IMPROVEMENT OF THE GRAPE GENOME ASSEMBLY

A. Canaguier1, S. Scalabrini2, M.-C. Le Paslier1, E. Duchêne2, N. Mohelili2, A. Berard3, A. Chaveau1, D. Scaglione2, C. Del Fabbro2, F. Cattonaro2, S. Vezzulli2, J.-M. Bouriquiot2, G. Di Gaspero2, L. Hausmann2, J.M. Martinez-Zapater3, M. Morgante2, A.-F. Adam-Blondon2

1 INRA, UMR1165 URGV, 2 rue Gaston Crémieux, BP 5708, 91057 Evry, France
2 IGA, via J.Jinusso 51, 33100 Udine, Italy
3 INRA, US1279 EPGV, CEA-IGC/CNG, 2 rue Gaston Crémieux, BP 5724, 91057 Evry, France
4 SVQU, 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 ICV, CSIC, UR, Gobierno de La Rioja, Madre de Dios 51, 26006, Logroño, Spain
9 Fundazione Edmund Mach, San Michele Al Tagliamento, Italy

Summary. The current chromosome assembly of the grapevine reference genome is under improvement using different 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 Vitisaceae genome to develop a 18k genotyping chip (http://urgi-versailles.inra.fr/Species/Vitis/GrapeReSeq_Illinum_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; unguriservillesinra.fr/Species/Vitis/DataSequences). Second, mate pairs sequences were generated from 26k DNA fragments of V. vinifera cv. Kishmish vatkana and used for further assembly of the scaffolds along the chromosome.

1. Maps development

Six parental maps were developed using 3 segregating populations: Riesling × Gewürztraminer (RxG), Syrah × Grenache (SyG) and Chardonnay × Bianca (ChB). The markers used were SSR markers (Ciprani et al. 2011), SNP markers developed from Sanger re-sequencing (Vezzulli et al. 2008; Canaguier et al. 2010) and, for the RxG population, the SNP markers form the 1Kb SNP sequence chip (http://urgi-versailles.inra.fr/Species/Vitis/GrapeReSeq_Illinum_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

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Riesling</th>
<th>Syrah</th>
<th>Grenache</th>
<th>Chardonnay</th>
<th>Bianca</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of markers</td>
<td>120</td>
<td>120</td>
<td>192</td>
<td>192</td>
<td>358</td>
</tr>
<tr>
<td>Total size (bp)</td>
<td>120</td>
<td>120</td>
<td>288</td>
<td>283</td>
<td>450</td>
</tr>
</tbody>
</table>

100% V2 version of the chromosome assembly 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 Ilinum 100bp reads from 51,731,807 inserts of average insert size 2kb were obtained from single library of Kishmish vatkana. These reads were re-paired using the 18K genotyping chip (http://urgi-versailles.inra.fr/Species/Vitis/GrapeReSeq_Illinum_20K). Illumina adapters were removed using cutadapt (http://journal.ncbi.nlm.nih.gov/article/0129). Quality trimming and contaminant removal was performed with filter (Del Fabbro C et al. PlosOne, in press). Successively reads with highly duplicated markers were removed with mercount (http://sourceforge.net/projects/mercount/) 377 segregating markers 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, XQ 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 sequences was reordered grapevine chromosomes in the resulting V2 assembly (fig. 2) and there is still a portion of these scaffolds which is poorly ordered especially on chromosomes 7, 10 and 16 (fig. 3). The consortium decided to include these scaffolds as their best putative placeinstead of generating a chr_random separate sequence like in the V0 version of the chromosome 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 sequences oriented in V2 compared to V0 and this improvement affects nearly all chromosome sequences (fig. 3). The pair mate approach contributed importantly to this improvement of scaffold orientation especially in regions split up in many small scaffolds (86 scaffolds totalling 5Mb of sequence were orientated this way). In the chromosome sequences finally generated, scaffold sequences are separated by 500N. The multilasta file can be retrieved at (http://urgi-versailles.inra.fr/Species/Vitis/DataSequences/GenomeSequences).

Fig. 2. Percentage of the genome sequence ordered on the chromosome, assigned 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)