Improvement of the Grapevine genome assembly

Aurelie A. Canaguier, Simone Scalabrin, Marie-Christine Le Paslier, Eric Duchêne, Nacer Mohellibi, Aurélie A. Berard, Aurelie A. Chauveau, Jean-Michel Boursiquot, Gabriele Di Gaspero, Ludger Hausmann, et al.

IMPROVEMENT OF THE GRAPE GENOME ASSEMBLY

A. Canaguier[1], S. Scalabrin[2], M-C. Le Paslier[3], E. Duchêne[4], N. Mohelibi[5], A. Berard[3], A. Chauveau[3],D. Scaglione[2], C. Del Fabbro[2], F. Cattonaro[2], S. Vezzulli[9], J-M. Boursiquot[6], G. Di Gaspero[2], L. Hausmann[7], J.M. Martinez-Zapater[8], M. Morgante[2], A-F. Adam-Blondon[5]

[1] *INRA, UMR1165 URGV, 2 rue Gaston Crémieux, BP 5708, 91057 Evry, France*
[2] *IGA, via J.Linussio 51, 33100 Udine, Italy*
[3] *INRA, US1279 EPGV, CEA-IG/CNG, 2 rue Gaston Crémieux, BP 5724, 91057 Evry, France*
[4] *SVQV, INRA, 28 rue de Herrlisheim, BP 20507, 68021 COLMAR Cedex, France*
[5] *INRA, UR1164 URGI, route de Saint-Cyr, RD 10, 78026 Versailles, France*
[6] *Montpellier SupAgro, INRA UMR 1334 AGAP, 2 place Pierre Viala, 34060 Montpellier, France*
[7] *JKI, Institute for Grapevine Breeding Geilweilerhof, 76833 Siebeldingen, Germany*
[8] *ICVV, CSIC, UR, Gobierno de La Rioja, Madre de Dios 51, 26006, Logroño, Spain*
[9] *Fundazione Edmund Mach, San Michele Al Adige, Italy*

Summary. The current chromosome assembly of the grapevine reference genome is under improvement using two strategies. First several maps have been densified using SNP markers developed both from a classic Sanger re-sequencing of the parents and from the re-sequencing of a panel of diversity in the Vitaceae genome to develop a 18K genotyping chip (http://urgi.versailles.inra.fr/Species/Vitis/GrapeReSeq_Illumina_20K). This allowed to develop 6 parental genetic maps that were aligned on the scaffold of the genome assembly (www.ncbi.nlm.nih.gov/assembly/317318/; urgi.versailles.inra.fr/Species/Vitis/Data-Sequences). Second, mate paire sequences were generated from 2kb DNA fragments of *V. vinifera cv.* Kishmish vatkana and used for further scaffolding. We were able to improve from 85% to 89% the percentage of the ordered sequence along the chromosomes and to improve the overall orientation of the scaffolds along the chromosomes.

Acknowledgements: Grant Plant-KBBE-2008-GrapeReSeq ; the authors thank PS Paul Stephen Raj[2], M Ponnaiah[2] for help in result analysis.

1. Maps development

Six parental maps were developped using 3 segregating populations: Riesling x Gewürtztraminer (Ri x Gw), Syrah x Grenache (Sy x Gr) and Chardonnay x Bianca (Ch x Bi). The markers used were SSR markers (Cipriani et al 2011), SNP markers developed from Sanger re-sequencing (Vezzulli et al, 2008; Canaguier et al 2010) and, for the RixGw progeny, the SNP markers form the 18K grapevine chip (http://urgi.versailles.inra.fr/Species/Vitis/GrapeReSeq_Illumina_20K). Only informative markers were kept, especially from the 18K set (table 1). The main characteristics of the resulting genetic maps are described in table 2 and Fig. 1.

Table 1. Characteristics of the 6 parental maps

Map	Gewürtztraminer	Riesling	Syrah	Genache	Chardonnay	Bianca
nb individuals	**120**	**120**	**192**	**192**	**358**	**358**
Nb SSR locus	117	128	288	283	450	466
Nb SNP locus	750	831	152	94	40	59
Total Nb locus	**867**	**959**	**440**	**377**	**490**	**525**

Table 2. Common locus in the pairs of parental maps

	Gr	Ch	Bi	Ri	Gw
Sy	245	154	154	60	55
Gr		150	148	73	56
Ch			318	83	76
Bi				70	64
Ri					84

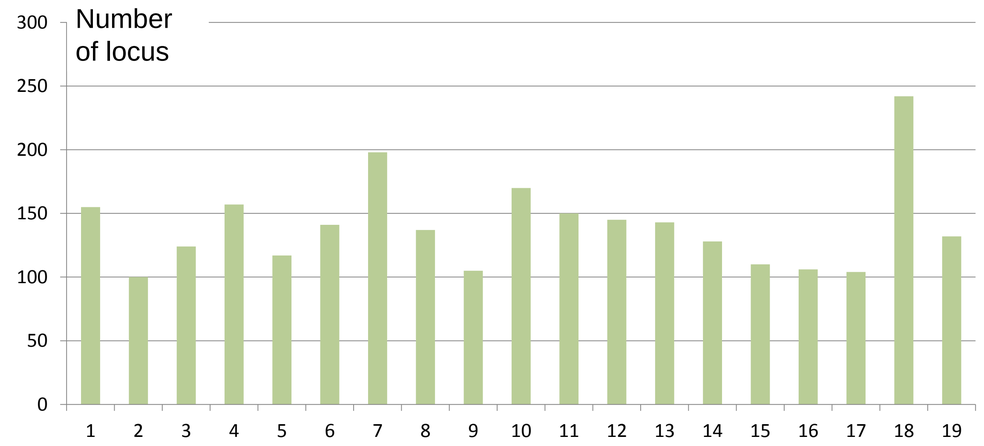

Fig. 1. Number of non redundant locus mapped on each grapevine chromosome using the three segregating populations

The 2664 non redundant markers were aligned on the scaffolds of the *V. vinifera* reference genome sequence (www.ncbi.nlm.nih.gov/assembly/317318/) by ePCR or BLAT.

2. Improvement of the scaffolding using mate pair sequences from *V. vinifera cv.* Kishmish vatkana

103,463,614 single Illumina 100bp reads from 51,731,807 inserts of average insert size 2kb were obtained from a single library of Kishmish vatkana. These reads were mate-pair produced using circularization by Cre-Lox recombination, the LoxP circularization linker was removed and used to classify reads with DeLoxer (http://nar.oxfordjournals.org/content/40/3/e24). Illumina adapter was removed using cutadapt (http://journal.embnet.org/index.php/embnetjournal/article/view/200). Quality trimming and contaminant removal was performed with erne-filter (Del Fabbro C et al, PlosOne, in press). Successively reads with highly duplicated kmers were removed with kmercounter (http://sourceforge.net/projects/kmercounter/). Then reads were aligned to the repeat masked reference genome with bowtie2. Reads not aligning at scaffold ends (max 5000bp from the ends) with mapping quality lower than 20, XM, XO and XG flags above, respectively 2, 1 and 4 were discarded with internally developed Perl scripts. Finally, alignments on scaffolds connected by multiple mate-pairs were visually inspected in order to discard further false positive alignments.
2,031 mate pairs respected above constraints and were used to join adjacent scaffolds.

3. V2 version of the chromosome assembly of the grapevine reference genome sequence

8% more of the genome sequence was ordered on grapevine chromosomes in the resulting V2 assembly (Fig. 2) and there is still a portion of the scaffofds which is poorly ordered especially on chromosomes 7, 10 and 16 (Fig. 3). The consortium decided to include these scaffolds at their best putative place instead of generating a chrX_random separate sequence like in the V0 version of the chromosomes assembly. The V2 chromosomes assembly therefore consists in 19 chromosome sequences (chr01 to chr19) and one chromosome random (chr00), which contains all scaffolds not mapped to a chromosome. 14% more of the total sequence is oriented in V2 compared to V0 and this improvement affects nearly all chromosome sequences (Fig. 3). The paire mate approach contributed importantly to this improvement of scaffold's orientation especially in regions splitted in many small scaffolds (86 scaffolds totalling 5Mb of sequence were oriented this way).. In the chromosome sequences finally generated, scaffold sequences are separated by 500N,. The multifasta file can be retrieved at (https://urgi.versailles.inra.fr/Species/Vitis/Data-Sequences/Genome-sequences).

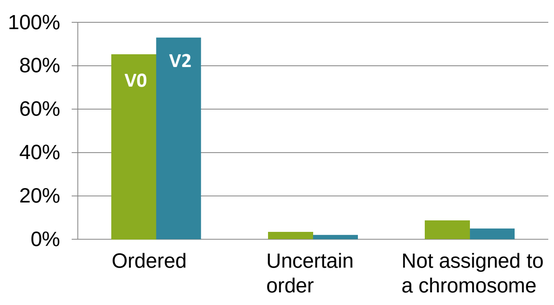

Fig. 2. Percentage of the genome sequence ordered on the chromosome, assigned but with uncertain order or not on a chromosome

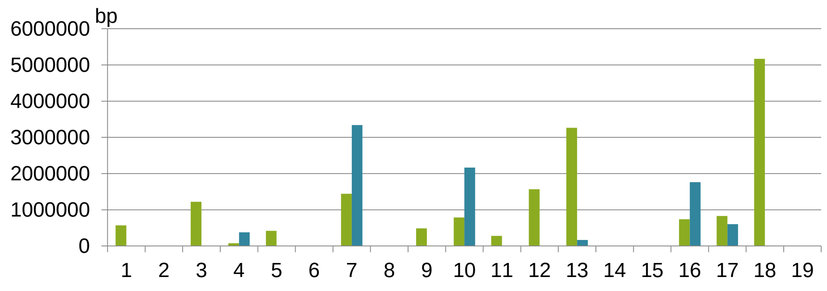

Fig. 3. Total size of the sequence scaffolds which order is uncertain for the 19 chromosomes in the V0 (green bars) compared to the V2 (blue bars)

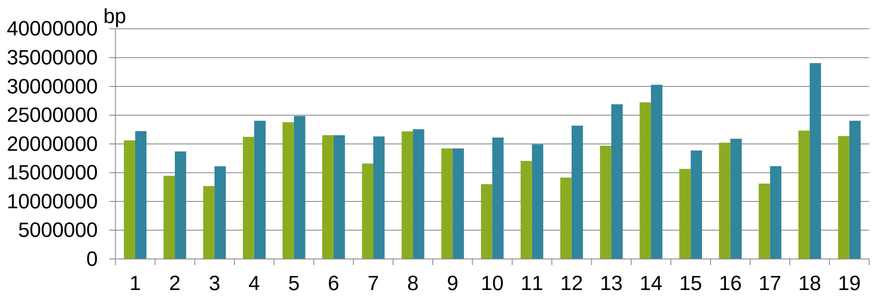

Fig. 4. Total size of the sequence scaffolds which are ordered and oriented for the 19 chromosomes in the V0 (green bars) compared to the V2 (blue bars)

[1] http://www.versailles.inra.fr/urgv
[2] http://www.appliedgenomics.org
[3] http://www.cng.fr/fr/organisation/laboratories/inra_epgv
[4] http://www6.colmar.inra.fr/svqv
[5] http://urgi.versailles.inra.fr
[6] http://umr-agap.cirad.fr/
[7] http://www.jki.bund.de/en/startseite/institute/rebenzuechtung.html
[8] http://www.icvv.es/