Characterisation of the pan-genome of *Vitis vinifera* using Next Generation Sequencing

Gabriele Magris[1,2], Michele Vidotto[1,2], Sara Pinosio[1,2], Eleonora Paparelli[1,2], Fabio Marroni[1,2], Giusi Zaina[1], Gabriele Di Gaspero[2], Michele Morgante[1,2]

[1] Dipartimento di Scienze Agroalimentari, Ambientali e Animali, Università di Udine, Udine, Italy - [2] Istituto di Genomica Applicata, Udine, Italy

e-mail: *gmagris@appliedgenomics.org*

1. INTRODUCTION

Genomes of individuals belonging to the same species can differ due to structural variation, encompassing both smaller variants due to transposable elements and larger ones, which modify the chromosomal structure. Based on these observations, a single genome might not reflect the entire genomic complement of a species. Therefore, the concept of pan-genome, originally introduced for bacteria [1], was extended to plants [2,3]. It is composed by a Core Genome (CG), shared by all individuals, and a Dispensable Genome (DG), present in some individuals, but not in all. To gain knowledge about the composition of the dispensable fraction, Structural Variants (SVs) ranging in size between 1 Kb and 25 Kb were identified in 50 grapevine varieties, based on the paired-end mapping information derived from the alignment of short reads to the reference genome sequence of *Vitis vinifera* [4].

2. METHODS

For the detection of SVs we integrated the results obtained using three different tools: DELLY [5], GASV [6] and a pipeline developed by our research group [7]. SVs were classified as insertions (Fig 1A) or deletions (Fig 1B), based on two outgroup species (*Vitis armata* and *Vitis rupestris*) and on their phylogenetic relationship with *V. vinifera*.

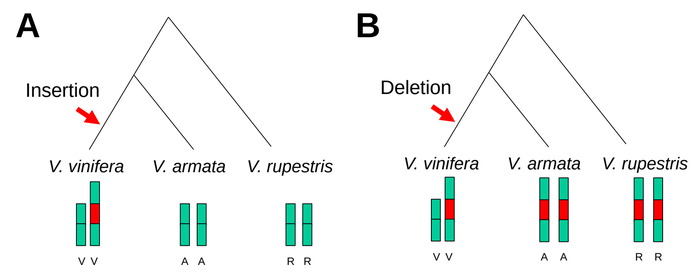

Fig 1. SV classification based on two outgroup species. A SV was classified as **insertion** if the sequence was present in *vinifera*, but absent in the outgroup species **(A)**. Vice versa, a SV was classified as **deletion**, if the sequence was present in the outgroup species, but missing in at least one *vinifera* variety **(B)**.

3. SV CLASSIFICATION & ANNOTATION

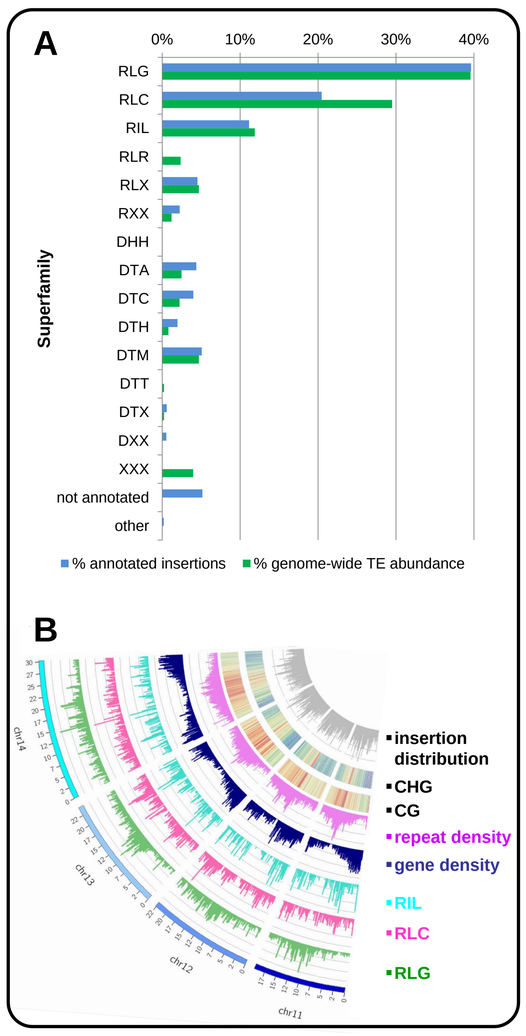

Fig 2. TE composition of insertions **(A)** and genome-wide distribution of class I TE's **(B)**.

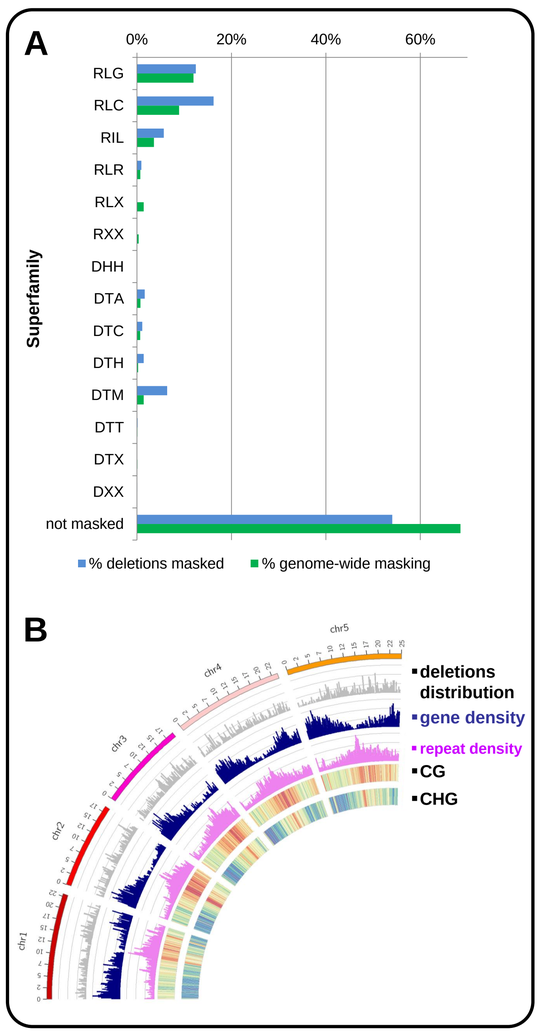

Fig 3. TE composition of deletions **(A)** and genome-wide distribution of deletions **(B)**.

Across the 50 grapevine varieties we identified a total of 7,856 deletions and 54,500 insertions. While insertions were mediated by the movement of Transposable Elements (TEs) (Fig 2), deletions were mostly unrelated to TE activity and represented the removal of random sequences mediated by double strand breaks and further defective repair events (Fig 3). LTR retrotransposons involved in SV were younger than the LTR-retros fixed in the population (Fig 4).

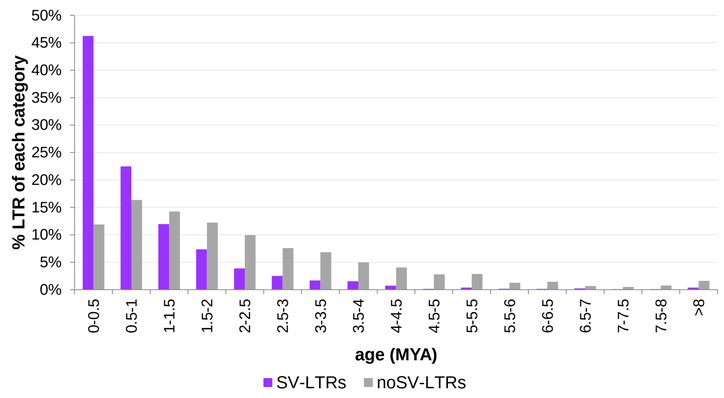

Fig 4. Insertion age estimation of LTR retrotransposons.

We further classified TEs involved in SV belonging to Copia (RLC) and Gypsy (RLG) superfamilies in families. For each family, we investigated their preference of localization compared to genomic locations [8] (Fig 5).

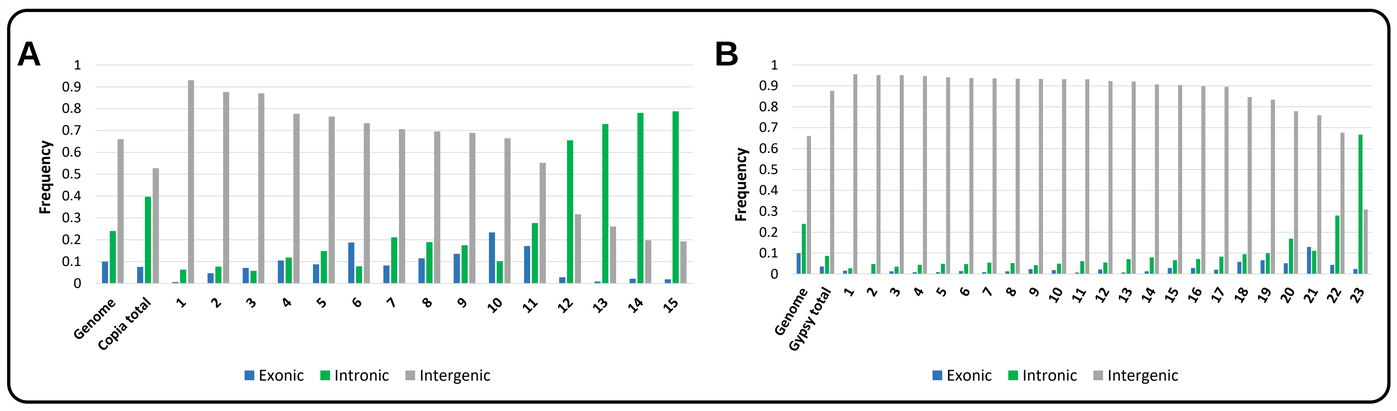

Fig 5. Localization of the most abundant Copia families **(A)** and Gypsy families **(B)**. Individual families representing the most abundant ones detected in SV events are indicated with numbers

4. SV vs GENE EXPRESSION

While promoting genetic variability, TEs may also disrupt genes, promoter or enhancer sequences or alter the status of epigenetic marks, such as cytosine methylation. On the other hand, TEs may also increase expression of transcription factors (Table 1,2) and modify the gene structure (Fig 7).

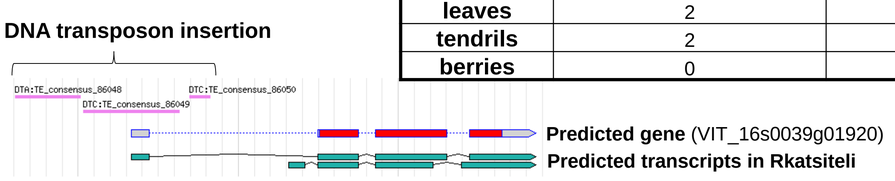

Fig 7. Predicted transcripts in Rkatsiteli, based on the presence or absence of the TE.

Table 1. Rkatsiteli allele-specific expression of the VIT_16s0039g01920 gene.

	TE absent (-)	TE present (+)
leaves	2	352
tendrils	2	714
berries	0	410

Table 2. FPKM levels of the VIT_16s0039g01920 gene in presence or absence of TE.

Variety		Sangiovese	Rkastiteli	Traminer	Kismish
TE genotype		-/-	-/+	-/+	+/+
FPKM	**leaves**	0.0	25.4	30.8	90.4
	tendrils	0.0	43.3	49.2	97.1
	berries	0.0	34.0	35.8	80.5

- absence of TE; **+** presence of TE.

5. PAN-GENOME ESTIMATION

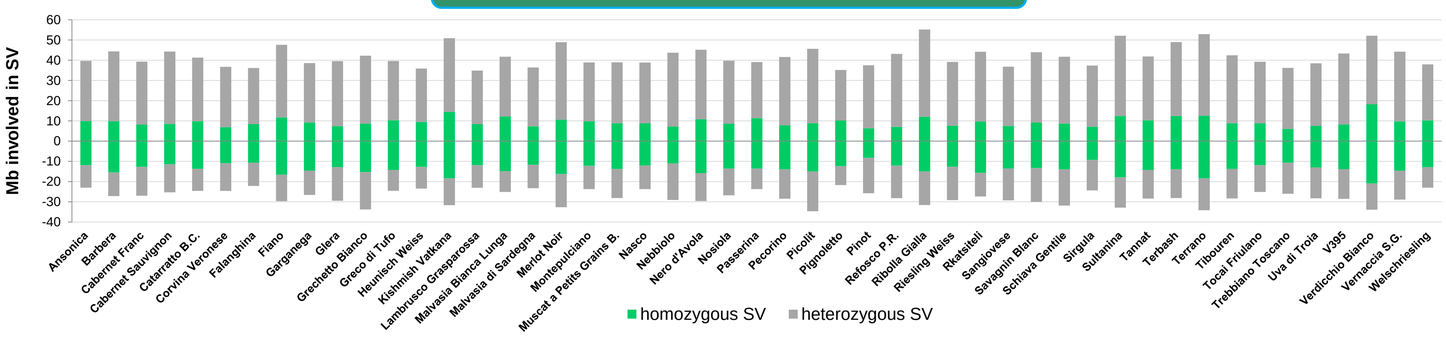

Fig 8. Heterozygous and homozygous SV extension (Mb). Negative values represent the Mb missing in a variety, while positive values the additional Mb.

Compared to the reference genome, each variety is characterised by a certain amount of base-pairs that are missing and, vice versa, by base-pairs that are in addition (Fig 8). Considering only the small SVs, we provided a first insight into the pan-genome size. Since grapevine is a diploid species, we considered dispensable only sequences completely missing in at least one individual of the species (Fig 9).

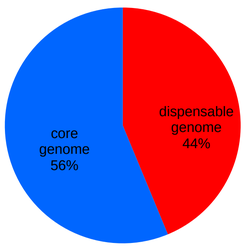

Fig 9. *Vitis vinifera* pan-genome estimation.

	Size (Mb)
Reference genome	486
Pangenome	752.11
Core Genome (CG)	423.06
Dispensable Genome (DG)	329.05

6. CONCLUSIONS AND OUTLOOKS

SVs are a very important source of genetic variation and widely contribute to the dispensable portion, which represents the 44% of the grape pan-genome. Not only intergenic regions were involved, but also gene space resulted affected by SVs (Fig 10). The movement of TEs contributed to a substantial increase of the genome size. Intronic regions were subjected to the widest expansion (46% of their size), followed by the intergenic space (31%). In the near future it would be of great interest to determine the phenotypic effects of SVs.

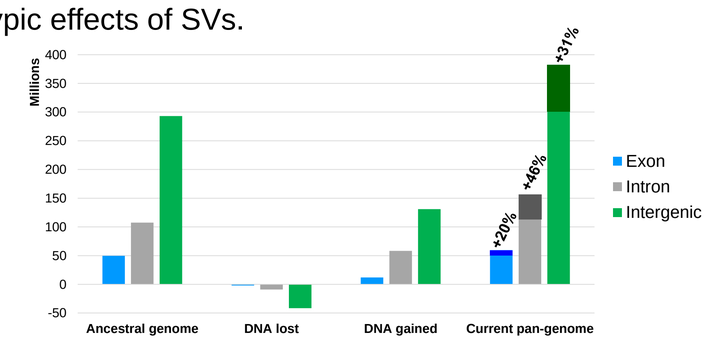

Fig 10. Genome expansion as a conseguence of TE expansion.

7. BIBLIOGRAPHY

[1] Tettelin et al. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: Implications for the microbial "pangenome". Proc Natl Acad Sci USA, 2005
[2] Morgante et al. Transposable elements and the plant pan-genomes. Curr Opin Plant Biol, 2007
[3] Marroni et al. Structural variation and genome complexity: is dispensable really dispensable? Curr Opin Plant Biol, 2014
[4] Jaillon, O. et al. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature, 2007.
[5] Rausch et al. DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics, 2012
[6] Sindi et al. A geometric approach for classification and comparison of structural variants. Bioinformatics, 2009
[7] Pinosio et al. Characterization of the Poplar Pan-Genome by Genome-Wide Identification of Structural Variation. Mol Biol Evol, 2016
[8] Vitulo et al. A deep survey of alternative splicing in grape reveals changes in the splicing machinery related to tissue, stress condition and genotype. BMC Plant Biology, 2014

Supported by: "Novel variation in plant breeding and the plant pan-genomes NOVABREED", Grant agreement no.: 294780, **European Research Council** (**ERC**)