

A new version of the grapevine reference genome assembly (12X.v2) and of its annotation (VCost.v3)

Aurelie Canaguier, Jérome Grimplet, Gabriele Di Gaspero, Simone Scalabrin, Eric Duchêne, Nathalie Choisne, Nacer Mohellibi, Cécile Guichard, Stéphane Rombauts, Isabelle Le Clainche, et al.

▶ **To cite this version:**

Aurelie Canaguier, Jérome Grimplet, Gabriele Di Gaspero, Simone Scalabrin, Eric Duchêne, et al.. A new version of the grapevine reference genome assembly (12X.v2) and of its annotation (VCost.v3). Genomics Data, 2017, 14, pp.56-62. 10.1016/j.gdata.2017.09.002 . hal-01619926

HAL Id: hal-01619926
https://hal.science/hal-01619926

Submitted on 19 Oct 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NonCommercial - NoDerivatives 4.0 International License

Genomics Data 14 (2017) 56–62

Contents lists available at ScienceDirect

Genomics Data

journal homepage: www.elsevier.com/locate/gdata

Data in Brief

A new version of the grapevine reference genome assembly (12X.v2) and of its annotation (VCost.v3)

A. Canaguier[a,b], J. Grimplet[c], G. Di Gaspero[d], S. Scalabrin[d], E. Duchêne[e], N. Choisne[f], N. Mohellibi[f], C. Guichard[a,g], S. Rombauts[h,i], I. Le Clainche[a,b], A. Bérard[b], A. Chauveau[b], R. Bounon[a,b], C. Rustenholz[e], M. Morgante[d], M.-C. Le Paslier[b], D. Brunel[b], A.-F. Adam-Blondon[a,f,*]

[a] UMR GV, INRA, UEVE, ERL CNRS, 2 rue Gaston Crémieux, 91000 Evry, France
[b] EPGV US 1279, INRA, CEA, IG-CNG, Université Paris-Saclay, 91000 Evry, France
[c] Instituto de Ciencias de la Vid y del Vino (CSIC, Universidad de La Rioja, Gobierno de La Rioja), Logroño 26007, Spain
[d] IGA, via J. Linussio 51, 33100 Udine, Italy
[e] SVQV, UMR 1131, INRA, Université de Strasbourg, 28 rue de Herrlisheim, 68000 Colmar, France
[f] URGI, UR 1164, INRA, Université Paris-Saclay, route de Saint-Cyr, 78026 Versailles, France
[g] IPS2, UMR 1403, INRA, Université Paris-Saclay, Rue de Noetzlin, bât. 630, 91190 Gif-sur-Yvette, France
[h] Ghent University, Department of Plant Biotechnology and Bioinformatics, Technologiepark 927, 9052 Ghent, Belgium
[i] VIB Center for Plant Systems Biology, Technologiepark 927, 9052 Ghent, Belgium

ARTICLE INFO

Keywords:
Vitis vinifera
Genome
Chromosomes assembly
Gene annotation

Specifications	
Organism/cell line/tissue	*Vitis vinifera* cv. PN40024
Sex	Hermaphrodite
Sequencer or array type	The scaffold sequences were obtained by whole genome sequencing using the Sanger technology on ABI3730xl sequencers (Applied BioSystems) according to the supplementary information of Jaillon et al., Nature, 2007, 449: 463–468, doi: http://dx.doi.org/10.1038/nature06148. Genotype data were obtained from the GrapeReSeq 20K *Vitis* genotyping chip (https://urgi.versailles.inra.fr/Species/Vitis/GrapeReSeq_Illumina_20K) following the Infinium HD Assay Ultra Protocol (Ilumina Inc.). The *V. vinifera* cv. Kishmish vatkana mate pair sequences were produced using an Illumina HiSeq 2500 sequencer (Illumina Inc.).
Data format	Analyzed
Experimental factors	Three mapping populations were used: • 120 individuals derived from two reciprocal crosses between *V. vinifera* cv. Riesling cl.49 and *V. vinifera* cv. Gewürztraminer cl.643 (Ri × Gw) • 358 individuals derived from a cross between *V. vinifera* cv. Chardonnay and *Vitis* spp. 'Bianca' (Ch × Bi) • 192 individuals derived from two reciprocal crosses between *V. vinifera* cv. Syrah and *V. vinifera* cv. Grenache (Sy × Gr)
Experimental features	Grapevine reference genome assembly and annotation *V. vinifera* cv. Kishmish vatkana was used for the generation of mate pair sequences.
Consent	Creative commons non copy left (cc-by): the data can be freely re-used at the condition to cite its authors
Sample source location	The Ri × Gw and the Sy × Gr populations were maintained in experimental units of the Institut

* Corresponding author at: UMR GV, INRA, UEVE, ERL CNRS, 2 rue Gaston Crémieux, 91000 Evry, France.
E-mail address: anne-francoise.adam-blondon@inra.fr (A.-F. Adam-Blondon).

http://dx.doi.org/10.1016/j.gdata.2017.09.002
Received 25 July 2017; Received in revised form 13 September 2017; Accepted 15 September 2017
Available online 18 September 2017
2213-5960/ © 2017 Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/BY-NC-ND/4.0/).

National de la Recherche Agronomique (INRA), respectively the Service Experimentation Agricole et Viticole (Colmar, France) and the Domaine de Vassal (Marseillan-Plage, France). The Ch × Bi population and the *V. vinifera* cv. Kishmish vatkana variety (VIVC no. 6277) were maintained in the germplasm collection of the University of Udine at the Experimental Farm A. Servadei (Udine, Italy).

1. Direct link to deposited data

http://doi.org/10.15454/1.4962347083032307E12.
http://doi.org/10.15454/1.5009072354498936E12.

2. Introduction

The grapevine reference genome was published by Jaillon et al. [1]. The sequence for the first version of the genome, called the 8X version, was obtained using a whole genome shotgun strategy and the Sanger sequencing technology and was assembled from reads representing 8X coverage. Soon after, the assembly was improved through the addition of 4X of additional coverage, including more Bacterial Artificial Chromosome end sequences that greatly improved the scaffolding of the sequence contigs [2,3]. The corresponding scaffolds and raw sequences were deposited in European Molecular Biology Laboratory (EMBL) archives (FN594950-FN597014, 2065 entries, release 102). A new chromosome assembly was also developed, based on an improved version of the maps used for the 8X genome version [2–5] and was also archived at EMBL (FN597015-FN597047, 33 entries, release 102): it is referenced in the grapevine community as the 12X.v0 version of the grapevine reference genome. The chromosome sequence scaffolding of this version still necessitated improvements as around 9% of the sequence was not anchored to chromosomes (with the corresponding scaffolds stacked in the “Unknown” chromosome) and 3.5% of the sequence could be assigned to a chromosome but without certain placement and orientation within the chromosome (stacked in additional “random” chromosomes). The chromosome assembly of the grapevine reference genome was therefore further improved using two strategies. First, six parental maps were saturated with SNP markers developed with different strategies. Second, a collection of mate paired sequences generated from 2 kb DNA fragments of *V. vinifera* cv. Kishmish vatkana was used for further scaffolding. This allowed producing the 12X.v2 version of the grapevine genome assembly presented here.

All these versions of the genome assembly have been accompanied by an automatic gene annotation. The annotation for the original 8X genome release included 30,434 genes predicted with the GAZE software [6]. For the 12X genome assembly, two versions of the annotation were distributed with the 12X.v0 release: the v0 version of the annotation was obtained with the GAZE software and the v1 version (CRIBIv1, 29,971 genes) was the result of the union of v0 and a gene prediction performed with the JIGSAW software [7]. Later, an update of the CRIBIv1, focused on the discovery of the splicing variants, was published by the same group [8]. Finally, National Center for Biotechnology Information (NCBI) Refseq released its own version of the gene prediction (27,043 putative genes) as for most of the species with published genomes. The NCBI Refseq was produced with the Gnomon-NCBI eukaryotic gene prediction tool [9]. For the 12X.v2 version of the genome assembly, an annotation was performed in the frame of the European Cooperation in Science and Technology project FA1106 (VCost) using the EUGENE software [10] and generating 33,568 genes. The design of this latter version was under the supervision of the Super-Nomenclature Committee for Grape Gene Annotation of the International Grapevine Genome Program (IGGP, www.vitaceae.org) fitting its recommendation for the gene nomenclature. The annotation initiatives by families that fitted these recommendations were integrated dynamically to the VCost annotation by curating their respective gene models when needed. So far, the following gene families were integrated to this annotation: the terpenoid synthase gene family [11], the stilbene synthases [12], the MADS box [13], the GRAS [14] and the MYB [15] transcription factors families. Here we describe the generation of the VCost.v3 version of the 12X.v2 version of the grapevine genome assembly, based on a comparison and merging of the NCBI-Refseq, VCost and CRIBIv1 annotations and a semi-manual curation and following the recommendations of the IGGP.

3. Materials and methods

3.1. Plant material

Three mapping populations were used to develop high density genetic maps: (i) a population of 120 individuals derived from two reciprocal crosses between *V. vinifera* cv. Riesling cl.49 and *V. vinifera* cv. Gewürztraminer cl.643 (Ri × Gw) and maintained at the experimental unit Service Experimentation Agricole et Viticole of the Institut National de la Recherche Agronomique (INRA, Colmar, France), (ii) a population of 358 individuals derived from a cross between *V. vinifera* cv. Chardonnay and *Vitis* spp. ‘Bianca’ (Ch × Bi) and obtained at Experimental Farm A. Servadei of the University of Udine but no longer maintained, (iii) a population of 192 individuals derived from two reciprocal crosses between *V. vinifera* cv. Syrah and *V. vinifera* cv. Grenache (Sy × Gr) maintained at the experimental unit Domaine de Vassal (INRA, Marseillan-Plage, France).

3.2. Genotyping the Ch × Bi, Sy × Gr and Gw × Ri populations

The development of a first version of the Ch × Bi and Sy × Gr parental maps is described in Cipriani et al. [4] and Canaguier et al. [5]. Possible errors in segregation data were carefully manually reviewed in these maps and their subsequent revised versions [dataset] [16] were used to generate the chromosome assembly presented in this data paper.

For the Gw × Ri maps, total DNA was extracted with Qiagen DNeasy Plant Maxi Kit (Qiagen, Hilden, Germany), according to the manufacturer's instructions except that 1% of polyvinylpyrrolidone (PVP 40,000) and 1% of β-mercaptoethanol were added to the AP1 buffer. DNA was quantified with Quant-it Picogreen dsDNA Assay Kits (InVitrogen, Life Technologies). The samples were normalized at 50 ng/µl in 96-well plates. Genotype data were obtained from the GrapeReSeq 20K Vitis genotyping chip (https://urgi.versailles.inra.fr/Species/Vitis/GrapeReSeq_Illumina_20K) following the Infinium HD Assay Ultra Protocol (Ilumina Inc., San Diego, CA, USA). Data were analyzed using the Genotyping Module V1.9.4 of Illumina's Genome Studio® software (Illumina Inc., San Diego, CA, USA). After genotyping quality check and automatic clustering the SNP allele callings were manually inspected and edited and the parental maps were generated from the data using the R/qtl software [17].

3.3. Mate pair sequencing and alignment on the scaffolds of the grapevine genome assembly

Illumina mate-pair reads were produced using circularization by Cre-Lox recombination. The LoxP circularization linker was removed and used to classify reads with DeLoxer [18]. Illumina adapter was removed using Cutadapt [19]. Quality trimming and contaminant removal was performed with erne-filter [20]. Reads with highly duplicated kmers were removed using Kmercounter (http://sourceforge.net/projects/kmercounter/). Reads were aligned to the repeat masked reference genome using the software bowtie2 [21]. Reads not aligning at scaffold ends (max 5000 bp from the ends), with mapping quality lower than 20, or 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 to discard further false positive alignments. Mate pairs were deposited in the NCBI Short Read Archive under the accession number SRR5712111.

3.4. Assembly of the chromosomes

Chromosome assembly was achieved in three steps. First, all markers were aligned on the scaffolds of the 12X genome assembly (FN594950-FN597014, EMBL release 102) by Blat [22] and ePCR [23] according to Jaillon et al. [1]. A first ordering was generated based on these results and taking into account only the parental maps. Then, junctions between adjacent scaffolds were confirmed using mate pair information. Only the scaffolds with multiple evidence of correct ordering (anchoring by at least two maps or at least one map and a mate pair junction) were retained in the assembly. Mate pair information was also used for orienting scaffolds. Finally, all the scaffolds tentatively placed at the extremities of the chromosomes were manually inspected for the presence of telomere repeats. This allowed also confirming the anchoring of these scaffold and sometimes to correct or confirm their orientation.

3.5. Development of the VCost.v3 version of the Vitis genome annotation

3.5.1. Dataset collection

The CRIBIv1, the NCBI Refseq (NCBI *Vitis vinifera* Annotation Release 101: https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Vitis_vinifera/101/) and the VCost annotation were collected. CRIBI v1 and Refseq were developed on the grapevine genome 12X.v0 while the VCost version was developed already on the 12X.v2 using the EUGENE software. In addition, the gene models predicted by GAZE software in the 8X assembly and by ESTs, used by Grimplet et al. [24], but absent from the CRIBI v1 annotation were used for validation of the models but were not considered in the final VCost.v3 annotation because they correspond to truncated, non-functional genes. The CRIBIv1 gene track includes 29,971 gene models, the Refseq one 27,043 gene models and the VCost one 33,568 models. Algorithm and method for annotations were described in Thibaud-Nissen [25] for Refseq, Foissac et al. [10] for the VCost and in Vitulo et al. [8] for the CRIBIv1.

Manually expert-based curated gene families were also mapped on the 12X.v2 genome version: the terpenoid synthases [11], the stilbene synthases and chalcone synthase [12], the MADS box [13], the GRAS [14] and the MYB [15] transcription factors.

3.5.2. Remapping of genes on the grapevine genome V2

CRIBIv1 and Refseq automatic annotations and the expert-based curated gene models were all transposed from genome sequence V0 to V2 using a homemade python script (free source code available at https://github.com/timflutre/VitisOmics/blob/master/src/transferAnnot_from_Vitis_12X_V0_to_V2.pl): since the 12X.v2 assembly was an improvement of the ordering of the scaffolds already used in the 12X.v0 assembly [5], the positions of the features could be deduced from the new position of the scaffolds on the V2 chromosomes (Fig. 1). A JBrowse (http://jbrowse.org/, version 1.11.5) was set up to visualize and give access to these results (https://urgi.versailles.inra.fr/jbrowse/gmod_jbrowse/?data=myData/Vitis/data_gff).

3.5.3. Comparison of annotations and definition of a unique set of gene models

The position of the gene models from the three annotations was compared with a homemade Perl script and overlapping models were grouped together for further analysis.

For each gene, a Blast search was performed against plant protein sequences of the UniProt database except sequences from the *Vitis*

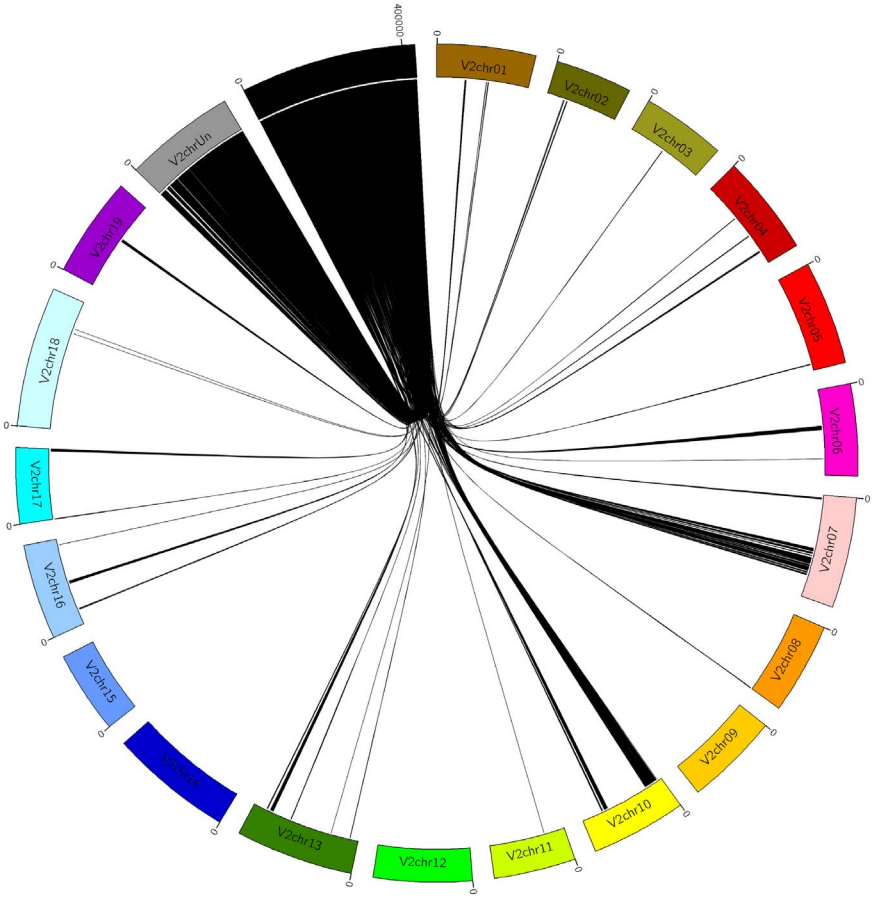


Fig. 1. Circular diagram of the transposition of the scaffolds from the unknown chromosome of the 12X.v0 genome assembly (black) to the chromosomes in the 12X.v2 assembly.

Table 1
Number of loci from the different categories of markers in the six parental maps.

Map	Gw	Ri	Sy	Gr	Ch	Bi
SSR	117	128	288	283	450	466
SNP	750	831	152	94	40	59
Total	867	959	440	377	490	525

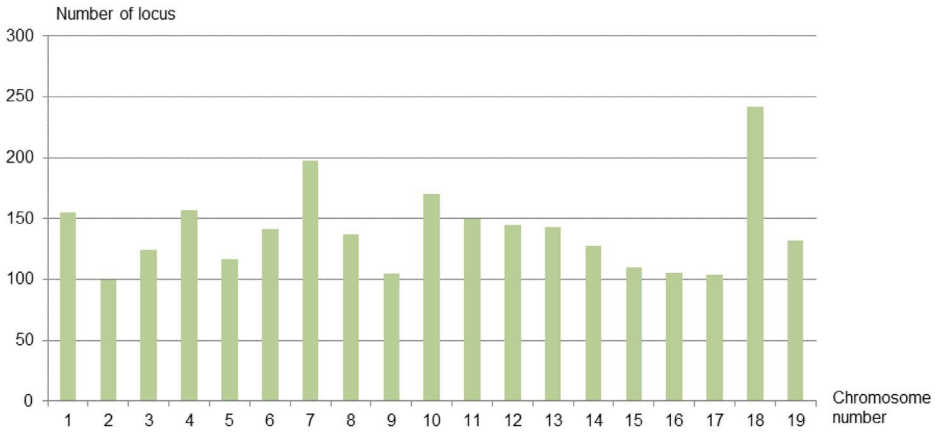


Fig. 2. Number of non-redundant loci mapped on each grapevine chromosome using the three segregating populations.

Table 2
Number of common loci in each pair 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

genus to avoid self-matching. The 30 best hits with an e-value lower that 1e-20 were kept for further analysis. Two indicators of quality were collected for each gene model: (i) the number of alignments showing an overlapping region of the subject (hit) sequence > 90% (hit overlap value: HO) and (ii) the number of alignments where the overlapping region of the query was > 90% (Query Overlap value: QO). High values of both HO and QO means that the exact structure of the grapevine gene model is frequently found in other species and is likely valid. If the HO number is low and the QO is high, a part of the correct sequence is probably missing in the annotation. If the QO is low and the HO is high, the gene models or known genes from the other plants do not fully cover the grapevine gene model, which may indicate a chimera in the annotation. When both values are low, or in case that there is no hit, the homology only occurs at best on portions of the gene models (subject and query) and keeping the grapevine gene model in the final annotation is questionable. It is important to note that the grapevine coding sequences might not have the same size than in other species but if high HO and high QO were observed for a grapevine gene model from an annotation, this model was preferred over alternative models with lower HO/QO value for inclusion in the final annotation.

If a gene model was only predicted in a single annotation, the locus was added to the final gene set with no further discriminative analysis.



Fig. 3. Percentage of the genome sequence (i) ordered on the 19 grapevine chromosomes in the current version of the assembly (12X.v0, in green) and in the new version (12X.v2, in blue), (ii) assigned to a chromosome but with uncertain order or (iii) not assigned to any chromosome.

If a gene model was predicted by two of the three annotations, the one with the highest HO and QO (> 90%) was chosen in the final set. When a gene model showed equivalent HO and QO scores in more than one annotation, the CRIBI V1 was favored over the VCost that was favored over the Refseq annotation. The main reason to do so, was that the CRIBI V1 was the most widely used version of annotation by the grapevine community, in particular in many published transcriptomic studies. The expert-based manually curated gene models were kept in preference to all the automatic annotations.

3.5.4. Specific case of split or merged gene models

Gene prediction methods can produce inaccurate models resulting in wrong split or merged versions of the actual genes. When such an error occurs in one annotation and not in the others, several genes from each annotation will belong to the same group. These groups were carefully visually inspected with the support of the IGV program [26] to visualize the gene structures from all the annotations. The sequence likely to be correct was conserved. If interpretation was still conflictive, shorter, possibly incomplete structures were favored over longer, possible chimeric, structures.

3.5.5. Construction of the final set of gene models of the Vcost.v3 annotation

Features from conserved gene models for each of the three annotation sets were extracted from their respective initial GFF file and merged into one single GFF file. Feature structure from the three automatic annotations and the six manually curated gene families were standardized and a Locus ID was allocated to each gene following the recommendations of Grimplet et al. [27]. Finally, a file containing both the new sequence and the V3 annotation was prepared at the GenBank sequence format [dataset] [28].

4. Results

4.1. Development of six parental genetic maps

Six parental maps were developed using three segregating populations, Ri × Gw, Sy × Gr and Ch × Bi, and 2664 non redundant loci. The markers used were SSR markers [4], SNP markers developed from

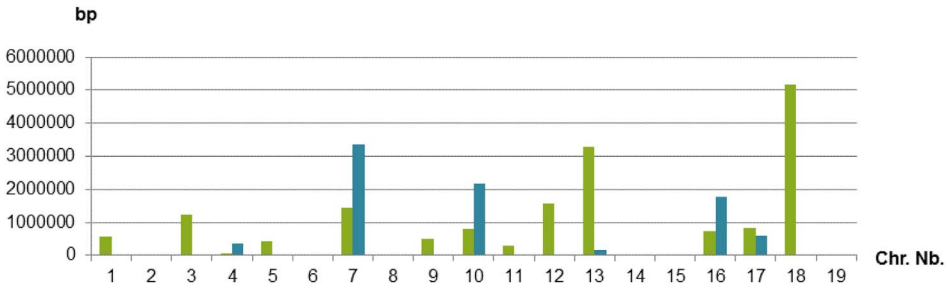


Fig. 4. Total size of the sequence scaffolds which order is uncertain for the 19 chromosomes in the 12X.v0 (green bars) compared to the 12X.v2 (blue bars) versions of the grapevine reference genome sequence.

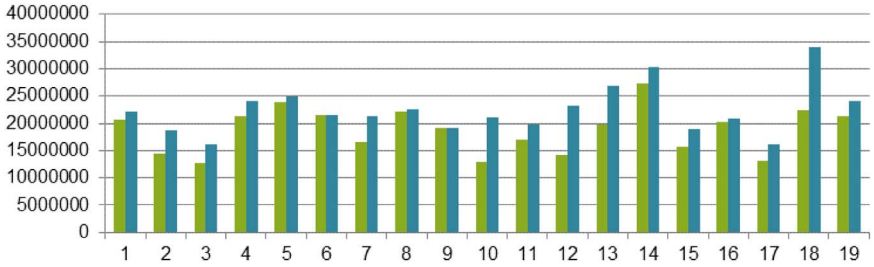


Fig. 5. Total size of the scaffolds which are ordered and oriented for each of the 19 chromosomes in the 12X.v0 version of the grapevine genome assembly (green bars) compared to the 12X.v2 version (blue bars).

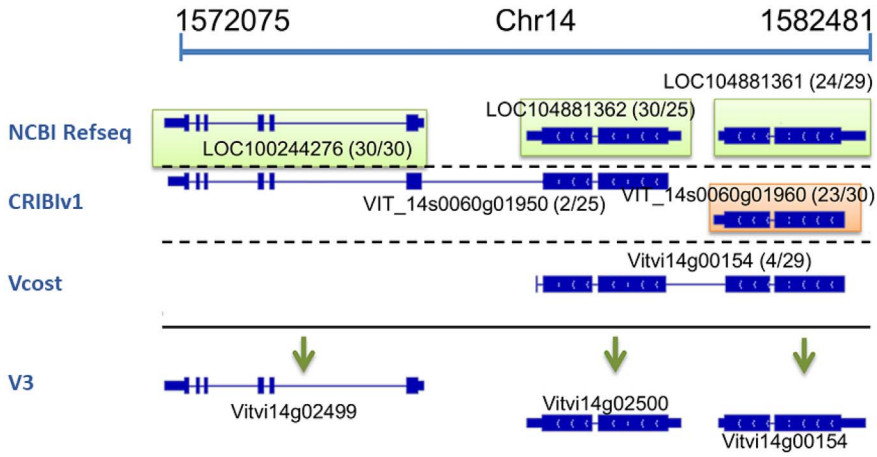


Fig. 6. Example of alignment between the gene models from the 3 annotations showing chimeric genes for pectinesterases genes. In brackets HO/QO scores.

Sanger re-sequencing [5] and for the Ri × Gw progeny, 1580 SNP markers from the 20K grapevine chip. The distribution of the different type of markers in the different maps is described in Table 1.

The mapped loci were quite well distributed across the chromosomes: from a 100 loci for the less covered (chromosome 2) to 242 loci for the most covered (chromosome 18; Fig. 2).

The common markers between the maps mainly corresponded to SSR markers (Table 2). These common markers were particularly important to obtain the relative order the contigs anchored in each individual parental maps.

The maps and description of the markers are available at [dataset] [16].

Table 3
Correspondence between gene models within the 3 annotations. In brackets possible occurrence in CRIBI V1, VCost and Refseq respectively.

	Before manual analysis	After curation
In only one annotation (1/0/0)	**17,325**	**16,444**
In 2 annotations (1/1/0)	**6535**	7555
In 3 annotations (1/1/1)	**13,233**	15,288
Group with multiple genes (ex:2/3/1)	**5761**	**3127**
Total	**42,854**	**42,414**

4.2. Development of the 12X.v2 chromosome assembly

The 2664 non redundant markers were aligned on the scaffolds of the *V. vinifera* reference genome sequence, resulting in a first draft assembly of the chromosomes. A total of 103,463,614 Illumina 100-bp reads were generated from 51,731,807 inserts of average 2 kb size from a single library of *V. vinifera* cv. Kishmish vatkana. These reads were aligned on the scaffolds sequence extremities of the *V. vinifera* reference genome sequence in order to generate links between scaffolds. The alignments were manually inspected, taking into account the data obtained from the genetic maps and resulting in the selection of 2031 mate pairs that joined adjacent scaffolds.

The combination of these two layers of information together with a manual check of the presence of telomeric repeats at the extremity of the chromosomes allowed developing the 12X.V2 chromosome assembly [dataset] [16]. It consists of 19 grapevine chromosomes containing 366 scaffolds totaling 458,641,822 bp. An additional 2,654,308 bp pseudomolecule, named chr00, consists of the remaining 1692 unanchored scaffolds. Compared to the previous version, 8% of unassigned genome sequence is ordered along grapevine chromosomes in the resulting V2 assembly (Fig. 3), although there is still a small portion of the scaffolds which is ordered with some degree of uncertainty, especially on chromosomes 7, 10 and 16 (Fig. 4).

The International Grapevine Genome Program consortium decided to insert these scaffolds at their most likely intra-chromosomal location

Table 4
Correspondence of gene models between the three versions of automatic annotation. In bold, the gene models specific of each of them. In blue: gene models appearing in two annotations. In brown, models that were split in the V1. In purple, models that were split in the VCost. In green, models that were split in the Refseq. Yellow: models for which not a single gene model from one annotation was conserved in the final set (0 or many genes in each annotation).

VCost	V1 / Refseq	0	1	2	3	4	5	6	7	8	9	10
0	0	32	**3948**	3	0	1	0	0	0	0	0	0
0	1	**2658**	4265	129	10	3	2	0	0	0	0	0
0	2	11	86	2	2	0	1	0	0	0	0	0
0	3	1	4	0	0	1	0	0	0	1	0	0
1	0	**9831**	2153	82	8	2	0	1	0	0	0	0
1	1	1137	15,288	497	55	17	5	5	1	3	0	1
1	2	13	220	22	5	3	0	0	0	0	0	0
1	3	0	11	0	1	0	1	0	0	0	0	0
1	4	0	1	1	0	0	0	0	1	0	0	0
2	0	2	119	1	0	0	0	0	0	0	0	0
2	1	82	1116	119	19	8	0	1	0	0	0	0
2	2	2	45	3	0	0	0	0	0	0	0	0
2	3	1	2	0	0	0	0	0	0	0	0	0
3	0	0	16	0	0	0	0	0	0	0	0	0
3	1	13	191	35	7	0	1	0	0	0	0	0
3	2	0	16	2	0	0	0	0	0	0	0	0
3	3	0	3	0	0	0	0	0	0	0	0	0
4	1	2	39	9	2	0	0	0	0	0	0	0
4	2	0	0	2	0	0	0	0	0	0	0	0
4	3	0	1	0	0	0	0	0	0	0	0	0
5	0	0	1	0	0	0	0	0	0	0	0	0
5	1	0	11	0	0	1	0	0	0	0	0	0
5	2	0	1	0	0	0	0	0	0	0	0	0
5	3	0	0	1	0	0	0	0	0	0	0	0
6	0	0	1	0	0	0	0	0	0	0	0	0
6	1	0	1	1	1	0	0	0	1	0	0	0
7	1	0	1	0	0	0	0	0	0	0	1	0
7	2	0	0	0	0	1	0	0	0	0	0	0
8	1	0	0	1	0	1	0	0	0	0	0	0
17	1	0	1	0	0	0	0	0	0	0	0	0
39	1	1	0	0	0	0	0	0	0	0	0	0

instead of generating a chrX random pseudomolecule, as we did in the v0 version of the chromosomes assembly. The v2 chromosome assembly therefore consists of 19 chromosome sequences (chr01 to chr19) and one chromosome random pseudo-molecule (chr00). The AGP (Assembly Golden Path) of the chromosomes and the level of uncertainties are described in details in [dataset] [16].

The 12X.v2 assembly contains more oriented sequence than the 12X.v0 (+14%) and nearly all chromosome sequences benefit from this improvement (Fig. 5). The pair-mate approach contributed importantly to the improvement of the orientation of the scaffolds in the new assembly, confirming the orientation of 75 scaffolds (156.8 Mb) and allowing the orientation of 90 scaffolds (5.3 Mb). This improvement was especially important in regions covered by many small scaffolds.

4.3. Development of the VCost.v3 version of the grapevine reference genome annotation

An initial blast comparison between the three sets of gene models proposed by the three gene annotations generated 5761 groups containing multiple genes from each of the annotations. The structure of each group was very specific and it was not possible to define an automatic procedure to properly identify the correct gene models within each group. In order to standardize the selection criterion, we defined indicators for each gene taking into account the occurrence of similar gene model in public database based on alignment with proteins from other plant species: the HO and QO described in the material and methods. As an example, Fig. 6 represent a group of adjacent pectinesterase that has been concatenated into chimeras in some annotations.

We observed that the 3 gene models from Refseq (LOC100244276 (30/30), LOC104881362 (30/25), LOC104881361 (24/29)) and one gene model from the CRIBIv1 (VIT14s0060g01960 (23/30)) showing high HO/QO scores whereas the VIT14s0060g01950 (2/25) and Vitvi14g00154 (4/29) models from the VCost did not, for both there are few genes in other species that fully overlap the *Vitis* sequence. These two gene models were likely chimeras from 2 artificially assembled coding sequences corresponding to the Refseq gene models. Besides, predicted proteins for LOC104881362 VIT14s0060g01960 were identical but LOC104881362 was retained in the final set over VIT14s0060g01960 because it contained a longer UTR on both sides.

Nine hundred and seventy gene models out of the 5761 could be chosen for the final set only based on the HO/QO scores. The other groups were visually inspected with IGV. Many groups contained more than one true gene model which were curated and split into smaller groups, leading to in an increase of genes appearing in 2 or 3 annotations (Table 3). The sequences from the versions older than Cribi v1 (8X, or EST) that did not overlap gene models, were removed because they did not correspond to functional gene models or because there was no proof of actual expression. The final set of putative genes contained 42,414 gene models. Nearly half of them however only appeared in one

single annotation, while 15,288 were constantly predicted in all 3 annotations.

A detail of the distribution of the genes models within groups is presented in Table 4. VCost was the version of annotation with the highest number of unique gene models (9831), many of these genes were very short and their existence needed to be confirmed. On the opposite, there were only 2665 Refseq specific gene models. The number of groups, for which not a single gene model from one annotation was conserved in the final set (0 or many genes in each annotation, in yellow in Table 4) was drastically reduced after curation. Among the remaining groups, two distinct cases could be distinguished. The most frequent case consisted of multiple gene models from the Refseq annotation overlapping on each other (the two other annotations algorithms did not allow overlapping). In that case, the largest gene was conserved: we only observed small gene models included in larger ones and never overlapping portions of different models. The other case consisted in genes from the families that were manually curated that were split in an annotation and not detected in the others.

Acknowledgements

This work was supported by the French National Institute for Agriculture (INRA, France), the University of Udine and the Institute of Applied Genomics (Italy), the Vlaams Institut voor Biotechnologie and the University of Ghent (Belgium), the Istituto de Ciencas de la Vid et del Vino (Logroño, Spain) and several grants: ANR-Plant-KBBE-2008-GrapeReSeq and ANR-2008-Muscares funded by the French National Research Agency (ANR), Valorizzazione dei Principali Vitigni Autoctoni Italiani e dei loro Terroir (Vigneto, no. COSVIR27129) funded by the Italian Ministry of Agriculture and the COST action FA1106 funded under the European FP7 Research Program. The authors thank the CEA-IG/CNG for allowing them to perform the DNA QC in its DNA and Cell Bank service and for providing access to their Illumina Genotyping Platform. The authors are grateful to Séverine Gagnot for developing an easy-access ePCR tool, to Manel Merimèche for her help in the setting up of a JBrowse allowing one to visualize the 12X.v2 genome and all the features mapped on it, to Nicoletta Felice, Giusi Zaina, Irena Jurman and Federica Cattonaro for the production of mate pair libraries and for Illumina sequencing, and to Gabriele Magris for sequence submission to short read archive. Finally, they warmly thank Jens Keilwagen for his detection of errors in the first gff releases of some expert-based curated genes.

References

[1] O. Jaillon, J.-M. Aury, B. Noel, A. Policriti, C. Clepet, A. Casagrande, N. Choisne, S. Aubourg, N. Vitulo, C. Jubin, A. Vezzi, F. Legeai, P. Hugueney, C. Dasilva, D. Horner, E. Mica, D. Jublot, J. Poulain, C. Bruyere, A. Billault, B. Segurens, M. Gouyvenoux, E. Ugarte, F. Cattonaro, V. Anthouard, V. Vico, C. Del Fabbro, M. Alaux, G. Di Gaspero, V. Dumas, N. Felice, S. Paillard, I. Juman, M. Moroldo, S. Scalabrin, A. Canaguier, I. Le Clainche, G. Malacrida, E. Durand, G. Pesole, V. Laucou, P. Chatelet, D. Merdinoglu, M. Delledonne, M. Pezzotti, A. Lecharny, C. Scarpelli, F. Artiguenave, E. Pé, G. Valle, M. Morgante, M. Caboche, A.-F. Adam-Blondon, J. Weissenbach, F. Quétier, P. Wincker, The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla, Nature 449 (2007) 463–468, http://dx.doi.org/10.1038/nature06148.

[2] A.-F. Adam-Blondon, O. Jaillon, S. Vezzulli, A. Zharkikh, M. Troggio, R. Velasco, Genome sequence initiatives, in: A.-F. Adam-Blondon, J.M. Martinez-Zapater, Chittaranjan Kole (Eds.), Genetics, Science Publishers and CRC Press, Genomics and Breeding of Grapes, 2011, pp. 211–234.

[3] A.F. Adam-Blondon, B.I. Reisch, J. Londo (Eds.), Grapevine genome update and beyond, X International Conference on Grapevine Breeding and Genetics, Geneva, August 2010, 1046 Acta Horticulturae, 2014, pp. 311–318.

[4] G. Cipriani, G. Di Gaspero, A. Canaguier, J. Jusseaume, J. Tassin, A. Lemainque, V. Thareau, A.-F. Adam-Blondon, R. Testolin, Molecular linkage maps: strategies, resources and achievements, in: A.-F. Adam-Blondon, J.M. Martinez-Zapater, Chittaranjan Kole (Eds.), Genetics, Genomics and Breeding of Grapes, Science Publishers and CRC Press, 2011, pp. 111–136.

[5] A. Canaguier, I. Le Clainche, A. Berard, A. Chauveau, M.S. Vernerey, C. Guichard, M.C. Le Paslier, G. Di Gaspero, O. Coriton, D. Brunel, A.-F. Adam-Blondon, B.I. Reisch, J. Londo (Eds.), Towards the deciphering of Chromosome structure in *Vitis vinifera*, X International Conference on Grapevine Breeding and Genetics, 1046 Acta Horticulturae, 2014, pp. 319–327.

[6] K.L. Howe, T. Chothia, R. Durbin, GAZE: a generic framework for the integration of gene-prediction data by dynamic programming, Genome Res. 12 (2002) 1418–1427, http://dx.doi.org/10.1101/gr.149502.

[7] J.E. Allen, S.L. Salzberg, JIGSAW: integration of multiple sources of evidence for gene prediction, Bioinformatics 21 (2005) 3596–3603, http://dx.doi.org/10.1093/bioinformatics/bti609.

[8] N. Vitulo, C. Forcato, E.C. Carpinelli, A. Telatin, D. Campagna, M. D'Angelo, R. Zimbello, M. Corso, A. Vannozzi, C. Bonghi, M. Lucchin, G. Valle, A deep survey of alternative splicing in grape reveals changes in the splicing machinery related to tissue, stress condition and genotype, BMC Plant Biol. 14 (2014) 99, http://dx.doi.org/10.1186/1471-2229-14-99.

[9] A.K.Y. Souvorov, B. Kiryutin, V. Chetvernin, T. Tatusova, D. Lipman, Gnomon-NCBI Eukaryotic Gene Prediction Tool, National Center for Biotechnology Information (US), 2010.

[10] S. Foissac, J. Gouzy, S. Rombauts, C. Mathe, J. Amselem, L. Sterck, Y.V. de Peer, P. Rouze, T. Schiex, Genome annotation in plants and fungi: EuGene as a model platform, Curr. Bioinforma. 3 (2008) 87–97.

[11] D.M. Martin, S. Aubourg, M.B. Schouwey, L. Daviet, M. Schalk, O. Toub, S.T. Lund, J. Bohlmann, Functional annotation, genome organization and phylogeny of the grapevine (*Vitis vinifera*) terpene synthase gene family based on genome assembly, FLcDNA cloning, and enzyme assays, BMC Plant Biol. 10 (2010) 226, http://dx.doi.org/10.1186/1471-2229-10-226.

[12] C. Parage, R. Tavares, S. Réty, R. Baltenweck-Guyot, A. Poutaraud, L. Renault, D. Heintz, R. Lugan, G.A. Marais, S. Aubourg, P. Hugueney, Structural, functional, and evolutionary analysis of the unusually large stilbene synthase gene family in grapevine, Plant Physiol. 160 (3) (2012) 1407–1419, http://dx.doi.org/10.1104/pp.112.202705.

[13] J. Grimplet, J.M. Martinez-Zapater, M.J. Carmona, Structural and functional annotation of the MADS-box transcription factor family in grapevine, BMC Genomics 17 (2016) 80, http://dx.doi.org/10.1186/s12864-016-2398-7.

[14] J. Grimplet, P. Agudelo Romero, R. Teixeira, J.M. Martinez Zapater, A.M. Fortes, Structural and functional analysis of the GRAS gene family in grapevine indicates a role of GRAS proteins in the control of development and stress responses, Front. Plant Sci. 7 (2016) 353, http://dx.doi.org/10.3389/fpls.2016.00353.

[15] D.C.J. Wong, R. Schlechter, A. Vannozzi, J. Höll, I. Hmmam, J. Bogs, G.B. Tornielli, S.D. Castellarin, J.T. Matus, A systems-oriented analysis of the grapevine R2R3-MYB transcription factor family uncovers new insights into the regulation of stilbene accumulation, DNA Res. 23 (2016) 451–466, http://dx.doi.org/10.1093/dnares/dsw028.

[16] A. Canaguier, M.C. LePaslier, E. Duchêne, S. Scalabrin, G. Di Gaspero, N. Mohellibi, C. Guichard, N. Choisne, A. Bérard, A. Chauveau, I. Le Clainche, R. Bounon, C. Guichard, C. Ruztenholtz, D. Brunel, M. Morgante, H. Quesneville, A.-F. Adam-Blondon, Development of a new version of the grapevine reference genome assembly (12X.v2) based on genetic maps and paired-end sequences, http://doi.org/10.15454/1.4962347083032307E12, (2017).

[17] K.W. Broman, H. Wu, S. Sen, G.A. Churchill, R/qtl: QTL mapping in experimental crosses, Bioinformatics 19 (2003) 889–890 12724300.

[18] F. Van Nieuwerburgh, R.C. Thompson, J. Ledesma, D. Deforce, T. Gaasterland, P. Ordoukhanian, Head SR (2012) Illumina mate-paired DNA sequencing-library preparation using Cre-Lox recombination, Nucleic Acids Res. 40 (3) (2012) e24, , http://dx.doi.org/10.1093/nar/gkr1000.

[19] M. Martin, Cutadapt removes adapter sequences from high-throughput sequencing reads, EMBnet J. 17 (1) (2011) 10–12.

[20] C. Del Fabbro, S. Scalabrin, M. Morgante, F.M. Giorgi, An extensive evaluation of read trimming effects on Illumina NGS data analysis, PLoS One 8 (12) (2013) e85024, , http://dx.doi.org/10.1371/journal.pone.0085024.

[21] B. Langmead, S. Salzberg, Fast gapped-read alignment with Bowtie 2, Nat. Methods 9 (2012) 357–359, http://dx.doi.org/10.1038/nmeth.1923.

[22] W.J. Kent, BLAT - the BLAST-like alignment tool, Genome Res. 12 (4) (2002) 656–664, http://dx.doi.org/10.1101/gr.229202.

[23] G.D. Schuler, Sequence mapping by electronic PCR, Genome Res. 7 (5) (1997) 541–550 (PMID: 9149949).

[24] J. Grimplet, J. Van Hemert, P. Carbonell-Bejerano, J. Diaz-Riquelme, J. Dickerson, A. Fennell, M. Pezzotti, J.M. Martinez-Zapater, Comparative analysis of grapevine whole-genome gene predictions, functional annotation, categorization and integration of the predicted gene sequences, BMC Res. Notes 5 (2012) 213, http://dx.doi.org/10.1186/1756-0500-5-213.

[25] F.S.A. Thibaud-Nissen, T. Murphy, M. DiCuccio, P. Kitts, Eukaryotic genome annotation pipeline, The NCBI Handbook [Internet], 2nd edition, National Center for Biotechnology Information (US), Bethesda (MD), 2013.

[26] H. Thorvaldsdottir, J.T. Robinson, J.P. Mesirov, Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration, Brief. Bioinform. 14 (2013) 178–192, http://dx.doi.org/10.1093/bib/bbs017.

[27] J. Grimplet, A.-F. Adam-Blondon, P.-F. Bert, O. Bitz, D. Cantu, C. Davies, S. Delrot, M. Pezzotti, S. Rombauts, G. Cramer, The grapevine gene nomenclature system, BMC Genomics 15 (2014) 1077, http://dx.doi.org/10.1186/1471-2164-15-1077.

[28] A. Canaguier, J. Grimplet, S. Scalabrin, G. Di Gaspero, N. Mohellibi, N. Choisne, S. Rombault, C. Ruztenholtz, M. Morgante, H. Quesneville, A.-F. Adam-Blondon, A new version of the grapevine reference genome assembly (12X.v2) and of its annotation (VCost.v3), http://doi.org/10.15454/1.5009072354498936E12, (2017).