Sequencing and *de-novo* Assembly of six grapevine cultivars

Michele Vidotto[1,2], Davide Scaglione[3], Gabriele Magris[1,2], Sara Pinosio[2,4], Giusi Zaina[1], Fabio Marroni[1,2], Gabriele Di Gaspero[1,2], Michele Morgante[1,2]

[1] Dipartimento di Scienze Agrarie e Ambientali, Università di Udine, Udine, Italy - [2] Istituto di Genomica Applicata, Udine, Italy
[3] IGA Technology Services, Udine, Italy - [4] CNR, Istituto di Genetica Vegetale, Sezione di Firenze, Sesto Fiorentino, Italy

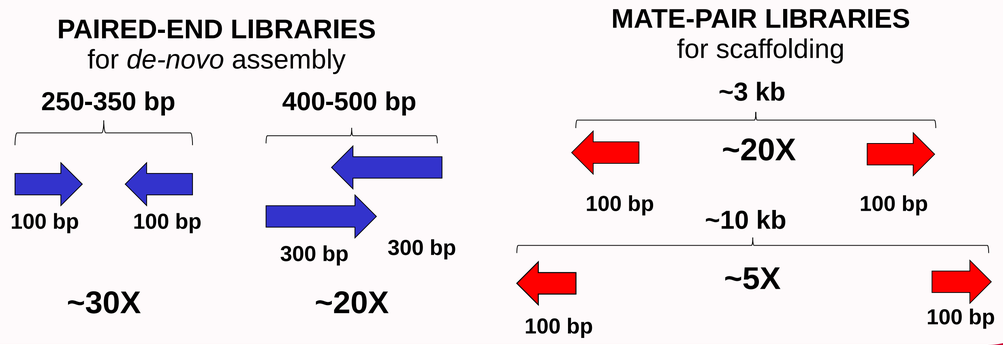
e-mail: *michele.vidotto@uniud.it*

1. INTRODUCTION

Genomes are characterized by high levels of structural variation, consisting of insertion/deletions, mostly due to recent insertions of transposable elements. Next-generation sequencing (NGS) allows re-sequencing the whole genome of several subjects to produce catalogs of structural variants (SVs), ultimately defining a species Dispensable Genome (DG) composed of partially shared and/or non-shared DNA sequence elements. The *Vitis vinifera* reference genome sequence of 485 Mb was obtained from PN40024, a highly inbred strain using a WGS approach with Sanger technology [1]. To detect those portions of the DG that are not present in the reference but that may be present in one or more individuals, we performed sequencing and *de-novo* assembly of 6 grapevine cultivars: Cabernet Franc, Gouais Blanc, Kishmish Vatkana, Rkatsiteli, Sangiovese and Traminer. Here we describe the procedure to obtain and validate the six genome assemblies and the comparison with the *V. vinifera* reference.

2. SEQUENCING

Sequencing machine: Illumina HiSeq2500; Illumina Miseq
***V. vinifera* genome size**: 485 Mb (diploid genome, 2n=19)



3. GENOME ASSEMBLY

Step 1: data preprocessing

Adapter sequences and low quality 3' ends were removed from raw reads using cutadapt [2] and ERNE-FILTER [3] respectively. The complexity of libraries, GC and base content were examined with FastQC [4]. The quality of the libraries was estimated by the k-mer distribution analysis (Figure 1). The tool Jellyfish [5] was used for the k-mers count. Reads were aligned to the *V. vinifera* reference genome with BWA [6] to estimate the achieved coverage, the duplication level and the insert size distribution.

Step 2: *de novo* assemblers comparison

Preliminary results showed that ALLPATHS-LG algorithm [7] given the appropriate mix of sequencing libraries at adequate coverage outperform the other algorithms by producing scaffolds with lower N50 and a greater L50. For this reason ALLPATHS-LG was used to assemble the 6 cultivars under study (table 1).

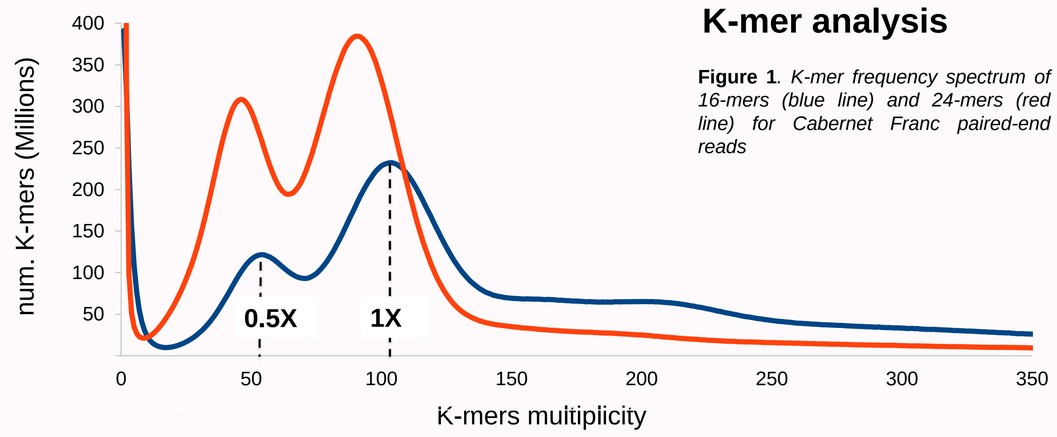


Figure 1. *K-mer frequency spectrum of 16-mers (blue line) and 24-mers (red line) for Cabernet Franc paired-end reads*

Table 1

Sample	Gouais Blanc		Cabernet Franc		Traminer		Kishmish Vatkana		Rkatsiteli		Sangiovese	
Assembler	**ALLPATHS-LG v. 48359**											
	contigs	**scaffolds**	**contigs**	**scaffolds**	**contigs**	**scaffolds**	**contigs**	**scaffolds**	**contigs**	**scaffolds**	**contigs**	**scaffolds**
Total (bp)	407,616,233	680,976,317	477,947,859	690,595,556	411,015,161	678,662,631	405,822,925	605,311,887	423,146,878	691,981,550	379,819,754	538,851,806
length improv.		67%		44%		65%		49%		64%		42%
Average (bp)	4,147	69,353	5,022	55,006	3,827	78,223	3,994	67,830	4,068	70,524	3,916	27,078
Max (bp)	182,376	3,087,250	222,515	1,641,804	137,674	2,817,510	109,311	2,959,549	158,393	2,334,202	131,077	3,422,443
Min (bp)	500	885	500	889	500	885	500	892	500	898	500	880
Sequences (#)	98,294	9,819	95,170	12,555	107,393	8,676	101,602	8,924	104,011	9,812	97,004	19,900
N50 (#)	10,090	588	9,056	961	12,267	484	12,010	514	11,788	553	10,436	514
N50 (bp/L50)	9,334	314,627	12,049	198,151	7,891	413,025	8,223	302,012	8,265	327,142	8,552	240,166

Table 1. *Assembly statistics from ALLPATHS-LG for the 6 cultivars under study. Contigs and scaffolds less than 500 bp in length were discarded.*

4. ASSEMBLY EVALUATION

Step 1: k-mers spectra comparison

The assemblies and the starting reads were decomposed into their component k-mers using Jellyfish [5]. The spectra were compared by their decomposed components related to copy number with the Kmer Analysis Toolkit (KAT) [8]. The sharp peaks at 0.5X (see figure 2) suggest the presence of high heterozygosity in each sample. The 0X component of histograms (black area) shows that ALLPATHS-LG was able to remove half of the heterozygous component and most of the errors from the starting data.

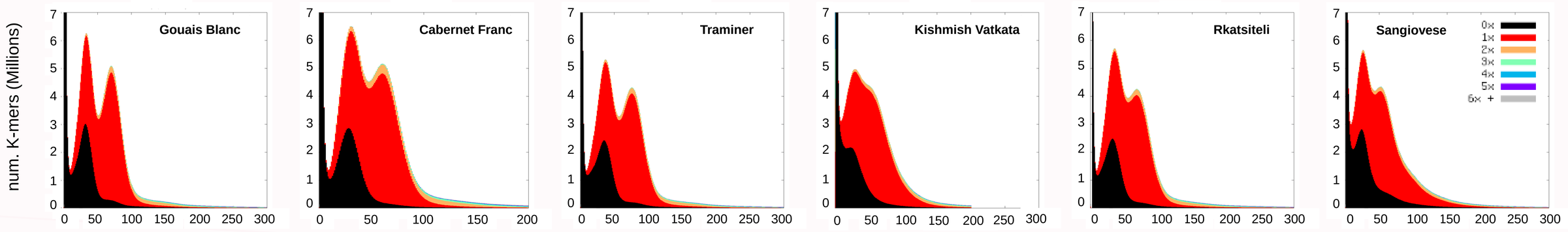


Figure 2. *Copy number spectra of assembled reads compared to the assembled contigs obtained with ALLPATHS-LG, at k-mer length 19 for the 6 samples.*

Step 2: alignment to the reference

The contigs of each assembly were aligned to the *V. vinifera* reference using DENOM [9]. A script that exploits both the contig alignments and the scaffolding information from the assembler, was developed to place scaffolds. We then estimated the fraction of genes, exons and repeats annotated in the *V. vinifera* reference, that are present in the placed scaffolds (table 2).

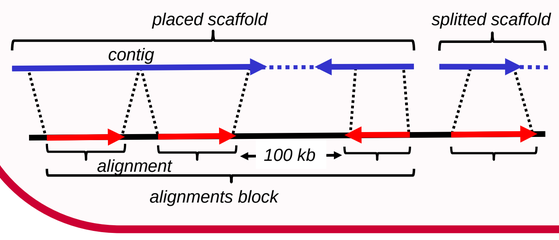


Table 2

	Cabernet Franc	Gouais blanc	Traminer	Kishmish	Rkatsiteli	Sangiovese
Placed scaffolds (#)	8,758	6,669	6,104	5,882	6,952	14,036
Anchored bp	537,828,400	540,482,725	546,673,623	469,984,893	543,200,320	427,312,604
Splitted scaffolds (#)	1,065	639	657	900	683	1,320
Unplaced scaffolds (#)	2,732	2,511	1,915	2,142	2,177	4,544
Unplaced bp	52,649,692	48,580,992	32,625,401	39,346,855	45,568,402	37,995,826
Total placed bp	90%	92%	94%	92%	92%	91%
Genome covered (bp)	411,800,306	399,822,352	436,070,441	425,621,682	432,784,394	413,363,890
Genes	89.99%	83.17%	90.72%	89.30%	90.08%	87.38%
Exons	95.74%	89.63%	96.19%	95.48%	95.74%	94.91%
Repeats	86.66%	79.45%	87.47%	83.88%	85.88%	80.15%

Table 2. *Statistics of scaffolds mapped on the V. vinifera reference (486,198,630 bp) and fraction of annotated genes, exons, and repeats covered by placed scaffolds.*

5. CONCLUSIONS AND OUTLOOKS

We assembled *de-novo* the genome of six heterozygous grape cultivars with ALLPATHS-LG, obtaining high accuracy and good assembly statistics even in complex genomic regions. The assemblies will be used to define the extent and composition of regions belonging to the *V. vinifera* dispensable genome as follows:

1. We will identify the transposable element categories that mostly contribute to the dispensable component
2. We will look at what type of genes are mainly present in CNV and PAV variants, according to the assembled genomes.

BIBLIOGRAPHY

[1] Jaillon, O. et al. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature 2007.
[2] Martin M, Cutadapt removes adapter sequences from high-throughput sequencing reads, EMBnet.journal 2011.
[3] http://erne.sourceforge.net.
[4] Schmieder, R. & Edwards, R. Quality control and preprocessing of metagenomic datasets. Bioinformatics 2011.
[5] Marcais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics 2011.
[6] Li H & Durbin R, Fast and accurate short read alignment with Burrows-Wheeler Transform, Bioinformatics 2009.
[7] Gnerre S et al., High-quality draft assemblies of mammalian genomes from massively parallel sequence data, PNAS 2011.
[8] https://github.com/TGAC/KAT
[9] http://mus.well.ox.ac.uk/19genomes/IMR-DENOM/#IMR

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