Genome sequence initiatives

Anne-Francoise A.-F. Adam-Blondon, Olivier Jaillon, Silvia Vezzulli, Andrey Zharkikh, Michela Troggio, Riccardo Velasco

9

Genome Sequence Initiatives

Anne-Françoise Adam-Blondon,[1,] Olivier Jaillon,[2] Silvia Vezzulli,[3] Andrey Zharkikh,[4] Michela Troggio[3]* and *Riccardo Velasco[3]*

ABSTRACT

The possibility to access a complete genome sequence allows the development of new resources for high-throughput functional analysis without *a priori*, whole genome syntenic approaches for the detection of badly predicted genes or of conserved regulatory boxes. It also facilitates the identification of key genes by combined genetic and genomics approaches. Two parallel projects using different strategies for the sequencing of the grapevine genome were initiated. The first one aimed at the development of a reference genome sequence using a near homozygous genotype and the second one at the sequencing of a highly heterozygous high quality cultivar. Beside giving access to two annotated draft sequences of the genome of *Vits vinifera* L., the two projects developed genomic resources such as BAC libraries, BAC end sequences, sequenced full length cDNA libraries, physical and genetic maps. They offer very good material to assess the influence of heterozygosity on the development of physical maps or whole genome shotgun sequences and a first rough view on the polymorphism between the two haplotypes of a grapevine variety. Finally, they provided a first picture of the grapevine highly heterozygous genome, with a structure not much rearranged from the genome of its ancient hexaploid ancestor and the expansion of specific gene families.

Keywords: genome sequence, grapevine, *Vitis vinifera*, heterozygosity, annotation, hexaploid

[1]UMR INRA UEVE ERL CNRS Génomique Végétale, 2 rue Gaston Crémieux, BP 5708, 91 057 Evry cedex, France.
[2]CEA, DSV, Institut de Génomique, Genoscope, 2 rue Gaston Crémieux, CP5706, 91057 Evry, France.
[3]IASMA Research and Innovation Centre, Fondazione Edmund Mach, Via E. Mach 1, 38010, S. Michele all'Adige, Trento, Italy.
[4]Myriad Genetics Inc., 320 Wakara Way, Salt Lake City, 84108 UT, USA.
*Corresponding author: *adam@evry.inra.fr*

9.1 Introduction

The possibility to access a complete genome sequence allows the development of new resources for (i) high-throughput functional analysis without *a priori* (see for instance *http://www.catma.org/*), (ii) whole genome syntenic approaches for the detection of badly predicted genes or of conserved regulatory boxes (see for instance Roest Crollius et al. 2000). Finally, it greatly facilitates the identification of key genes by combined genetic and genomics approaches (Morgante and Salamini 2003).

Three main strategies can be developed for the sequencing of complex genomes: a clone-by-clone approach, a whole genome shotgun approach and a mix of both (for review, see Green 2001). The BAC by BAC approach consists of building an integrated genetic and physical map of the genome, then to select the minimum set of large insert clones that allows to cover the genome (minimum tiling path) and to obtain the finished sequence of these clones. The quality of such a genome sequence is the highest possible among these three approaches, and the number of remaining gaps in the sequence depends on the quality of the integrated genetic and physical map used. This usually necessitates many exchanges between the sequencing teams and the mapping teams during the sequencing process to improve the tiling path. In plants, the *Arabidopsis* genome sequence (Arabidopsis Genome Initiative 2000) and the rice genome sequence (International Rice Genome Sequencing Project, IRGSP 2005) were obtained using this approach. The second method, called Whole Genome Shotgun (WGS) is now the most widely used because it is less labor intensive, faster, and thus much cheaper than the BAC by BAC approach. In this case, different libraries are prepared with contrasting average insert size (plasmid, cosmid and BAC). The extremities of these clones are sequenced and the pairs of reads are assembled in sequence contigs and super-contigs. The super-contigs can then be aligned along an integrated genetic and physical map. The resulting genomic sequence contains sizeable gaps between the contigs of a super-contig, gaps of unknown size between the super-contigs and the total gap number is heavily dependent on the genome coverage. In Angiosperms, the poplar, rice and papaya genomes were sequenced using this approach (Goff et al. 2002; Yu et al. 2002; Tuskan et al. 2006; Ming et al. 2008). The recent development of ultra-high throughput new technologies for sequencing (Shendure et al. 2004; see also Chapter 10) has definitely lowered the price of WGS approaches and are now increasingly used for such projects (see for instance Huang et al. 2009). A comparison was made between the clone-by-clone rice sequence and a 6 × WGS sequence of the same subspecies (IRGSP 2005). The WGS draft sequence covered 78% of the rice genome and 75.3% of the full-length cDNA supported genes models, probably due to local bad quality of the WGS sequence or of the

cDNA sequence, and also to contamination. Another issue is the level of heterozygosity of the genome sequenced. In WGS approaches, it may cause a bad assembly of the sequence, leading to artificial duplications of variable size (Vinson et al. 2005).

Two parallel projects using different strategies for the sequencing of the grapevine genome were initiated. The first one was established within the frame of the International Grapevine Genome Program (IGGP; *www.vitaceae.org*) and the second one resulted from collaboration between the Istituto Agrario di San Michele all'Adige and Myriad Genetics.

9.2 Strategies for Sequencing of the Grapevine Genome

9.2.1 Development of BAC Library Resources

A growing resource of grapevine BAC libraries is available (Table 9-1), mainly for *Vitis vinifera* L. well known cultivars but also for a *V. vinifera* genotype introgressed with disease resistance genes. The two *V. vinifera* cv. Cabernet Sauvignon BAC libraries were chosen as reference libraries by the IGGP, as the initial plan was to use Cabernet Sauvignon as the model genotype for grapevine, due to its wide distribution around the world. Now, the development of new BAC resources are more targeted to other species of the *Vitis* or *Muscadinia* genus, aiming at identifying efficient alleles for resistance to diseases or for the adaptation to abiotic stresses through comparative genomics approaches.

9.2.2 Development of Physical Maps of the Grapevine Genome

Physical maps consist of the assembly into contigs of overlapping large insert clones based on fingerprint similarities and on the presence of common markers. These assemblies can be produced using the software FPC (Nelson and Soderlund 2005) and the distance is in base pairs. Most of the physical maps developed in plants are now based on the assembly of BAC clones (see Meyers et al. 2004 for review).

Two physical maps of the *V. vinifera* genome have been recently constructed: one using a Pinot Noir ENTAV 115 BAC library (Scalabrin et al. 2010) and one using the Cabernet Sauvignon reference BAC library (Moroldo et al. 2008). The fingerprinting method that was used for the construction of both maps is based on the fluorescent labeling of four different restriction sites and was adapted by Moroldo (2008) from the original paper of Luo et al. (2003). 49,104 BAC clones from the Pinot Noir library representing 11 times the grape genome were fingerprinted (Scalabrin et al. 2010) whereas 30,828 clones, representing eight times the genome were processed from the Cabernet Sauvignon BAC library (Moroldo et al. 2008; Table 9-2).

Table 9-1 Grape BAC libraries published in grapevine.

Genotype	Species	Number of clones	Genome coverage	Publicly available BES	Reference	Distribution[$]
Syrah	*V. vinifera*	55296	16.5	No	Tomkins et al. 2001	CUGI
Syrah cl.ENTAV 173	*V. vinifera*	23040	7.7		Adam-Blondon et al. 2005	
Cabernet-Sauvignon cl.412	*V. vinifera*	44544	13.2	Yes	Adam-Blondon et al. 2005	CNRGV
Pinot Noir cl.777	*V. vinifera*	48384	14.8	No	Adam-Blondon et al. 2005	
Pinot Noir cl.115	*V. vinifera*	23040	4.6	Yes	Adam-Blondon et al. 2005	
Pinot Noir cl. 115	*V. vinifera*	26064	6.8	Yes	Velasco et al. 2007	
VHR3294-R23	*V. vinifera**	55295	9.4	No	Barker et al. 2005	
PN40024	*V. vinifera*	70656	~14	Yes	Jaillon et al. 2007	CNRGV

**V. vinifera* introgressed with a resistance gene carrying genomic region from *Muscadinia rotundifolia*
$CUGI: *www.genome.clemson.edu*; CNRGV: *http://cnrgv.toulouse.inra.fr/*

Table 9-2 Features of the Cabernet Sauvignon (Moroldo et al. 2008) and of the Pinot Noir (Scalabrin et al. 2010) physical maps. The average band size (1.1 kbp) was calculated by dividing the average insert size of the BAC clones by the average number of bands *per* clone. This value was used to estimate the physical length of the contigs. The physical map of the Cabernet Sauvignon and its links with the grapevine genetic map and the reference genome sequence can be queried at *http://urgi.versailles.fr/cmap*. The physical map of the Pinot Noir and its links with the Pinot noir genome assembly and the grapevine genetic maps developed at IASMA can be viewed at *http://genomics.research.iasma.it*.

Whole physical map	Cabernet S	Pinot Noir
Number of clones fingerprinted	44,544	44,544
Number of clones used for map assembly	30,828	38,983
Number of singletons	1,111	3,372
Number of contigs	1,770	1,804
Physical length of the contigs (kbp)	715,684	888,000
Contigs anchored on the genetic map:		
No. of contigs	395	436
Coverage (kbp)	255,476	341,621

Collections of BAC end sequences (BES) were also developed for both genotypes allowing the development of new genetic markers to improve the integration between the genetic and the physical maps (Lamoureux et al. 2006; Troggio et al. 2007): in particular, 77,231 Cabernet Sauvignon BES contained 3,601 SSR loci (V. Tharreau, unpublished results), which are very efficient molecular markers for mapping in grapevine.

A common feature of both maps is that the total length of the contigs is larger than the estimated size of the grape genome (1.5–1.6 fold) and that this expansion could mainly be attributed to the effects of heterozygosity (Moroldo et al. 2008; Scalabrin et al. 2010). Indeed, genotyping with simple sequence repeat (SSR) markers of a *V. vinifera* germplasm collection showed a high level of heterozygosity in *V. vinifera* (Aradhya et al. 2003). Moreover, it was later shown that an important part of the sequence variation is due to insertion/deletion events and that the frequency of Single Nucleotide Polymorphisms (SNPs) is uneven, by comparing the two haplotype sequences in "Cabernet Sauvignon" over two different genomic regions encompassing 182 kbp (F. Chopin et al. unpublished) or as a result of the WGS sequencing of the cultivar Pinot Noir (Velasco et al. 2007). These features affect the banding pattern of the fingerprints produced from two allelic regions and their assignment to overlapping regions by the FPC software. Even when the BAC clones corresponding to the two different haplotypes are assembled within the same contig, their local order has to be considered with caution as many false duplications are observed (Moroldo et al. 2008; Scalabrin et al. 2010).

Up to now, only three fingerprinting-based physical maps were assembled for heterozygous plants other than grape: *Prunus* (Jung et al. 2004), apple (Han et al. 2007) and poplar (Kelleher et al. 2007). The peach map was constructed using the same fingerprinting method as the one

used for grapevine, but some other differences impair a homogeneous comparison of the two assemblies (Moroldo et al. 2008). The poplar and apple maps both showed an 1.2-fold genome size expansion (Han et al. 2007; Kelleher et al. 2007). This value is lower compared to the ones found in grapevine and it may be related to the use of agarose gels in fingerprinting (Nelson et al. 2007; discussed in Moroldo et al. 2008), as at least in poplar, the sequence variation was shown to have the same features as in grapevine (Kelleher et al. 2007). Another concern was an unusually high percentage of chimeric contigs that was found compared to physical maps of other species despite a stringent filtering of possible plate cross contaminations before the physical map assembly (Moroldo et al. 2008). Such chimerism may arise from partial sequence similarity within the BAC clones caused by shared fingerprint bands corresponding to repetitive sequences or large-scale duplications.

Despite these limitations, in absence of a whole genome sequence, such integrated maps are invaluable resources for the quick development of new markers in targeted regions using BAC-end sequences (see Barker et al. 2005; Castellarin et al. 2006; Troggio et al. 2007; Moroldo et al. 2008 for examples in grapevine), for candidate gene approaches by establishing links between genetic maps where quantitative trait loci (QTLs) for traits of interest have been located and gene-containing BACs and to prepare and accelerate map base cloning projects. For instance, it was proved to be very useful to understand the genome organization of a gene family like the flavonoid hydroxylase ones (F3′H and F3′5′H) at a fine scale. Three groups recently cloned and characterized these genes in grape, based on available expressed sequence tags (ESTs) and sequence homologies with known F3′H and F3′5′H genes in other plants and analyzed their expression in relation with the composition in particular classes of flavonoids like cyanidin and delphinidin-based anthocyanins (Jeong et al. 2005; Bogs et al. 2006; Castellarin et al. 2006). Two groups showed that both F3′H and F3′5′H belong to a multigene family, either by sequencing several fragments amplified with degenerate primers (Jeong et al. 2005), or by a physical mapping of sequences obtained with the same approach combined with a single-strand conformational polymorphism (SSCP) based evidence of the different members of the gene family (Castellarin et al. 2006). The second approach revealed bigger and tandemly duplicated gene families (Castellarin et al. 2006). The fact that different forms of the F3′5′H genes showed different patterns of transcription (Jeong et al. 2005; Castellarin et al. 2006), underlines the importance of obtaining the complete family of sequences, which is now easy to implement with the availability of the grapevine whole genome sequence. Finally, the establishment of finely

integrated local genetic and physical maps will remain the key step of map-based cloning projects for genes of interest. The long generation time and the space needed for large progenies make such approaches much more difficult in perennial species than in annual species, even though it has already been proved possible (see for instance Patocchi et al. 1999 in apple tree and Claverie et al. 2004 in plum) and especially for target genes expressed in plantlets. The first local fine map in grapevine was published by Barker et al. (2005) for a region containing a major gene for resistance to powdery mildew (see Chapter 8).

9.2.3 Strategy for the Development of a Reference Genome Sequence for Grapevine by the French-Italian Public Consortium for the Grapevine Genome Characterization

In order to have a WGS sequence of the highest quality possible, the international consortium decided to sequence a nearly homozygous genotype, the PN40024 with a 12 × coverage. The near homozygous line was derived from Pinot Noir at the INRA station of Colmar (Bronner and Oliveira 1990) by nine successive selfing steps. Its analysis with SSR markers led Jaillon et al. (2007) to suspect that an outcross might have occurred in any of the three first generations of selfing. A rough paternity search showed that the most probable parent would be Helfensteiner (Jaillon et al. 2007). Even if the number of selfing generations the PN40024 was derived from may be lower than nine, its level of homozygosity was showed to be quite good: 7/102 tested SSR loci were heterozygous and the heterozygous genome sequence contigs did not represent more than 7% of the genome size (Jaillon et al. 2007).

In 2005, an agreement was signed between France and Italy to share the costs of this public initiative (*http://www.agriculture.gouv.fr/spip/IMG/pdf/cp_vigne170505-1.pdf*). The necessary funds have been provided in France by the French Ministry of Education and Research, the Genoscope (*www.genoscope.cns.fr*), INRA and the French National Agency for Research (*www.agence-nationale-recherche.fr*) and in Italy by the Italian Ministry for Agriculture and Forestry, the Friuli Venezia Giulia County and a consortium of private companies and banks. The sequence production was shared between Genoscope, IGA (Istituto di Genomica Applicata; *www.appliedgenomics.org*) and CRIBI (Centro Ricerca Interdipartimentale Biotechnologie Innovative; *www.cribi.unipd.it*) and the rest of the work was shared by other French (Institut National de Recherche Agronomique; INRA; *www.inra.fr*) and Italian laboratories (VIGNA consortium; *www.vitisgenome.it*).

9.2.4 Strategy for Sequencing a Highly Heterozygous Genome: The Pinot Noir Genome Project

The Grapevine Genome Initiative, collaboration between the Fondazione Edmund Mach-Istituto Agrario di San Michele all'Adige–(FEM-IASMA) and two private companies, Myriad Genetics, Inc. and the 454 Life Science, has focused on sequencing the elite cultivar Pinot Noir to provide insight into the structural nature of heterozygosity in an outcrossing species known to be highly heterozygous. The Sanger sequences (7x) and the 454 sequences (4.2x) were assembled by a Myriad proprietary software able to handle heterozygous sequences, revealing a large amount of polymorphism. Of special interest to biologists and breeders are polymorphisms in and around the coding regions representing a substantial resource for molecular breeding programs, as well as trait and quantitative trait loci (QTL) marker association. This project adopted a new approach for the sequencing and assembly of a large heterozygous eukaryotic genome, which consisted of a whole genome shotgun combining paired reads produced by Sanger sequencing (Sanger et al. 1977) and unpaired reads obtained through sequencing by synthesis (SBS, Margulies et al. 2005). The coverage of SBS reads aimed at identifying polymorphic sites and at closing most gaps between DNA contigs. A novel approach to genomic alignment was developed to generate a single consensus sequence from both chromosomes of Pinot Noir.

The project was funded entirely by the Province of Trento (Italy) while the sequence production was shared by Myriad Genetics, Inc., IASMA Research Centre and 454 Life Science.

Both genome sequences were recently published by Jaillon et al. (2007) for the 8 × version of the reference genome sequence and by Velasco et al. (2007) for the Pinot Noir genome sequence (10.7×), revealing interesting features of the grapevine genome sequence.

9.3 The Reference Grapevine Genome Sequence: The Grapevine Genome Descends from an Ancient Event of Polyploidization

All the results so far generated (maps, markers, shotgun reads, genome sequence and its annotations), once analyzed and published, were submitted to the relevant database at NCBI and can also be viewed in the two genome browsers of the project: *http://www.genoscope.cns.fr/externe/GenomeBrowser/Vitis/* and *http://www.appliedgenomics.org/*.

9.3.1 Assembling the Reads of the WGS

The first version of the reference sequence of the grapevine genome has been obtained from a set of sequences covering 8.4 times the genome size, which

would *a priori* lead to a relatively high quality sequence compared with other similar projects (Table 9-3, Goff et al. 2002; Tuskan et al. 2006; Velasco et al. 2007; Ming et al. 2008). Several assemblies were produced along the project and the total size of the genome sequence (487 Mb) was obtained at a six-fold coverage. After this stage, the improvements that were observed concerned an increase of the length of the contigs and super-contigs and the diminution of their respective numbers (Table 9-3). The objective of a 12 × coverage of the grapevine genome is now reached, a new assembly has been produced with an improvement of the assembly method and is now available to the public with its annotation at NCBI.

The anchorage of the sequence scaffolds along the grapevine genome was realized in two steps (Jaillon et al. 2007): (i) some super-contigs were joined together into 45 ultra-contigs using paired BAC-end sequences (BES) from Cabernet-Sauvignon and BAC contigs of the same BACs from our Cabernet-Sauvignon physical map and (ii) the ultra-contigs and remaining scaffolds were then aligned along the *V. vinifera* genetic map published by Doligez et al. (2006). This resulted in 61% of the assembly being anchored and orientated and 69% anchored (Jaillon et al. 2007). This aspect is being improved in the 12 × version of the genome sequence by a combination of improvements of the genetic map and of the assembly: currently, around 85% of the assembly is anchored and orientated and 91% anchored (The French-Italian Public Consortium for the Grapevine Genome Characterization, unpublished results).

9.3.2 Annotation of the Grapevine Genome Sequence

Around 41% of the grapevine genome sequence was found to correspond to repetitive sequences, which are higher in proportion to in rice and *Arabidopsis*, but surprisingly lower than in papaya, which has also a genome smaller in size than grapevine (Ming et al. 2008). A striking feature in grapevine was that introns are rich in repeats, with 12.4% of intron sequence that contains TE, mainly LINEs (Jaillon et al. 2007). Unlike in Arabisopsis, but as in poplar, in grapevine the TE density is non-homogeneous along the sequence and is opposite to gene density (Arabidopsis Genome Initiative 2000; Tuskan et al. 2006; Jaillon et al. 2007). Recently, an in-depth re-annotation of the Class II elements was carried out based on the two published genome sequences and the authors reported 1,160 complete elements and 2,086 degenerated elements distributed in four families: *hAT*, *PIF*, *Mutator* and *CACTA* (Benjak et al. 2008). The authors also found transcription evidence of Class II elements in the EST databases and showed that they contribute to inter-varietal polymorphism (Benjak et al. 2008). They also suggested that the percentage of genome corresponding to TE by Jaillon et al. (2007) and Velasco et al. (2007) may be underestimated (Benjak et al. 2008).

Table 9-3 Comparison of WGS assemblies in different plant genome sequencing projects.

	Grapevine[1]	Grapevine[1]	Grapevine[1]	Grapevine[1]	Rice[2]	Poplar[3]	Papaya[4]
Number reads produced	-	-	-	8,200,000	5,500,000	7,600,000	2,800,000
Coverage	2.9	3.7	5	8.4	6	7.5	3
Number of contigs	170,000	150,000	125,000	19,557	42,109	45,970	47,483
N50 contigs[6]	2.9 kb	4.3 kb	7.7 kb	65.9 kb	20 kb	2-29-126 kb[5]	11 kb
Number of super-contigs	140,000	100,000	77,000	3,514	-	-	17,764
Final size	-	-	477 Mb	487 Mb	390 Mb	410 Mb	370 Mb

[1]Near inbred line; Jaillon et al. 2007.
[2]Inbred line; Goff et al. 2002.
[3]Heterozygous individual; Tuskan et al. 2006.
[4]Near inbred line; Ming et al. 2008.
[5]The 26,363 contigs of 1–20 kb have a N50 = 2 kb; the 8,245 contigs with a size > 20 kb by unanchored on the genetic map have a N50 = 29 kb and the 11,362 contigs anchored on the genetic map have a N50 = 126 kb (Tuskan et al. 2006).
[6]N50 contigs is the median of the contig size.

The non-coding RNAs were also analyzed, showing that the tRNA, snRNA and ribosomal RNA sequences have similar numbers and distributions as in *Arabidopsis*, rice and poplar (Jaillon et al. 2007). As for the miRNA, 22 families were detected by MicroHarvester (Dezulian et al. 2006), with the noticeable specific expansion of two of them in grapevine (miR845 and miR395) and the first record of the miR535 and miR1213 families in a dicotyledon. As the miR395 family is thought to be involved into sulfate metabolism, these observations may lead to interesting research in relation with agronomical traits. Similar findings were derived from the Pinot Noir genome project, yet with grapevine specific miRNA families and targets (Velasco et al. 2007).

Automatic genome structural and functional annotations rely on the integration of different sets of data: *de novo* gene prediction, homologies with ESTs, cDNA, proteins, or sequences conserved across evolution (see for instance Town 2006 for the Medicago genome project and Jaillon et al. 2007 for the grapevine genome project). With regard to the latter point, in animals, the comparison of the human and of *Tetraodon* genome sequences allowed to improve or define gene models (Roest-Crolius et al. 2000; Jaillon et al. 2004). But the key resources with regard to the quality of the output of an automatic annotation are EST and full length cDNA collections. Full length cDNA sequences allow globally improving the quality of the protein coding gene models proposed by the process of automatic annotation. For instance, in *Arabidopsis*, although the genome had been sequenced four years ago and several rounds of annotation had already been done, a collection of 31,558 flcDNAs allowed to discover 326 unpredicted genes and to correct the structure of 45% models with no previous full length cDNA support (Castelli et al. 2004). Finally, it is an important resource for functional genomics (see as an example *http://www.evry.inra.fr/public/projects/orfeome/orfeome.html*). Large sets of public *Vitis* ESTs were available to Jaillon et al. (2007; around 190,000) and the consortium produced in addition around 48,000 sequences from libraries enriched in full length cDNAs. These sequences, mapped on available versions of the genome assembly, also allowed to quickly develop a set of 650 genes for which both the coding and genomic sequences were fully known and manually inspected, necessary for the training of the software used for *de novo* gene prediction (S. Aubourg and C. Clepet unpublished results).

The current number of proposed protein coding gene models in grapevine is 30,434 (Jaillon et al. 2007). The next version of the annotation should be improved by several means: an increase of the genome coverage will reduce the number of gaps and errors that still can affect gene predictions (IRGSP 2005); over 100,000 full length cDNA sequences have now been produced and finally the deep resequencing of grapevine transcriptome has also been included in the process (Denoeud et al. 2008; The French-

Italian Public Consortium for the Grapevine Genome Characterization, unpublished results).

9.3.4 Whole Genome Comparison between the Grapevine Genome and Itself or Other Angiosperm Genomes

One of the striking features of the grapevine genome is that it was found to be an ancient hexaploid with each genomic region always aligning (at the protein level) with two others (Jaillon et al. 2007; Velasco et al. 2007; Fig. 9-1), which confirmed and improved previous conclusions derived from cytological observations of F_1 hybrids between *V. vinifera* ($2n = 38$) and *Muscadinia rotundifolia* ($2n = 40$) and suggesting a polyploid origin of the *Vitis* genome (Chapter 1 Section 5.4; Patel and Olmo 1955). Moreover, the blocks of synteny between the three ancestral genomes have undergone few rearrangements (Jaillon et al. 2007).

A comparative analysis of the grapevine genome with the four angiosperm genomes available to date, showed that the hexaploid ancestor of grapevine is also the ancestor of *Arabidopsis*, papaya and poplar, but would not be at the origin of the monocot lineage (Jaillon et al. 2007; Ming et al. 2008). This result was only based on the observation of blocks of synteny between species like those represented in Fig. 9-1. A different timing of the whole genome duplication or hybridization events in the dicotyledon lineage, based on studies of substitution rates between pairs of othologous genes was proposed by Velasco et al. (2007), stressing the difficulty to date gene divergence between species where substitutions rates may be different (e.g., comparing annual and perennial species). Altogether these results confirmed the importance of whole genome duplications/polyploidy in genome evolution (Ohno 1970; Jaillon et al. 2004; Aury et al. 2006) and especially in plants (Bowers et al. 2003, Arabidopsis Genome Initiative 2000; IRGSP 2005; Tuskan et al. 2006; Ming et al. 2008; Rensing et al. 2008).

9.4 The DNA of Pinot Noir Discloses Features of a Highly Heterozygous Genome

9.4.1 Sequencing and Assembling a Heterozygous Genome: Problems and Solutions

As already discussed, the assembly of a whole genome shotgun sequence is more problematic for a highly heterozygous genotype than for an inbred genotype. For the Pinot Noir project, two sequencing techniques were adopted. The Sanger method was used to generate 6.5 × coverage of the genome. This was integrated with a 4.2 × coverage of sequence reads generated by a scalable, highly parallel sequencing by synthesis (SBS) method.

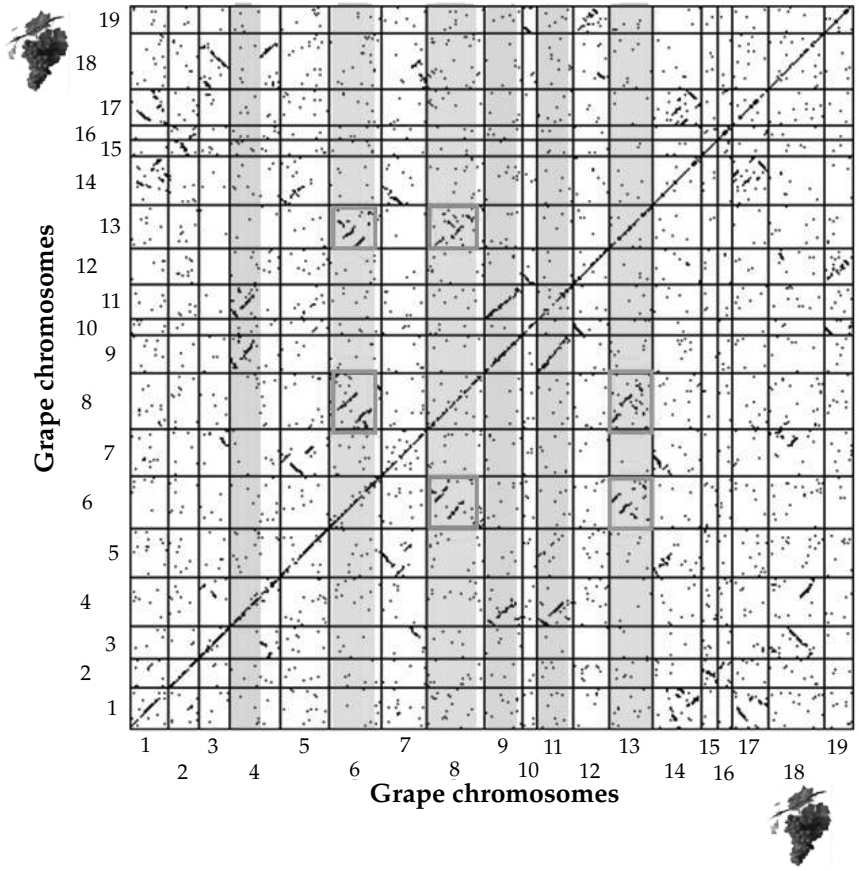

Figure 9-1 Topological distribution of paralogous genes in the grapevine genome showing a 3 to 3 relationship of blocks. The 19 chromosomes of grape are represented on both the x and y axis. Dots represent the positions of paralogous pairs of genes. Paralogous genes were computed according to a Reciprocal Best Hit (Jaillon et al. 2007). In blue is highlighted the exclusive correspondence between a half of chromosome 4 with chromosomes 9 and 11 and in red the same kind of relationships between the chromosomes 6, 8 and 13. Each and every one of the chromosomes have a relationship with two others.

Color image of this figure appears in the color plate section at the end of the book.

Existing software and strategies were not adequate for the assembly of this highly heterozygous genome. A modified version of the assembly pipeline developed for rice genome sequencing (Goff et al. 2002) was developed to handle the high level of heterozygosity in the Pinot Noir genome (Zharkikh et al. 2008). As a result, most of the assembled contigs were represented by a consensus sequence derived from the alignment of the two existing haplotypes. In addition to the regions in which such a merge was possible, many regions were found to be chromosome-

specific, i.e., either segments with different DNA sequences flanked by orthologous regions of the two homologous chromosomes (Fig. 9-2a) or gaps corresponding to sequences missing in one chromosome but not in the other (Fig. 9-2b).

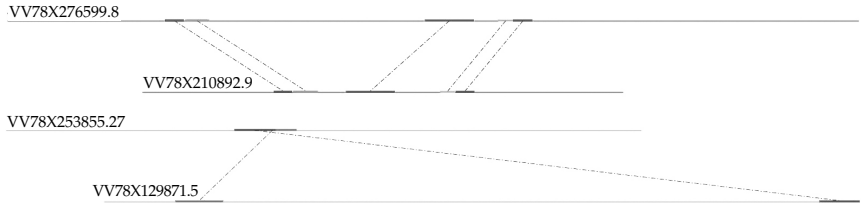

Figure 9-2 Two examples of chromosome-specific regions of the Pinot Noir genome. (a) segments with different DNA sequences (in green) flanked by orthologous regions of the two homologous chromosomes (same color for both chromosomes); (b) Gaps corresponding to a sequence missing in one chromosome but not in the other.

Color image of this figure appears in the color plate section at the end of the book.

The assembly started with aligning the unique sequences and then progressively included sequences with a higher degree of repetitiveness. Applying the procedure to about 6.6 M reads from Sanger sequencing, 90.6% of which represented paired clone ends, 211,374 initial seed contigs of unique sequences were generated. By using long clone links with non-repetitive clone ends, seed contigs were ordered into metacontigs (ordered assembly of contigs, referred to as supercontigs or scaffolds in other publications) and the overlapping members of the metacontigs were iteratively merged. After merging the sequences into 120,000 contigs, the data were combined with 4.2 genome-equivalents of SBS sequences. This helped identifying polymorphic sites in the assembled sequences and closing about 25% of the gaps between contigs. After removal of 10,847 contigs composed only of tandemly repeated sequences and disposal of 7,003 contigs shorter than 1,000 bp, the iterative assembly produced 58,611 contigs corresponding to 530.9 Mb of genomic DNA. 44,179 of the 58,611 contigs were assembled into 2,093 metacontigs covering 477.1 Mb of genomic DNA and the remaining 14,432 contigs were singletons.

9.4.2. Metaconting Alignment on a Dense Genetic Map

The next phase of the assembly involved positioning metacontigs along the chromosomes using a genetic map. A sample of validated single nucleotide polymorphisms (SNPs) in non-repetitive regions of the genome was used to develop 799 markers (Pindo et al. 2008; Vezzulli et al. 2008) for the genotyping of a Syrah x Pinot Noir full sib family (Troggio et al. 2007). This genetic map currently includes 1,767 molecular markers arranged in 19 linkage groups and covering 1,276 cM (Velasco et al. 2007).

By matching the sequences of 1,356 markers to the assembled contigs, the 397 largest metacontigs were anchored to the genetic map, thus assigning to the linkage groups about 435.1 Mb of sequence or more than 91% of the metacontig sequences. Over 81% of these sequences have been anchored by two or more markers and, therefore, are oriented. The largest metacontig, mapped to linkage group 18, was anchored with 21 markers and covered almost 32.3 cM of the map. Large metacontigs were mostly associated with non-centromeric regions of the chromosomes with a relatively high recombination rate: about 150–200 kb per cM. Pericentromeric regions, abundant in TEs and tandem repeats, were covered with multiple short metacontigs, most of which were poorly anchored to the genetic map (not oriented or ambiguous order). The recombination rate in these regions was proved to be low and the DNA content per cM to increase from 300 to 1,000 kb. Sequences which are still unmapped are mostly represented by short metacontigs and singleton contigs containing multiple tandem repetitive sequences (Zharkikh et al. 2008).

SNP-based markers were essential in improving the metacontig assembly. Many adjacent metacontigs that were not initially merged because of non-significant links between them were associated based on neighboring genetic markers and could therefore be successfully merged into a single larger metacontig. On the other hand, if a metacontig was associated with several markers from different linkage groups or with distant markers from the same linkage group, it was considered chimeric and was split into separate metacontigs by a semi-automated procedure (Velasco et al. 2007). The anchored sequence of the cultivated grapevine genome together with the large arsenal of SNP loci are available at the public databases and at *http://genomics.research.iasma.it* (Velasco et al. 2007).

9.4.3 The Pinot Noir Genome Structure

The set of Pinot Noir chromosome pairs included a considerable number of haplotype-specific gaps (sequences present in one haplotype but not in the other). The total length of the 1,042,174 identified gaps corresponded to 48.9 Mb (Velasco et al. 2007). In some chromosomal regions, the two alternative haplotypes were too different for the algorithm employed during assembly to combine them into a single contig and were considered by Velasco et al. (2007) as hemizygous DNA (22,061 contigs with a total length of 65.1 Mb). All these data allowed the authors to conclude that the homologous chromosomes of Pinot Noir differed on average by 11.2% of their DNA sequences and that the grapevine genome exists in a dynamic state, mediated at least in part by TE activity, as reported for other species (Morgante et al. 2005). Indeed, the large grapevine genomic gaps are frequently bordered by 5 bp direct repeats, reminiscent of a type of

DNA excision mediated by a precise process of transposition (Chandler and Mahillon 2002). A total of 2 million SNPs (1,751,176 of which are identified in the anchored contigs) were discovered along the grapevine genome together with more than a million of insertion/deletion events (Velasco et al. 2007) for an overall rate of 6.3 polymorphisms per kilobase. The level of heterozygosity observed in Pinot Noir was much higher compared to that reported for the heterogozygous Nisqually-1 genome in poplar (Tuskan et al. 2006) where 1,241,251 SNPs or indels were identified for an overall rate of 2.6 polymorphisms per kilobase. An average value of 4.0 SNP per kb was observed, with coding and non-coding regions demonstrating different degrees of polymorphism, 2.5 and 5.5 SNPs per kb respectively (Velasco et al. 2007), while in poplar 83% of the polymorphisms discovered occurred in non-coding portions of the genome. One or more SNPs were found in 86.7% of anchored genes and 71.4% of genes had more than four SNPs (Table 9-4). Those gene-based markers are valuable tools, as SNPs present in functional genes may cause natural phenotypic variation (Fridman et al. 2000; Thornsbery et al. 2001) and help in genetic diagnosis.

A reduction of SNP frequency in gene desert regions was observed as described for the dog genome (Lindblad-Toh et al. 2005). Independent of the gene density, SNP frequency was found not homogeneous along the grapevine genome, with regions where SNP frequency peaked between 5 and 7.5 per 1 kb, and others displaying dramatically reduced frequencies (Fig. 9-3). However, the sparseness of putative quasi-homozygous haplotypic blocks indicates that heterozygosity prevails (Velasco et al. 2007).

The existence of structural diversity between homologous chromosomes within plant species involving different contents in coding regions has been reported in maize (Brunner et al. 2005; Clark et al. 2007) as well as in allogamous plants (Rafalski 2002) such as grapevine (Velasco et al. 2007).

9.4.4 Comparison with the Nearly Homozygous Genotype PN40024 Assembly

The 58,611 contigs obtained from sequencing the heterozygous genome of grape (HTA = heterozygous assembly) were compared with the 57,634 contigs obtained from sequencing the near homozygous line of grapevine (PN40024) (HMA = homozygous assembly based on 8.4 × read coverage). About 57% of contigs from each set showed at least 1,000 bp non-repetitive sequence overlaps with contigs from the alternative set. More than 50% of the overlaps could be perfectly aligned with the number of sequence discrepancies corresponding to the estimated level of heterozygosity. The remaining overlaps contained multiple large heterozygous insertions and small (0.5–1.5 kb) sequence inversions which prevent reliable alignment of contigs. 15,851 HTA contigs and 14,148 HMA contigs were completely

Table 9-4 SNPs in exons and non-coding DNA (A) and percentage of anchored genes tagged with SNPs (B).

A.

Region	SNPs No.	SNPs No./kb
In non-coding	998,149	3.1
Intron	642,219	5.5
Exon	110,808	2.5

B.

SNPs/gene	No. anchored Genes	(%)
0	3,773	13.3
1-3	4,362	15.3
≥ 4	20,292	71.4
Total	28,427	100.0

overlapped by the corresponding alternative contigs, whereas 26,193 HTA contigs and 21,613 HMA contigs closed gaps in the alternative assembly. Adding all these numbers, the combined HTA+HMA assembly closed about 63.3% of HTA gaps and 70.0% of HMA gaps (Zharkikh et al. 2008).

9.5 Perspectives

The automatic annotation has then to be progressively refined by an expert-based manual annotation, which is made by a community of scientists and ideally should be centralized and organized (See as an example Aubourg et al. 2005; Menda et al. 2008). A first glance on the protein-coding sequences in grapevine showed an expansion of some gene families involved into the biosynthesis in terpen and flavonoid compounds, both important for the berry taste, before and after vinification. For instance, 89 genes and 27 pseudogenes were found in grapevine, arranged in clusters, sometimes of important size (33 genes in a 672 kb tandem array on chromosome 18; Jaillon et al. 2007) to be compared to around 40 genes in *Arabidopsis*, poplar and rice (Aubourg et al. 2002; Jaillon et al. 2007). Similarly, the stilbene synthase gene family is arranged in 9 clusters in grapevine, totaling 43 genes and 9 pseudogenes while in *Arabidopsis*, only 4 genes can be found and in rice and poplar, no more than 30 genes (Jaillon et al. 2007). It is now possible to set up a global strategy to decipher the individual biochemical function of each family member and its physiological role. In a more general point of view it is now possible to set up tools for the genome wide analysis of gene transcription and translation in grapevine.

Having a reference genome sequence allows the high throughput resequencing of varieties or wild species to extend the first insight on the molecular basis of polymorphism in grapevine given by the Pinot Noir genome sequence (Velasco et al. 2007). This was recently started by Myles et al. (2010) using a Solexa/Illumina platform and is necessary to fully

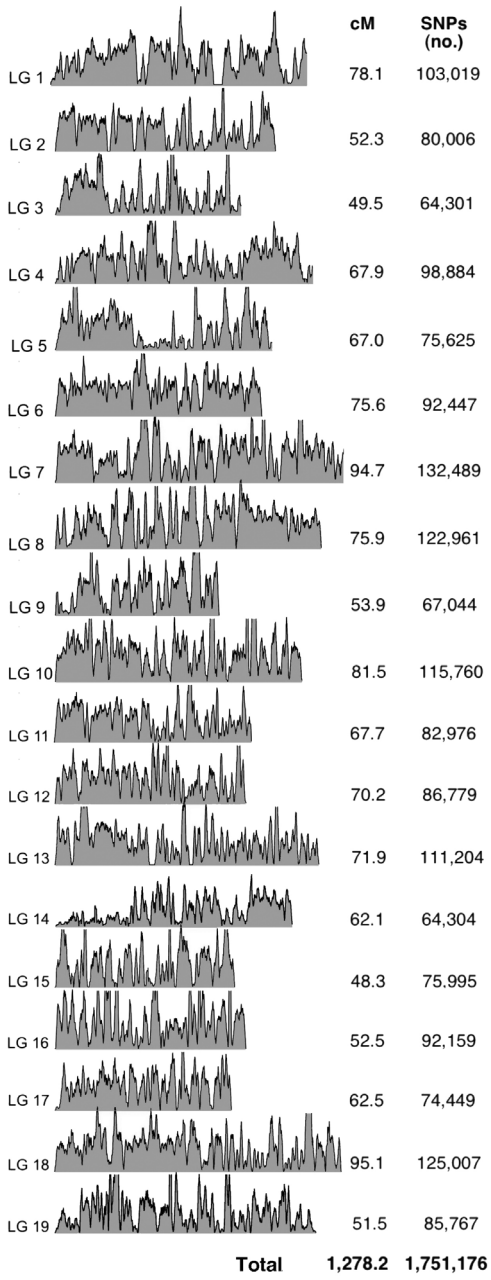

Figure 9-3 SNP density along the 19 chromosomes (LG) of *V. vinifera*. Left and right of the figure correspond respectively to top and bottom of chromosomes according to Troggio et al. (2007). The SNP values reported do not consider gaps in and among metacontigs.

understand the role of different structural variations (transposition, unequal recombination, mutations, etc.) to the phenotype variation (Hurles et al. 2008). It is also the ground for the development of efficient tools for the management and use of clonal variation (Benjak et al. 2008; Schellenbaum et al. 2008). The recent genome wide analysis of the class II TEs present in the grapevine genome showed a contribution of these elements to the high polymorphism observed among grapevine varieties (Benjak et al. 2008); these elements are therefore good candidates for scavenging the clonal variation within *V. vinifera* varieties. Finally, it allows studying the epigenetic regulation of grapevine at a genome scale and its interplay with the structural variations to shape the polymorphism in grapevine (Vaughn et al. 2007; Beck and Rakhyan 2008).

References

Adam-Blondon A-F, Bernole A, Faes G, Lamoureux D, Pateyron S, Grando MS, Caboche M, Velasco R, Chalhoub B (2005) Construction and characterization of BAC libraries from major grapevine cultivars. Theor Appl Genet 110: 1363–1371.

Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature 408: 796–815.

Aradhya MK, Dangl GS, Prins BH, Boursiquot J-M, Walker MA, Meredith CP, Simon CJ (2003) Genetic structure and differentiation in cultivated grape, *Vitis vinifera* L. Genet Res Camb 81: 179–192.

Aubourg S, Lecharny A, Bohlman J (2002) Genomic analysis of the terpenoid synthase (AtTPS) gene family of Arabidopsis thaliana. Mol Genet Genom 267: 730–747.

Aubourg S, Brunaud V, Bruyère C, Cock M, Cooke R, Cottet A, Couloux A, Déhais P, Deléage G, Drouard L, Duclert A, Echeverria M, Eschbach A, Falconet D, Filippi G, Gaspin C, Geourjon C, Grienenberger J-M, Guermann B, Houlné G, Jamet E, Lechauve F, Leleu O, Leroy P, Mache R, Meyer C, Nedjari H, Negrutiu I, Orsini V, Peyretaillade E, Pommier C, Raes J, Risler J-L. Rivière S, Rombauts S, Rouzé P, Schneider M, Schwob P, Small I, Soumayet-Kampetenga G, Stankovski D, Toffano C, Tognolli M, Caboche M, Lecharny A (2005) The GENEFARM project: structural and functional annotation of Arabidopsis gene and protein families by a network of experts. Nucl Acids Res 33: D641–D646.

Aury JM, Jaillon O, Duret L, Noel B, Jubin C, Porcel BM, Ségurens B, Daubin V, Anthouard V, Aiach N, Arnaiz O, Billaut A, Beisson J, Blanc I, Bouhouche K, Câmara F, Duharcourt S, Guigo R, Gogendeau D, Katinka M, Keller AM, Kissmehl R, Klotz C, Koll F, Le Mouël A, Lepère G, Malinsky S, Nowacki M, Nowak JK, Plattner H, Poulain J, Ruiz F, Serrano V, Zagulski M, Dessen P, Bétermier M, Weissenbach J, Scarpelli C, Schächter V, Sperling L, Meyer E, Cohen J, Wincker P (2006) Global trends of whole-genome duplications revealed by the ciliate Paramecium tetraurelia. Nature 444: 171–178.

Barker CL, Donald T, Pauquet J, Ratnaparkhe A, Bouquet A, Adam-Blondon A-F, Thomas MR, Dry I (2005) Genetic and physical mapping of the grapevine powdery mildew resistance gene, *Run1*, using a bacterial artificial chromosome library. Theor Appl Genet 111: 370–377.

Beck S, Rakyan VK (2008) The Methylome: approaches for global DNA methylation profiling. Trends Genet 24: 231–237.

Benjak A, Forneck A, Casacuberta JM (2008) Genome-wide analysis of the 'Cut-and-Paste" transposons of grapevine. PLOSone 3: e3107.

Bogs J, Ebadi A, McDavid D, Robinson SP (2006) Identification of the Flavonoid Hydroxylases from grapevine and their regulation during fruit development. Plant Physiol 140: 279–291.

Bowers JE, Chapman BA, Rong J, Paterson AH (2003) Unravelling Angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. Nature 422: 433–438.

Bronner A, Oliveira J (1990) Creation and study of the Pinot Noir variety lineage. In: Proc 5th Int Symp Grape Breeding, St Martin/Pflaz, Germany, Sept 1989, Vitis, spl issue, pp 69–80.

Brunner S, Fengler K, Morgante M, Tingey S, Rafalski A (2005) Evolution of DNA sequence non homologies among maize inbreds. Plant Cell 17: 343–360.

Castellarin SD, Di Gaspero G, Marconi R, Nonis A, Peterlunger E, Paillard S, Adam-Blondon A-F, Testolin R (2006) Colour variation in red grapevines (*Vitis vinifera* L.): genomic organisation, expression of flavonoid 3′-hydroxylase, flavonoid 3′, 5′-hydroxylase genes and related metabolite profiling of red cyanidin-/blue delphinidin-based anthocyanins in berry skin. BMC Genom 7: 12.

Castelli V, Aury JM, Jaillon O, Wincker P, Clepet C, Menard M, Cruaud C, Quetier F, Scarpelli C, Schachter V, Temple G, Caboche M, Weissenbach J, Salanoubat M (2004) Whole genome sequence comparisons and "full-length" cDNA sequences: a combined approach to evaluate and improve Arabidopsis genome annotation. Genome Res 14: 406–413.

Chandler M, Mahillon J (2002) Mobile DNA II. In: NL Craig, R Craigie, M Gellert, AM Lambowitz (eds) Insertion Sequences revisited. Am Soc Microbiol. Washington DC, USA, pp 305–366.

Clark RM, Schweikert G, Toomajian C, Ossowski S, Zeller G, Shinn P, Warthmann N, Hu TT, Fu G, Hinds DA, Chen H, Frazer KA, Huson DH, Schölkopf B, Nordborg M, Rätsch G, Ecker JR, Weigel D (2007) Common sequence polymorphisms shaping genetic diversity in *Arabidopsis thaliana*. Science 317: 338–342.

Claverie M, Dirlewanger E, Cosson P, Bosselut N, Lecouls AC, Voisin R, Kleinhentz M, Lafargue B, Caboche M, Chalhoub B, Esmenjaud D (2004) High-resolution mapping and chromosome landing at the root-knot nematode resistance locus Ma from Myrobolan plum using a large-insert BAC DNA library. Theor Appl Genet 109: 1318–1327.

Denoeud F, Aury J-M, Da Silva C, Noel B, Rogier O, Delledonne M, Morgante M, Valle G, Wincker P, Scarpelli C, Jaillon O, Artiguenave F (2008) Annotating genomes with massive-scale RNA-sequencing. Genome Biol 9: R175.

Dezulian T, Remmert M, Palatnik JF, Weigel D, Huson DH (2006) Identification of plant microRNA homologs. Bioinformatics 22: 359–360.

Doligez A, Adam-Blondon A-F, Cipriani G, Di Gaspero G, Laucou V, Merdinoglu D, Meredith CP, Riaz S, Roux C, This P (2006) An integrated SSR map of grapevine based on five mapping populations. Theor Appl Genet 113: 369–382.

Fridman E, Pleban T, Zamir D (2000) A recombination hotspot delimits a wild-species quantitative trait locus for tomato sugar content to 484 bp within an invertase gene. Proc Natl Acad Sci USA 97: 4718–4723.

Goff SA, Ricke D, Lan TH, Presting G, Wang R, Dunn M, Glazebrook J, Sessions A, Oeller P, Varma H, Hadley D, Hutchison D, Martin C, Katagiri F, Lange B.M, Moughamer T, Xia Y, Budworth P, Zhong J, Miguel T, Paszkowski U, Zhang S, Colberrt M, Sun WL, Chen L, Cooper B, Park S, Wood TC, Mao L, Quail P, Wing R, Dean R, Yu Y, Zharkikh A, Shen R, Sahasrabudhe S, Thomas A, Cannings R, Gutin A, Pruss D, Reid J, Tavtigian S, Mitchell J, Eldredge G, Scholl T, Miller RM, Batnagar S, Adey N, Rubano T, Tusneem N, Robinson R, Feldhaus J, Macalma T, Oliphant A, Briggs S (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). Science 296: 92–100.

Green ED (2001) Strategies for the systematic sequencing of complex genomes. Nat Rev 2: 573–583.

Han Y, Gasic K, Marron B, Beever JE, Korban SS (2007) A BAC-based physical map of the apple genome. Genomics 89: 630–637.

Huang S, Li R, Zhang Z, Li L, Gu X, Fan W, Lucas WJ, Wang X, Xie B, Ni P, Ren Y, Zhu H, Li J, Lin K, Jin W, Fei Z, Li G, Staub J, Kilian A, van der Vossen EAG, Wu Y, Guo J, He J, Jia Z, Ren Y, Tian G, Lu Y, Ruan J, Qian W, Wang M, Huang Q, Li B, Xuan Z, Cao J, Asan, Wu Z, Zhang J, Cai Q, Bai Y, Zhao B, Han Y, Li Y, Li X, Wang S, Shi Q, Liu S, Cho WK, Kim J-Y, Xu Y, Heller-Uszynska K, Miao H, Cheng Z, Zhang S, Wu J, Yang Y, Kang H, Man Li, Liang H, Ren X, Shi Z, Wen M, Jian M, Yang H, Zhang G, Yang Z, Chen R, Liu S, Li J, Ma L, Liu H, Zhou Y, Zhao J, Fang X, Li G, Fang L, Li Y, Liu D, Zheng H, Zhang Y, Qin N, Li Z, Yang G, Yang S, Bolund L, Kristiansen K, Zheng H, Li S, Zhang X, Yang H, Wang J, Sun R, Zhang B, Jiang S, Wang J, Du Y, Li S (2009) The genome of the cucumber, *Cucumis sativus* L. Nat Genet 41: 1275–1282.

Hurles ME, Dermitzakis ET, Tyler-Smith C (2008) The functional impact of structural variation in humans. Trends Genet 24: 238–245.

International Rice Genome Sequencing Project -IRGSP-(2005) The map-based sequence of the Rice genome. Nature 436: 793–800.

Jaillon O, Aury JM, Brunet F, Petit JL, Stange-Thomann N, Mauceli E, Bouneau L, Fischer C, Ozouf-Costaz C, Bernot A, Nicaud S, Jaffe D, Fisher S, Lutfalla G, Dossat C, Segurens B, Dasilva C, Salanoubat M, Levy M, Boudet N, Castellano S, Anthouard V, Jubin C, Castelli V, Katinka M, Vacherie B, Biemont C, Skalli Z, Cattolico L, Poulain J, De Berardinis V, Cruaud C, Duprat S, Brottier P, Coutanceau JP, Gouzy J, Parra G, Lardier G, Chapple C, McKernan KJ, McEwan P, Bosak S, Kellis M, Volff JN, Guigo R, Zody MC, Mesirov J, Lindblad-Toh K, Birren B, Nusbaum C, Kahn D, Robinson-Rechavi M, Laudet V, Schachter V, Quetier F, Saurin W, Scarpelli C, Wincker P, Lander ES, Weissenbach J, Roest Crollius H (2004) Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. Nature 431: 946–957.

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

Jeong ST, Goto-Yamamoto N, Hashizume K, Esaka M (2005) Expression of the flavonoid 3′-hydroxylase and flavonoid composition in grape (*Vitis vinifera*). Plant Sci 170: 61–69.

Jung S, Jesudurai C, Staton M, Du Z, Ficklin S, Cho I, Abbott A, Tomkins J, Main D (2004) GDR (Genome Database for *Rosaceae*): integrated web resources for *Rosaceae* genomics and genetics research. BMC Bioinformat 5: 130.

Kelleher CT, Chiu R, Shin H, Bosdet IE, Krzywinski MI, Fjell CD, Wilkin J, Yin T, DiFazio SP, Ali J, Asano JK, Chan S, Cloutier A, Girn N, Leach S, Lee D, Mathewson CA, Olson T, O'connor K, Prabhu AL, Smailus DE, Stott JM, Tsai M, Wye NH, Yang GS, Zhuang J, Holt RA, Putnam NH, Vrebalov J, Giovannoni JJ, Grimwood J, Schmutz J, Rokhsar D, Jones SJ, Marra MA, Tuskan GA, Bohlmann J, Ellis BE, Ritland K, Douglas CJ, Schein JE (2007) A physical map of the highly heterozygous *Populus* genome: integration with the genome sequence and genetic map and analysis of haplotype variation. Plant J 50: 1063–1078.

Lamoureux D, Bernole A, Le Clainche I, Tual, S, Thareau V, Paillard S, Legeai F, Dossat C, Wincker P, Oswald M, Merdinoglu D, Vignault C, Delrot S, Caboche M, Chalhoub B, Adam-Blondon A-F (2006) Anchoring a large set of markers onto a BAC library for the development of a draft physical map of the grapevine genome. Theor Appl Genet 113: 344–356.

Lindblad-Toh K, Wade CM, Mikkelsen TS, Karlsson EK, Jaffe DB, Kamal M, Clamp M, Chang JL, Kulbokas III EJ, Zody MC, Mauceli E, Xie X, Breen M, Wayne RK, Ostrander EA, Ponting CP, Galiber F, Smith DR, deJong PJ, Kirkness E, Alvarez P, Biagi T, Brockman W, Butler J, Chin CW, Cook A, Cuff J, Daly MJ, DeCaprio D, Gnerre S, Grabherr M, Kellis

M, Kleber M, Bardeleben C, Goodstadt L, Heger A, Hitte C, Kim L, Koepfli KP, Parker HG, Pollinger JP, Searle SMJ, Sutter NB, Thomas R, Weber C, Broad Sequencing Platform members, Lander ES. (2005) Genome sequence, comparative analysis and haplotype structure of the domestic dog. Nature 438: 803–819.

Luo MC, Thomas C, You FM, Hsiao J, Ouyang S, Buell CR, Malandro M, McGuire PE, Anderson OD, Dvorak J (2003) High-throughput fingerprinting of bacterial artificial chromosomes using the snapshot labeling kit and sizing of restriction fragments by capillary electrophoresis. Genomics 82: 378–389.

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer ML, Jarvie TP, Jirage KB, Kim JB, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu P, Begley RF, Rothberg JM (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437: 376–380.

Menda N, Buels RM, Tecle I, Mueller LA (2008) A community-based annotation framework for linking Solanaceae genomes with phenomes. Plant Physiol 147: 1788–1799.

Meyers BC, Scalabrin S, Morgante M (2004) Mapping and sequencing complex genomes: let's get physical! Nat Rev Genet 5: 578–588.

Ming R, Hou S, Feng Y, Yu Q, Dionne-Laporte A, Saw JH, Senin P, Wang W, Ly BV, Lewis KL, Salzberg SL, Feng L, Jones MR, Skelton RL, Murray JE, Chen C, Qian W, Shen J, Du P, Eustice M, Tong E, Tang H, Lyons E, Paull RE, Michael TP, Wall K, Rice DW, Albert H, Wang ML, Zhu YJ, Schatz M, Nagarajan N, Acob RA, Guan P, Blas A, Wai CM, Ackerman CM, Ren Y, Liu C, Wang J, Wang J, Na JK, Shakirov EV, Haas B, Thimmapuram J, Nelson D, Wang X, Bowers JE, Gschwend AR, Delcher AL, Singh R, Suzuki JY, Tripathi S, Neupane K, Wei H, Irikura B, Paidi M, Jiang N, Zhang W, Presting G, Windsor A, Navajas-Pérez R, Torres MJ, Feltus FA, Porter B, Li Y, Burroughs AM, Luo MC, Liu L, Christopher DA, Mount SM, Moore PH, Sugimura T, Jiang J, Schuler MA, Friedman V, Mitchell-Olds T, Shippen DE, dePamphilis CW, Palmer JD, Freeling M, Paterson AH, Gonsalves D, Wang L, Alam M (2008) The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). Nature 452: 991–996.

Morgante M, Salamini F (2003) From plant genomics to breeding practice. Curr Opin Biotechnol 14: 214–219.

Morgante M, Brunner S, Pea G, Fengler K, Zuccolo A, Rafalski A (2005) Gene duplication and exon shuffling by helitron-like transposons generate intraspecies diversity in maize. Nat Genet 37: 997–1002.

Moroldo M, Paillard S, Marconi R, Legeai F, Canaguier A, Cruaud C, De Berardinis V, Guichard C, Brunaud V, Le Clainche I, Scalabrin S, Testolin R, Di Gaspero G, Morgante M, Adam-Blondon AF (2008) A physical map of the heterozygous grapevine 'Cabernet Sauvignon' allows mapping candidate genes for disease resistance. BMC Plant Biol 8: 66.

Myles S, Chia J-M, Hurwitz B, Simon C, Zhong GY, Buckler E, Ware D (2010) Rapid Genomic Characterization of the Genus Vitis. PLOS ONE 5: e8219.

Nelson W, Soderlund C (2005) Software for restriction fragment physical maps. In: K Meksem, G Kahl (eds) The Handbook of Genome Mapping: Genetic and Physical Mapping. (eds K. Meksem and G. Kahl), Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, FRG pp 285–306 .

Nelson WM, Dvorak J, Luo MC, Messing J, Wing RA, Soderlund C (2007) Efficacy of clone fingerprinting methodologies. Genomics 89: 160–166.

Ohno S (1970) Evolution by Gene Duplication. Springer-Verlag, New York, USA.

Patel GI, Olmo HP (1955) Cytogenetics of Vitis 1 The hybrid *V. vinifera* x *V. rotundifolia*. Am J Bot 42: 141–159.

Patocchi A, Vinatzer BA, Gianfranceschi L, Tartarini S, Zhang HB, Sansavini S, Gessler C (1999) Construction of a 550 kb BAC contig spanning the genomic region containing the apple scab resistance gene Vf. Mol Gen Genet 4–5: 884–891.

Pindo M, Vezzulli S, Coppola G, Cartwright D, Zharkikh A, Velasco R, Troggio M (2008) SNP High throughput screening in grapevine using the SNPlexTM genotyping system. BMC Plant Biol 8: 12.

Rafalski A (2002) Applications of single nucleotide polymorphisms in crop genetics. Curr Opin Plant Biol 5: 94–100.

Rensing RA, Lang DD, Zimmer AD, Terry A, Salamov A, Shapiro H, Nishiyama T, Perroud P-F, Lindquist EA, Kamisugi Y, Tanahashi T, Sakakibara K, Fujita T, Oishi K, Shin-IT, Kuroki Y, Toyoda A, Suzuki Y, Hashimoto S-I, Yamaguchi K, Sugano S, Kohara Y, Fujiyama A, Anterola A, Aoki S, Ashton N, Barbazuk WB, Barker E, Bennetzen JL, Blankenship R, Cho SH, Dutcher SK, Estelle M, Fawcett JA, Gundlach H, Hanada K, Heyl A, Hicks KA, Hughes J, Lohr M, Mayer K, Melkozernov A, Murata T, Nelson DR, Pils B, Prigge M, Reiss B, Renner T, Rombauts S, Rushton PJ, Sanderfoot A, Schween G, Shiu S-H, Stueber K, Theodoulou FL, Tu H, Van de Peer Y, Verrier PJ, Waters E, Wood A, Yang L, Cove D, Cuming AC, Hasebe M, Lucas S, Mishler BD, Reski R, Grigoriev IV, Quatrano RS, Boore JL (2008) The *Physcomitrella* genome reveals evolutionary insights into the conquest of the land by plants. Science 319: 64–69.

Roest Crollius H, Jaillon O, Bernot A, Dasilva C, Bouneau L, Fischer C, Fizames C, Wincker P, Brottier P, Quétier F, Saurin W, Weissenbach J (2000) Human gene number estimate provided by genome wide analysis using *Tetradon nigroviridis* genomic DNA. Nature Genet 25: 235–238.

Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA 74(12): 5463–5467.

Scalabrin S, Troggio M, Moroldo M, Pindo M, Felice N, Coppola G, Prete G, Malacarne G, Marconi R, Faes G, Jurman I, Grando S, Jesse T, Segala C, Valle G, Policriti A, Fontana P, Morgante M, Velasco R (2010) Physical mapping in highly heterozygous genomes: a physical contig map of the Pinot Noir grapevine cultivar BMC Genom 11: 204.

Schellenbaum P, Mohler V, Wenzel G, Walter B (2008) Variation in DNA methylation patterns of grapevine somaclones (*Vitis vinifera* L.). BMC Plant Biol 8: 78.

Shendure J, Mitra RD, Varma C, Church GM (2004) Advanced sequencing technologies: methods and goals. Nat Rev Genet 5: 335–344.

Tomkins JP, Peterson DG, Yang TJ, Ablett ER, Henry RJ, Lee LS, Holton TA, Waters D, Wing RA (2001) Grape (*Vitis vinifera* L.) BAC library construction, preliminary STC analysis, and identification of clones associated with flavonoid and stilbene biosynthesis. Am J Enol Vitic 52: 287–291.

Town CD (2006) Annotating the genome of *Medicago truncatula.* Curr Opin Plant Biol 9: 122–127.

Thornsberry JM, Goodman MM, Doebley J, Kresovich S, Nielsen D, Buckler IV ES (2001) *Dwarf8* polymorphisms associate with variation in flowering time. Nat Genet 28: 286–289.

Troggio M, Malacarne G, Coppola G, Segala C, Cartwright DA, Pindo M, Stefanini M, Mank R, Moroldo M, Morgante M, Grando MS, Velasco R (2007) A dense single-nucleotide polymorphism-based genetic linkage map of grapevine (*Vitis vinifera* L.) anchoring Pinot Noir bacterial artificial chromosome contigs. Genetics 176: 2637–2650.

Tuskan GA, Difazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov A, Schein J, Sterck L, Aerts A, Bhalerao RR, Bhalerao RP, Blaudez D, Boerjan W, Brun A, Brunner A, Busov V, Campbell M, Carlson J, Chalot M, Chapman J, Chen GL, Cooper D, Coutinho PM, Couturier J, Covert S, Cronk Q, Cunningham R, Davis J, Degroeve S, Déjardin A, Depamphilis C, Detter J, Dirks B, Dubchak I, Duplessis S, Ehlting J, Ellis B, Gendler K, Goodstein D, Gribskov M, Grimwood J, Groover A, Gunter L, Hamberger B, Heinze B, Helariutta Y, Henrissat B, Holligan D, Holt R, Huang W, Islam-Faridi N, Jones S, Jones-Rhoades M, Jorgensen R, Joshi C, Kangasjärvi J, Karlsson J, Kelleher C, Kirkpatrick R, Kirst M, Kohler A, Kalluri U, Larimer F, Leebens-Mack

J, Leplé JC, Locascio P, Lou Y, Lucas S, Martin F, Montanini B, Napoli C, Nelson DR, Nelson C, Nieminen K, Nilsson O, Pereda V, Peter G, Philippe R, Pilate G, Poliakov A, Razumovskaya J, Richardson P, Rinaldi C, Ritland K, Rouzé P, Ryaboy D, Schmutz J, Schrader J, Segerman B, Shin H, Siddiqui A, Sterky F, Terry A, Tsai CJ, Uberbacher E, Unneberg P, Vahala J, Wall K, Wessler S, Yang G, Yin T, Douglas C, Marra M, Sandberg G, Van de Peer Y, Rokhsar D (2006) The genome of black cottonwood, *Populus trichocarpa* (Torr.& Gray). Science 313: 1596–1604.

Vaughn MW, Tanurd IcM, Lippman Z, Jiang H, Carrasquillo R, Rabinowicz PD, Dedhia N, McCombie WR, Agier N, BulskiA, Colot V, Doerge RW, Martienssen RA (2007) Epigeneic natural variation in Arabidopsis thaliana. PLOS Biol 5: e174.

Velasco R, Zharkikh A, Troggio M, Cartwright DA, Cestaro A, Pruss D, Pindo M, Fitzgerald LM, Vezzulli S, Reid J, Malacarne G, Iliev D, Coppola G, Wardell B, Micheletti D, Macalma T, Facci M, Mitchell JT, Perazzolli M, Eldredge G, Gatto P, Oyzerski R, Moretto M, Gutin N, Stefanini M, Chen Y, Segala C, Davenport C, Demattè L, Mraz A, Battilana J, Stormo K, Costa F, Tao Q, Si-Ammour A, Harkins T, Lackey A, Perbost C, Taillon B, Stella A, Solovyev V, Fawcett JA, Sterck L, Vandepoele K, Grando SM, Toppo S, Moser C, Lanchbury J, Bogden R, Skolnick M, Sgaramella V, Bhatnagar SK, Fontana P, Gutin A, Van de Peer Y, Salamini F, Viola R (2007) A high quality draft consensus sequence of the genome of a heterozygous grapevine variety. PLoS ONE 2: e1326.

Vezzulli S, Micheletti D, Riaz R, Pindo M, Viola R, This P, Walker MA, Troggio M, Velasco R (2008) A SNP transferability survey within the genus *Vitis*. BMC Plant Biol 8: 128.

Vinson JP, Jaffe DB, O'Neill K, Karlsson EK, Stange-Thomann N, Anderson S, Mesirov JP, Satoh N, Satou Y, Nusbaum C, Birren B, Galagan JE, Lander ES (2005) Assembly of polymorphic genomes: algorithms and application to Ciona savignyi. Genome Res 15: 1127–1135.

Yu J, Hu S, Wang J, Wong GK-S, Li S, Liu B, Deng Y, Dai L, Zhou Y, Zhang X, Cao M, Liu J, Sun J, Tang J, Chen Y, Huang X, Lin W, Ye C, Wei Tong W, Cong L, Geng G, Han Y, Li L, Li W, Hu G, Huang X, Li W, Li J, Liu Z, Li L, Liu J, Qi Q, Liu J, Li L, Li T, Wang X, Lu H, Wu T, Zhu M, Ni P, Han H, Dong W, Ren X, Feng X, Cui P, Li X, Wang H, Xu X, Zhai W, Xu Z, Zhang J, He S, Zhang J, Xu J, Zhang K, Zheng X, Dong J, Zeng W, Tao L, Ye J, Tan J, Ren X, Chen X, He J, Liu D, Tian W, Tian C, Xia H, Bao Q, Li G, Gao H, Cao T, Wang J, Zhao W, Li P, Chen W, Wang X, Zhang Y, Hu J, Wang J,Liu S, Yang J, Zhang G, Xiong Y, Li Z, Mao L, Zhou C, Zhu Z, Chen R, Hao B, Zheng W, Chen S, Guo W, Li G, Liu S, Tao M, Wang J, Zhu L, Yuan L, Yang H (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). *Science* 296: 79–92.

Zharkikh A, Troggio M, Pruss D, Cestaio A, Eldrdge G, Pindo M, Mitchell JT, Vezzulli S, Bhatnagar S, Fontana P, Viola R, Gutin A, Salamini F, Skolnick M, Velasco R (2008) Sequencing and assembly of highly heterozygous genome of *Vitis vinifera*: problems and solutions. J Biotechnol 36: 38–43.