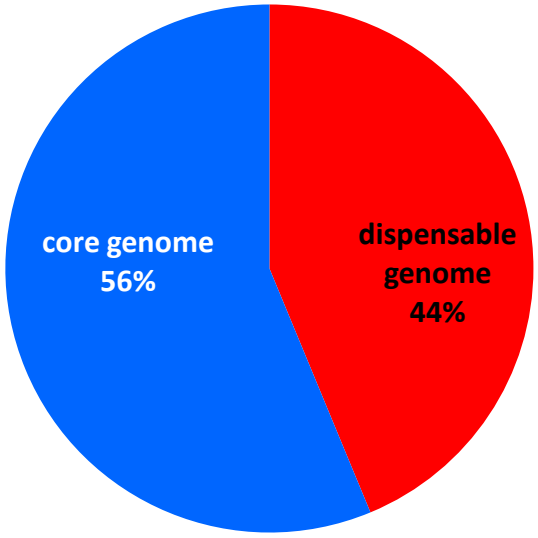
Broad definition

	Size (Mb)
Reference genome	486.21
Pan genome	752.11
Core genome	386.82
Dispensable genome	365.29



Strict definition

	Size (Mb)
Reference genome	486.21
Pan genome	752.11
Core genome	423.06
Dispensable genome	329.05



Larger SVs (of the CNV type) involving genes are also observed

De novo assembly of 6 grape varieties

ALLPATHS-LG recipe: *de novo* assembly of 6 varieties

	Cov. (X)
Traminer	
overlapping fragment	20
fragment	33
	53
jumping	19
long jumping	16
	36
Gouais Blanc	
overlapping fragment	28
fragment	35
	62
jumping	16
long jumping	14
	30
Cabernet Franc	
overlapping fragment	55
	55
jumping	23
long jumping	10
	33

	Cov. (X)
Rkatsiteli	
overlapping fragment	22
fragment	32
	55
jumping	29
long jumping	12
	40
Kishmish Vatkana	
overlapping fragment	29
fragment	23
	52
jumping	24
long jumping	10
	35
Sangiovese	
overlapping fragment	33
fragment	34
	68
jumping	23
long jumping	22
	45

a) ~30X Overlapping fragment libraries

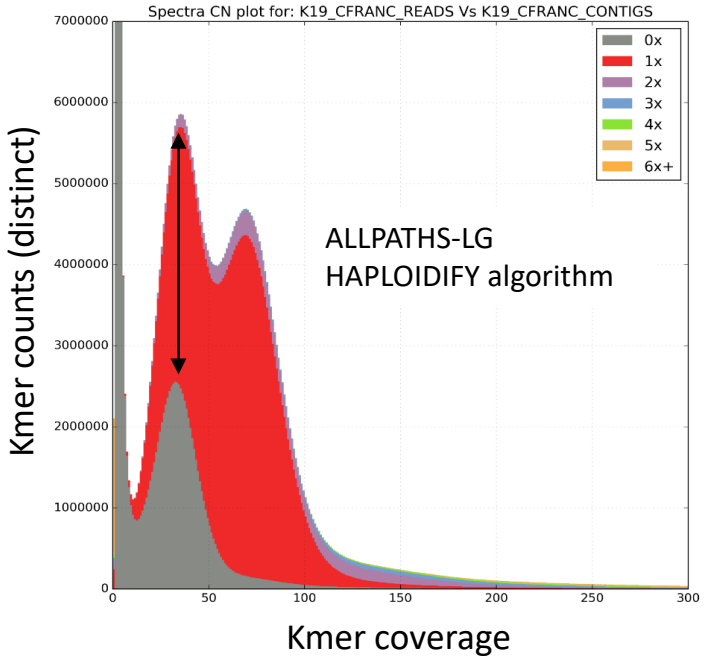
b) ~20X Non-overlapping fragment libraries

c) ~20X 3kb Jumping libraries + ~10X 10kb (long) Jumping libraries

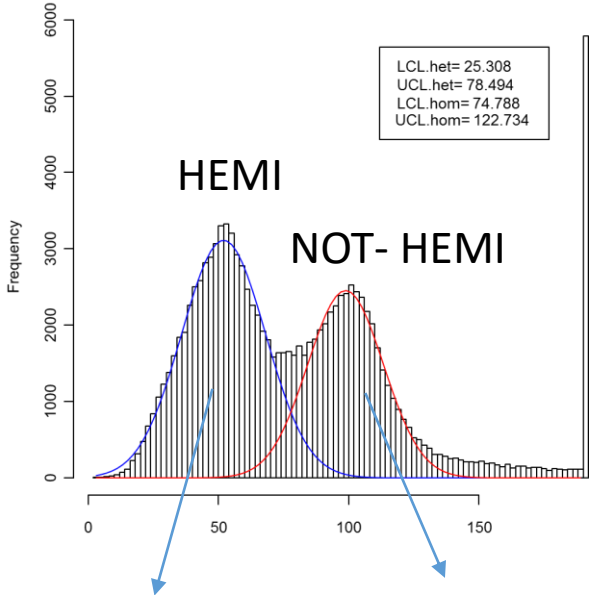
ALLPATHS-LG performances

Scaffold metrics	traminer	gouais blanc	cabernet franc	rkatsiteli	kishmish vatkana	sangiovese	pn40024
Total (bp)	678,662,631	743,730,069	711,707,043	753,744,202	610,410,426	691,235,812	521,491,794
Sequences (#)	8,676	9,564	12,485	10,089	8,546	15,667	6,016
N50 (#)	484	524	1,178	612	457	627	553
L50 (bp)	413,025	408,071	178,391	352,572	364,852	301,183	264,099
Gaps (#)	67,484	66,217	55,190	69,862	64,039	61,417	40,378
Median gap length (bp)	4,012	5,545	4,405	5,200	5,728	3,581	3,825

Final scaffold length is artificially inflated by overestimation of gap size

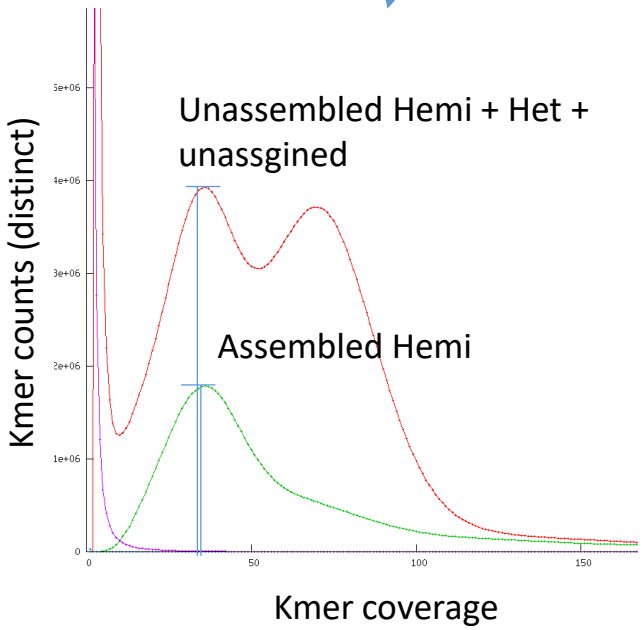


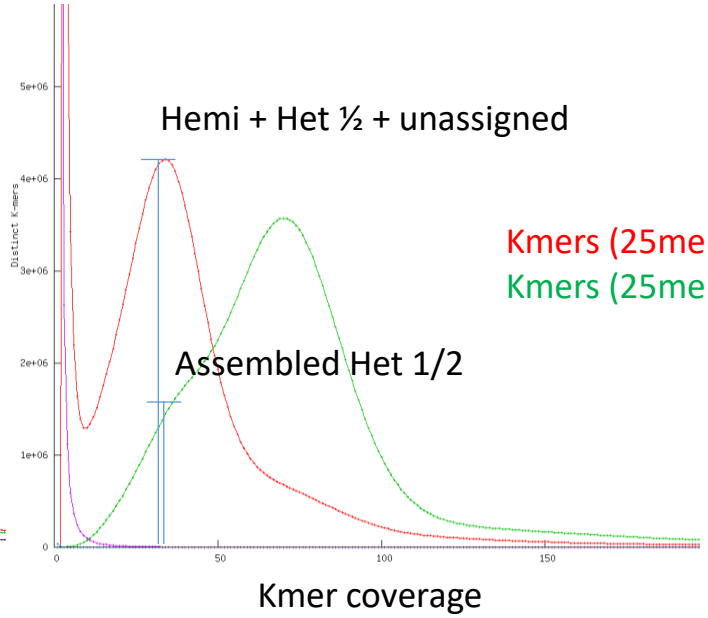
Alignment coverage as a proxy to catalog hemizygous contigs



Coverage plot (contigs)

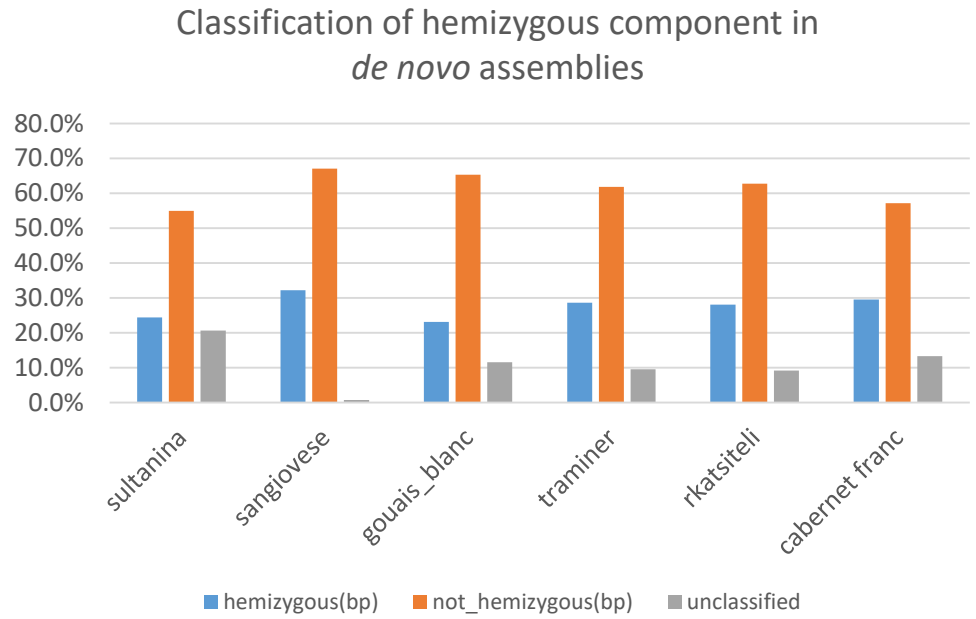
Fact-checking with kmers:
sequencing reads VS HEMI/non-HEMI contigs

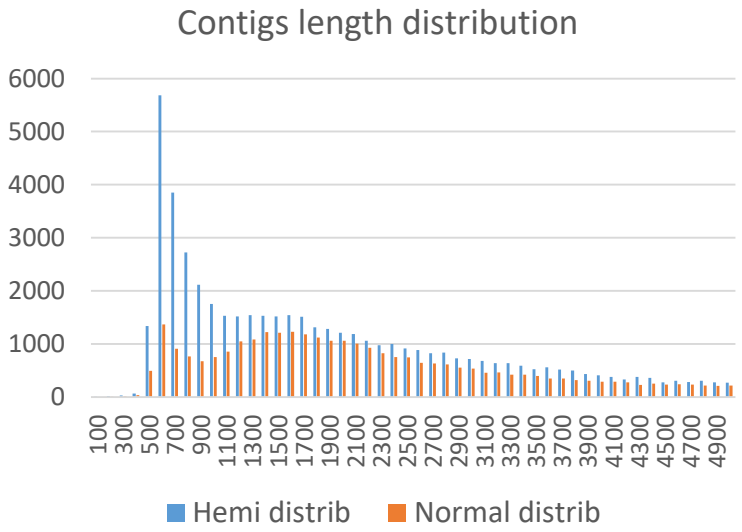




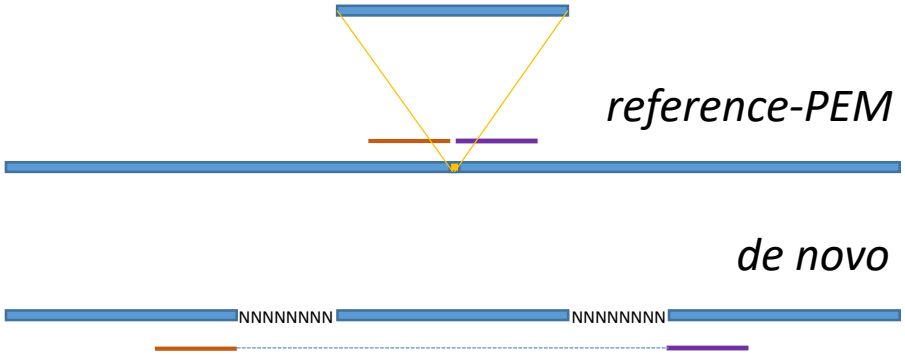
Kmers (25mer) exclusive to reads dataset
Kmers (25mer) shared with the reference (subset)

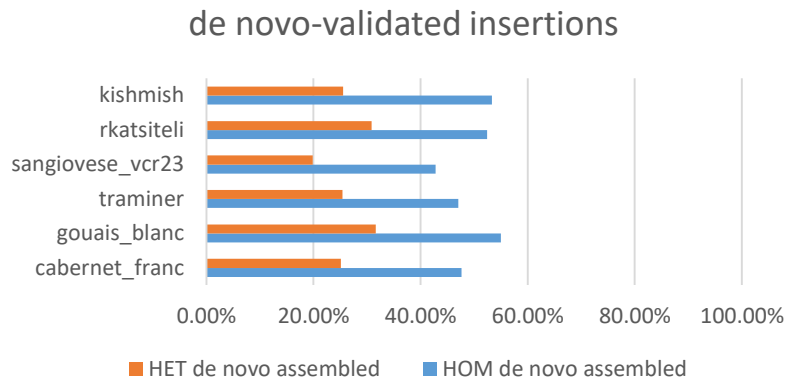
Classification of hemizygous contigs

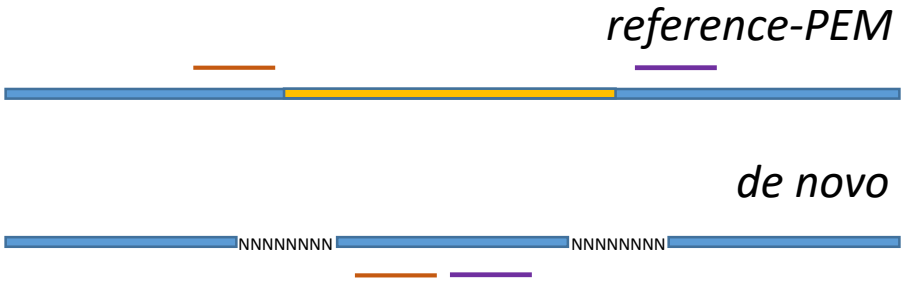


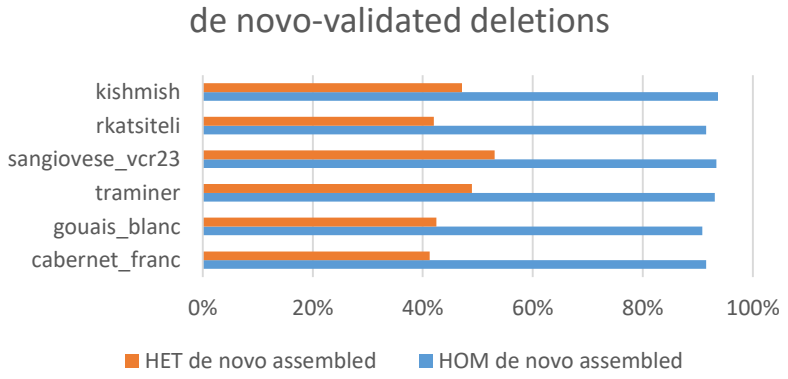


Validation of PEM-detected SV with de novo assemblies



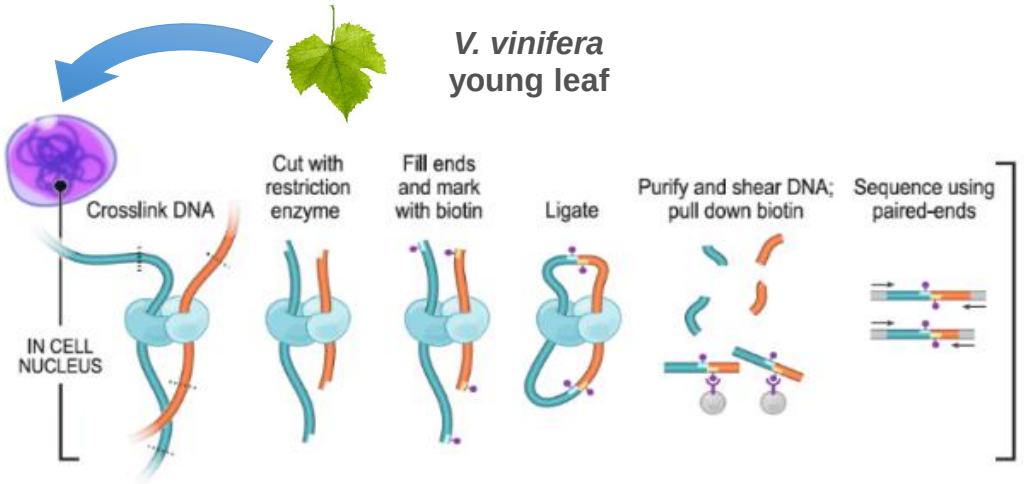






Reconstruction/improvement of reference chromosomes with Hi-C data

In situ Hi-C and TCC in Vitis vinifera

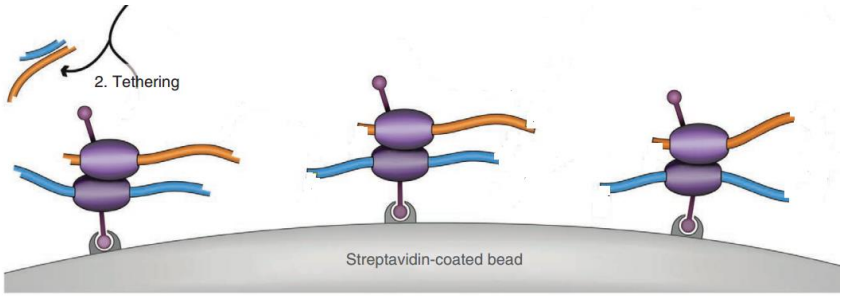


(Rao, *et al.* 2014)

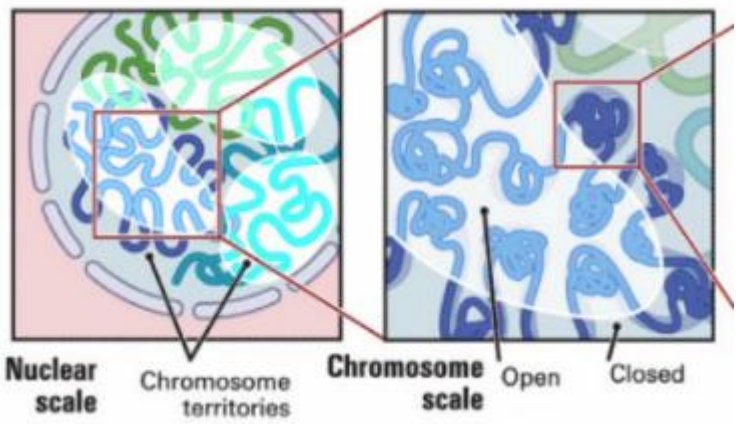
Advantages:

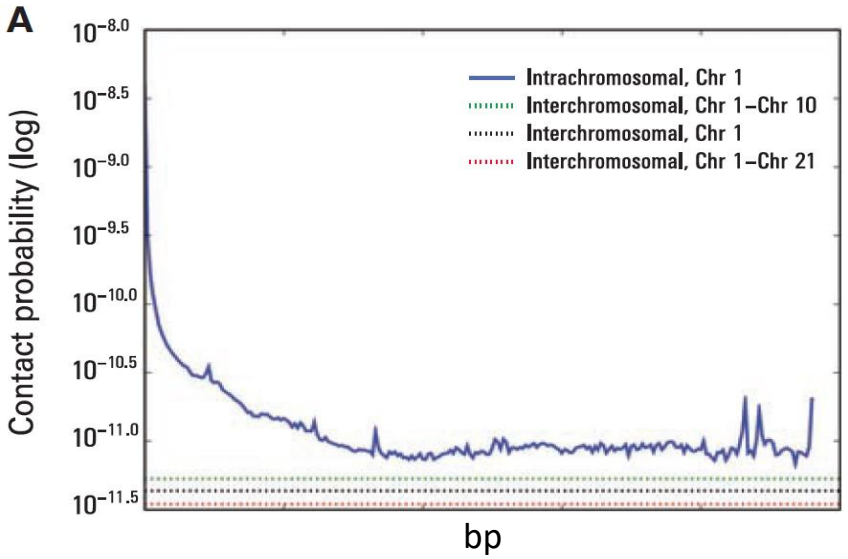
- Reduced frequency of spurious contacts due to random ligation in diluted solution
- Faster protocol (requiring 3 days instead of 7)
- Enables higher resolution (up to ~1Kb)

TCC (tethered conformation capture)

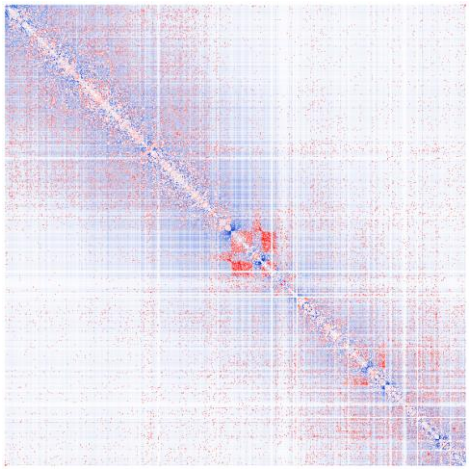


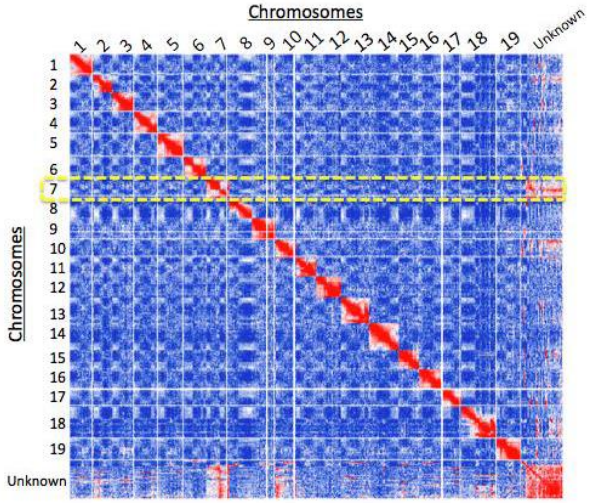
Proximity information and interaction information



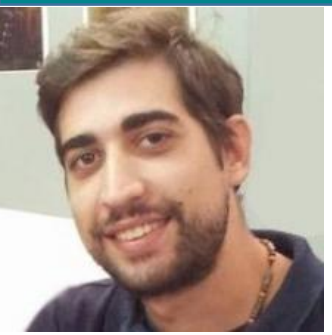


Lieberman-Aiden et al., Science 2009





Improvement of PN40024 reference with Hi-C (TCC) data

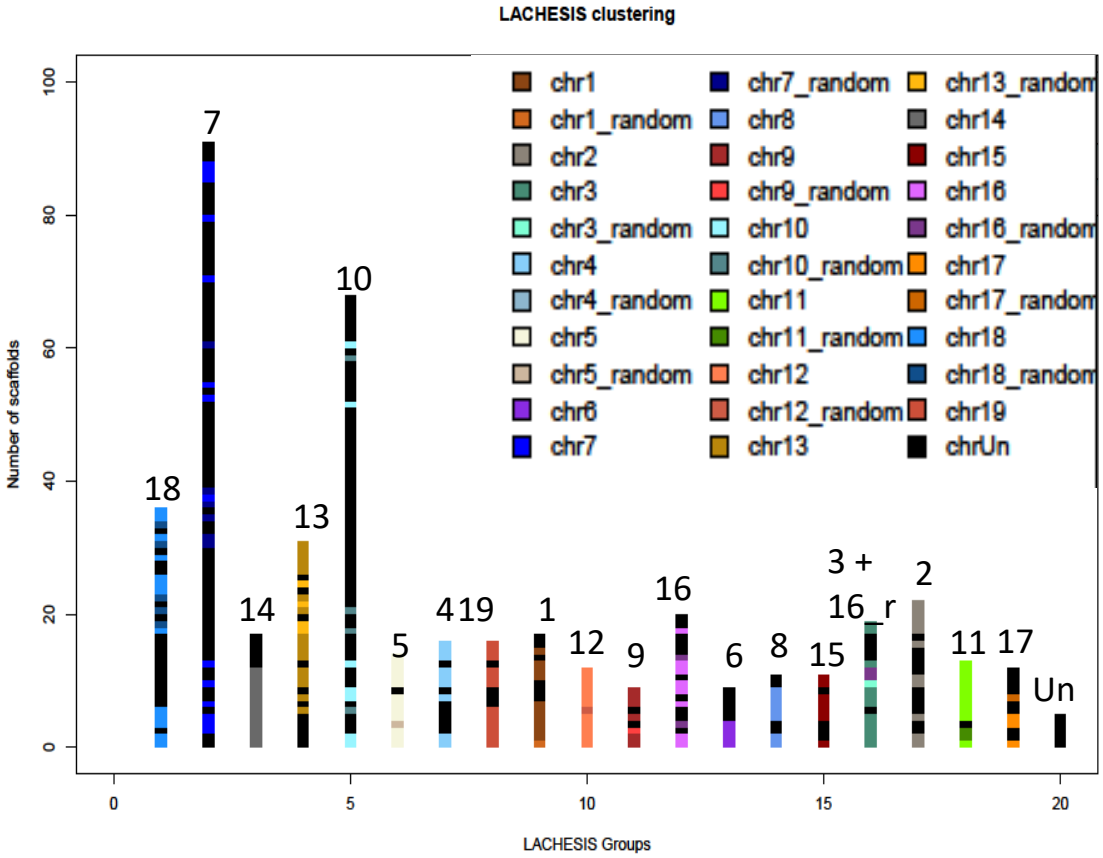


Clustering improvement:

Total scaffolds: 2059

chrUn scaffolds: 1849/2059

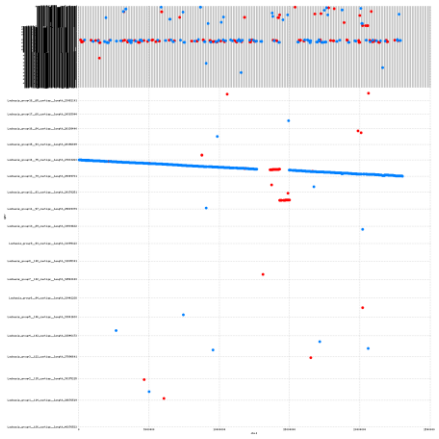
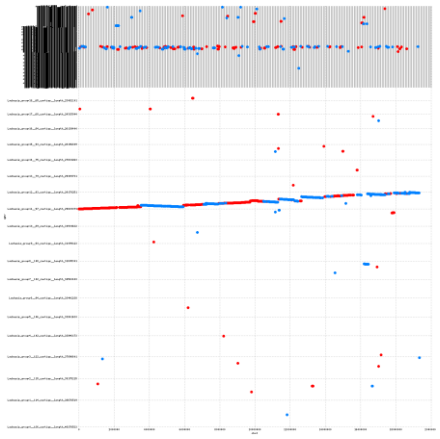
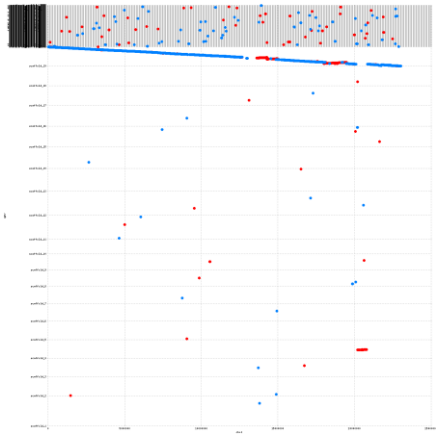
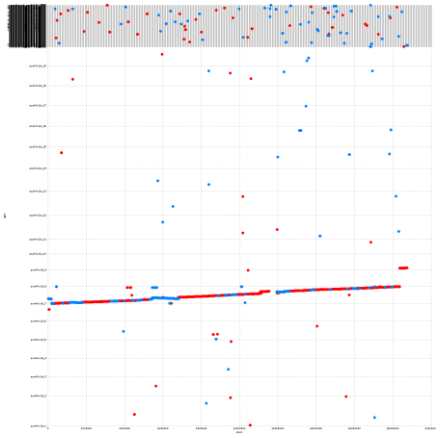
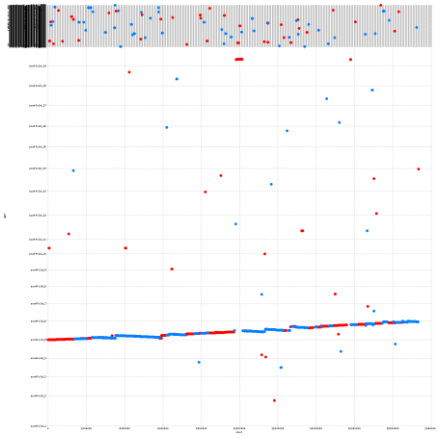
Assigned chrUn scaffolds: 1834/1849
~ 39 Mb added



- Hi-C
- Genetic maps
- Mate-pairs

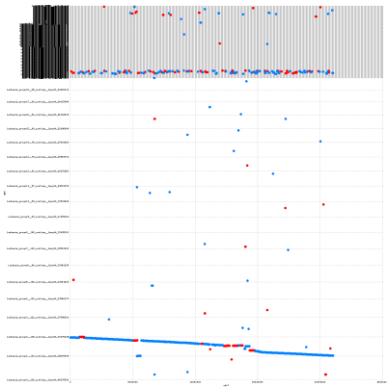
Release of an impoved golden path for 12X Sanger scaffolds (no changes to scaffolds sequences)
Candidate release date: April 1st 2017

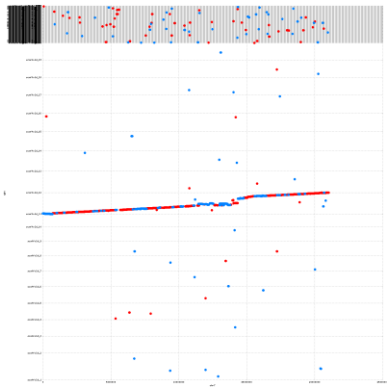
Chromosome-scale reconstruction of rkatsiteli assembly

	Chr01 (ref)	Chr02 (ref)	Chr03 (ref)
HindIII			
MboI			

Chromosome-scale reconstruction of rkatsiteli assembly

Chr07





	HindIII	MboI
Clustering efficiency (19 groups) (bp)	95.16%	95.38%
Ordering efficiency (bp):	80.64%	**90.05%**

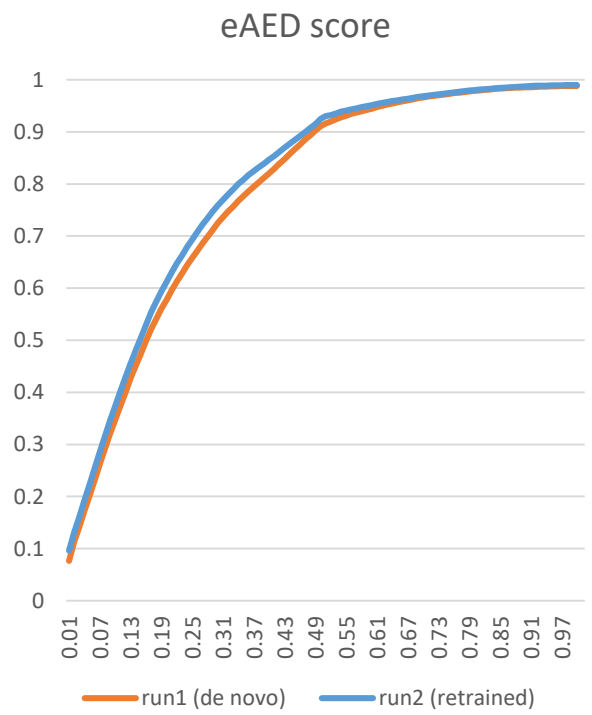
Rkatsiteli: Gene prediction with Maker-P

Evidences:

- RNA-Seq from leaves, roots and berries
- V. vinifera EST collection (NCBI)
- PN40024 reference transcripts (v2.1)
- V. vinifera proteins (NCBI)

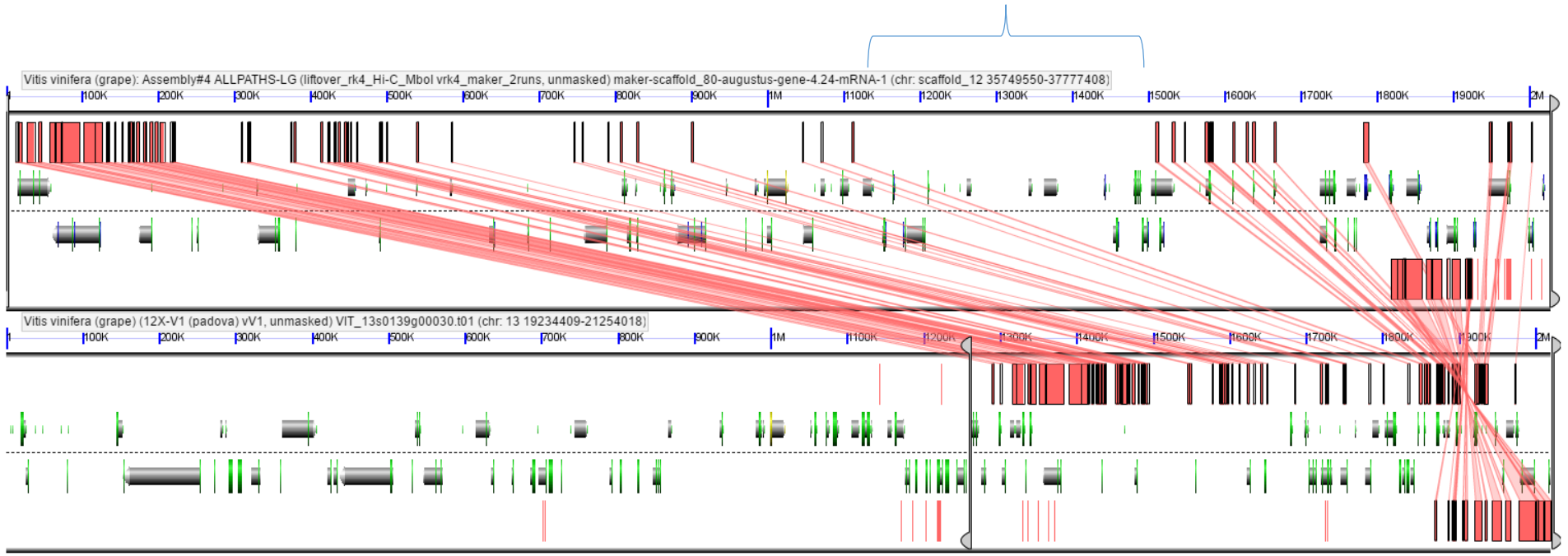
Training of Augustus and SNAP *ab initio* predictors

32,527 gene models
(30,406 on reconstructed pseudomolecules)



Sinteny check and nclusion of previously un-ordered gene-space

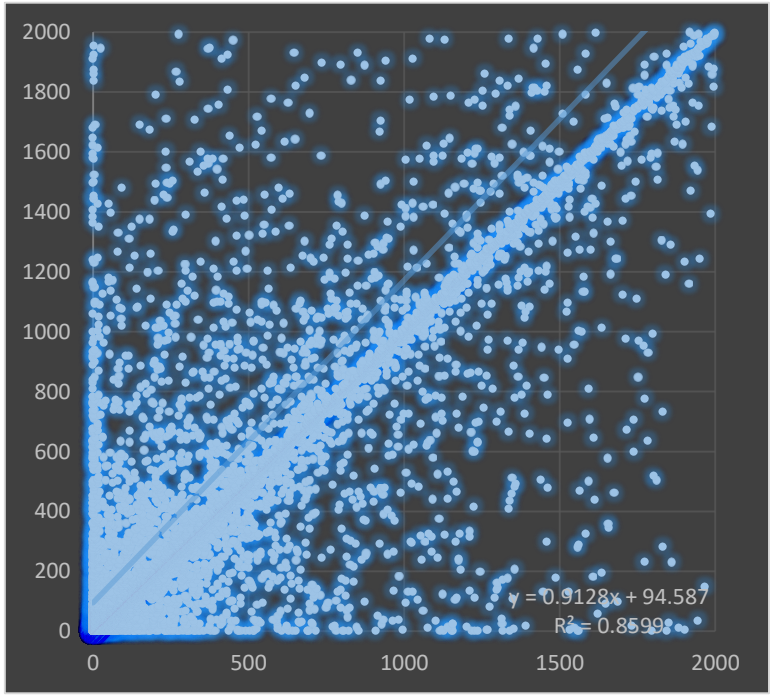
Segments belonging to “random” chromosomes in PN40024



Impact of SVs in gene-space and RNA-Seq analyses

Impact of gene-models in RNA-Seq counting

Raw fragment count per gene model



Raw frag. count on RK4

8690 unique syntelogs
[-3 < log(FC) < 3]

3675 higher on PN40024 models
5015 higher on RK4 models

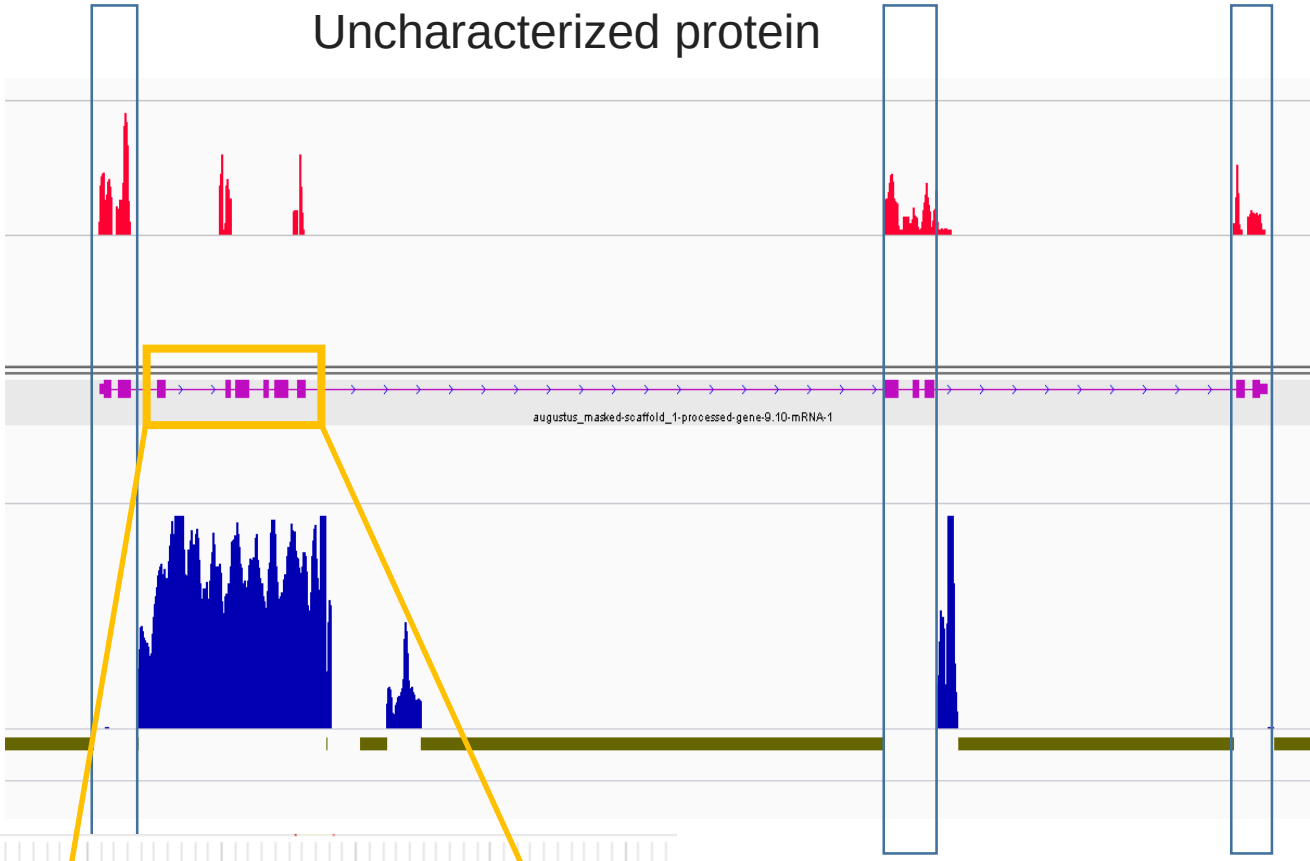
Missing exons with impact in RNA-Seq counts

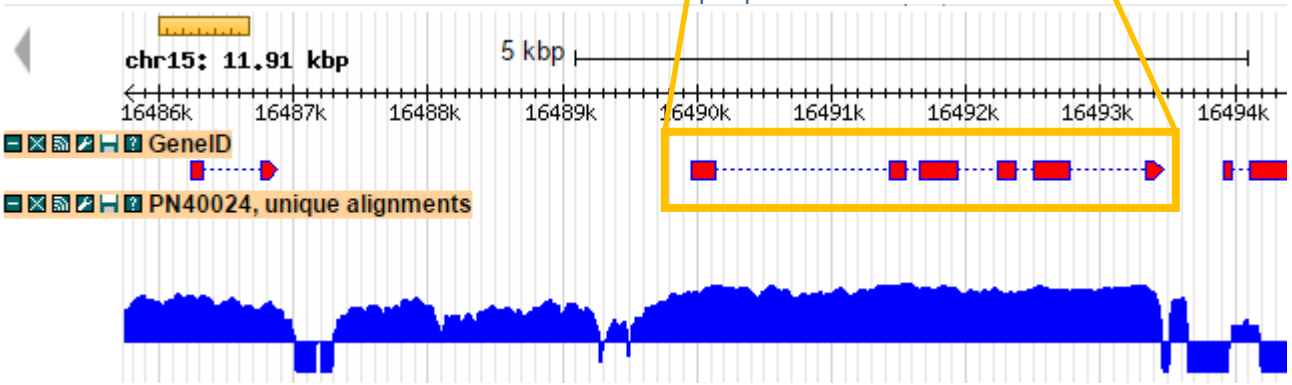
RNA-Seq counts

Ab initio
gene model

PN40024 reads coverage

Rkatsiteli assembly gaps





Missing exons with impact in RNA-Seq counts

Vitis vinifera putative disease resistance
RPP13-like protein 1 (LOC104881047)
95% identity

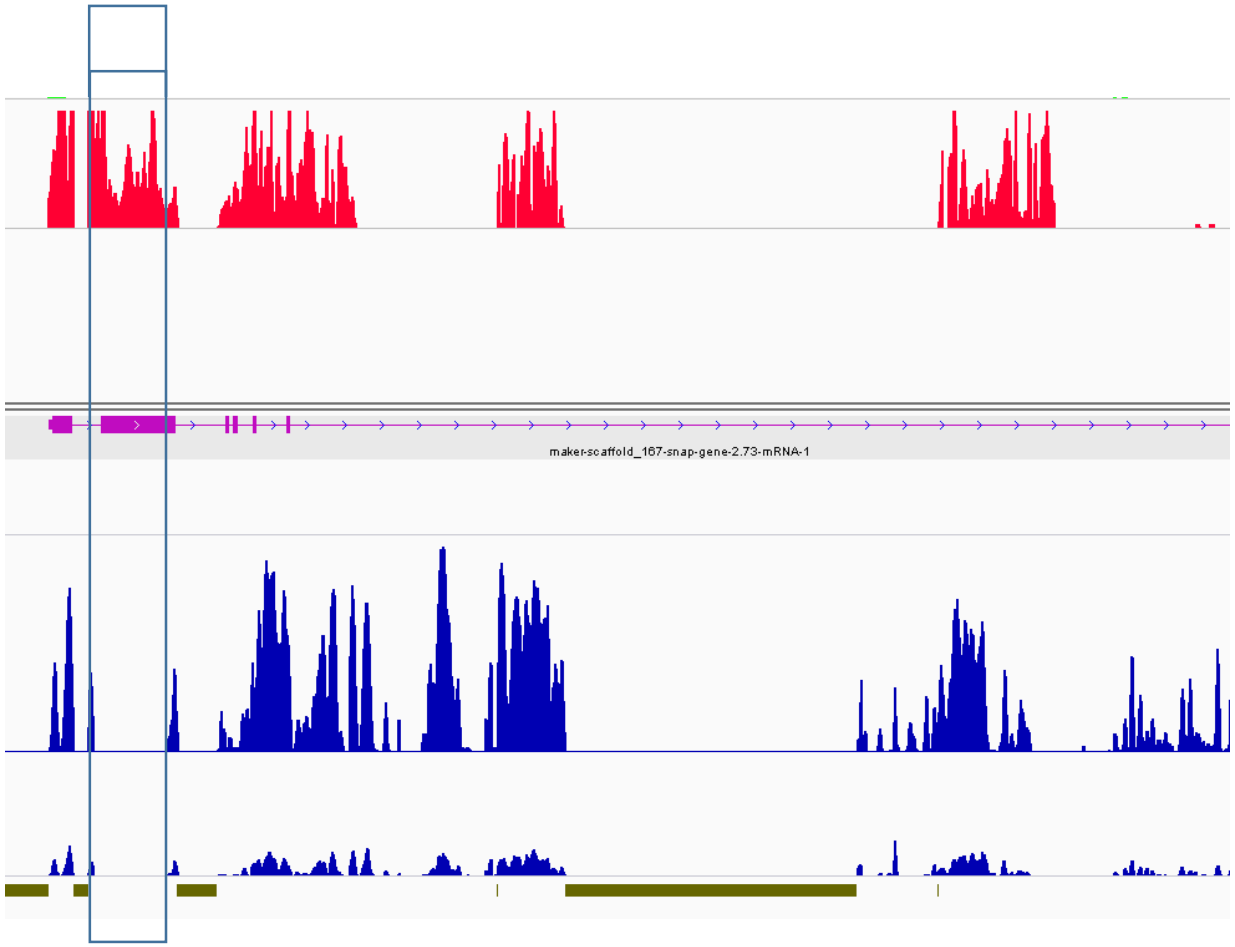
RNA-Seq counts

Ab initio
gene model

PN40024 reads coverage

12X assembly simul. reads

Rkatsiteli assembly gaps



Or even complete missing genes

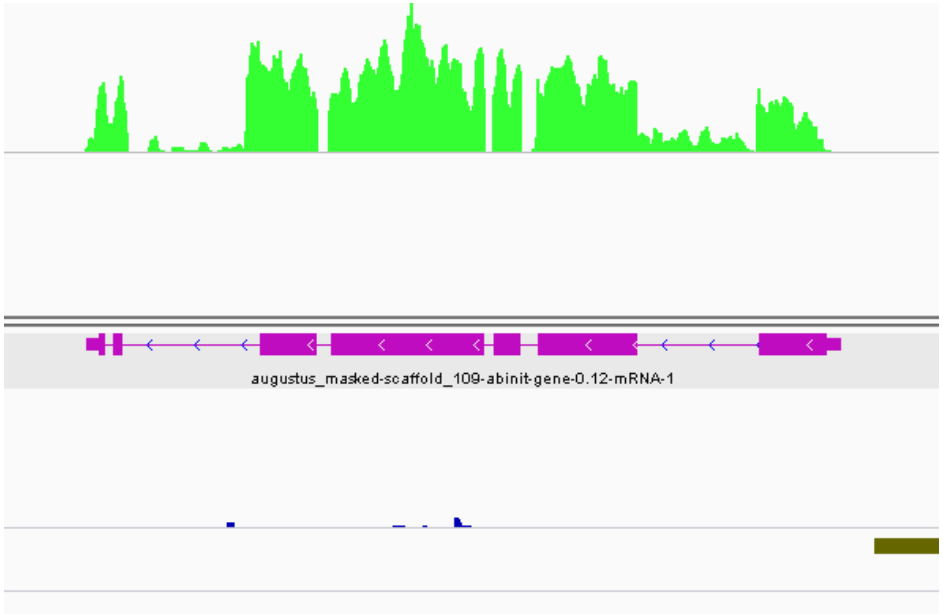
Vitis vinifera TMV resistance protein N (LOC100241404), transcript variant X2, mRNA
79% identity

RNA-Seq counts

Ab initio
gene model

PN40024 reads coverage

Rkatsiteli assembly gaps



Exons with >25% missing coverage
from reference PN40024 &&
>5 reads of RNA-Seq (berry) evidence

1669 exons
932 genes
~380 Kbp of expressed PAV coding sequence (rkatsiteli)

Conclusion

•(short reads sepaking) Identification of TE-related SVs is far better accomplished via PEM strategies

•*De novo* assembly can be a proxy to detect contigs involved in hemizygosity

•Hi-C (in situ) can provide high quality data to reconstruct and order chromosome even starting with relatively short scaffolds

•Variety-specific gene models can actually improve (also depending on prediction accuracy) expression analysis and the investigation a previosuly uncharacterized gene-space

•Releasing an improved PN40024 golden path

•FUTURE: to implement varietal-assemblies with 10X and PacBio

Acknowledgments

Michele Vidotto
Aldo Tocci
Rachel Schwope
Gabriele Magris
Alice Fornasiero
Eleonora Paparelli
Fabio Marroni
Gabriele Di Gaspero
Michele Morgante

Federica Cattonaro
Irena Jurman

Nicoletta Felice
Giusi Zaina

Novabreed
ERC-2011-ADG