LETTERS

The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla

The French–Italian Public Consortium for Grapevine Genome Characterization*

The analysis of the first plant genomes provided unexpected evidence for genome duplication events in species that had previously been considered as true diploids on the basis of their genetics[1–3]. These polyploidization events may have had important consequences in plant evolution, in particular for species radiation and adaptation and for the modulation of functional capacities[4–10]. Here we report a high-quality draft of the genome sequence of grapevine (*Vitis vinifera*) obtained from a highly homozygous genotype. The draft sequence of the grapevine genome is the fourth one produced so far for flowering plants, the second for a woody species and the first for a fruit crop (cultivated for both fruit and beverage). Grapevine was selected because of its important place in the cultural heritage of humanity beginning during the Neolithic period[11]. Several large expansions of gene families with roles in aromatic features are observed. The grapevine genome has not undergone recent genome duplication, thus enabling the discovery of ancestral traits and features of the genetic organization of flowering plants. This analysis reveals the contribution of three ancestral genomes to the grapevine haploid content. This ancestral arrangement is common to many dicotyledonous plants but is absent from the genome of rice, which is a monocotyledon. Furthermore, we explain the chronology of previously described whole-genome duplication events in the evolution of flowering plants.

All grapevine varieties are highly heterozygous; preliminary data showed that there was as much as 13% sequence divergence between alleles, which would hinder reliable contig assembly when a whole-genome shotgun strategy was used for sequencing. Our consortium therefore selected the grapevine PN40024 genotype for sequencing. This line, originally derived from Pinot Noir, has been bred close to full homozygosity (estimated at about 93%) by successive selfings, permitting a high-quality whole-genome shotgun assembly.

A total of 6.2 million end-reads were produced by our consortium, representing an 8.4-fold coverage of the genome. Within the assembly, performed with Arachne[12], 316 supercontigs represent putative allelic haplotypes that constitute 11.6 million bases (Mb). These values are in good fit with the 7% residual heterozygosity of PN40024 assessed by using genetic markers. When considering only one of the haplotypes in each heterozygous region, the assembly (Table 1a) consists of 19,577 contigs (N_{50} = 65.9 kilobases (kb), where N_{50} corresponds to the size of the shorter supercontig or contig in a subset representing half of the assembly size) and 3,514 supercontigs (N_{50} = 2.07 Mb) totalling 487 Mb. This value is close to the 475 Mb previously reported for the grapevine genome size[13].

Using a set of 409 molecular markers from the reference grapevine map[14], 69% of the assembled 487 Mb, arranged into 45 ultracontigs

Table 1 | Global statistics on the genome of *Vitis vinifera*

(a) Assembly

	Status	Number	N_{50} (kb)	Longest (kb)	Size (Mb)	Percentage of the assembly
Contigs	All	19,577	65.9	557	467.5	–
Supercontigs	All	3,514	2,065	12,675	487.1	100
	Anchored on chromosomes	191	3,189	12,675	335.6	68.9
	Anchored on chromosomes and oriented	143	3,827	12,675	296.9	60.9

(b) Annotation

	Number	Median size (bp)	Total length (Mb)	Percentage of the genome	%GC
Gene	30,434	3,399	225.6	46.3	36.2
Exons CDS	149,351	130	33.6	6.9	44.5
Introns CDS	118,917	213	178.6	36.7	34.7
Intergenic	30,453	3,544	261.5	34.7	33.0
tRNA*	600	73	0.04	NS	43.0
miRNA†	164	103.5	0.002	NS	35.9

(c) Orthology

	Number of orthologous proteins	Mean identity (%)
P. trichocarpa	12,996	72.7
A. thaliana	11,404	65.5
O. sativa	9,731	59.8
Common to eudicotyledons‡	10,547	
Common to Magnoliophyta§	8,121	

* Transfer RNA (tRNA) values were computed on exons.
† Micro RNAs (miRNAs) are members of known conserved miRNA families.
‡ Eudicotyledons are represented by *P. trichocarpa* and *A. thaliana*.
§ Magnoliophyta (most flowering plants) are represented by *P. trichocarpa*, *A. thaliana* and *O. sativa*.

*A list of participants and their affiliations appears at the end of the paper.

and 51 single supercontigs, were anchored along the 19 linkage groups. Thirty-seven ultracontigs and 22 single supercontigs were oriented, representing 61% of the genome assembly (Supplementary Tables 2 and 3).

This assembly has been annotated by using a combination of evidence. The major features of the genome annotation are presented in Table 1b. The 8.4-fold draft sequence of the grapevine genome contains a set of 30,434 protein-coding genes (an average of 372 codons and 5 exons per gene). This value is considerably lower than the 45,555 protein-coding genes reported for the poplar (*Populus trichocarpa*) genome, which has a similar size, at 485 Mb (ref. 1), and even lower than the 37,544 protein-coding genes identified in the 389 Mb of the rice genome[2].

Three different approaches revealed that 41.4% (average value) of the grapevine genome is composed of repetitive/transposable elements (TEs), a slightly higher proportion than that identified in the rice genome, which has a somewhat smaller size[2]. The distribution of repeats and TEs along the chromosomes is quite uneven (see below). All classes and superfamilies of TEs are represented in the grapevine genome, with a large prevalence of class I elements over class II and helitrons (rolling-circle transposons) (Supplementary Table 7). An analysis of the distribution of the repetitive elements in the different fractions of the grapevine genome based on the current annotation shows that introns are quite rich in repeats and TEs (data not shown). In addition, 12.4% of the intron sequence contains transposons as determined using our set of manually annotated elements, most of which (75%) correspond to LINE (long interspersed element) retrotransposons, which therefore seem to have contributed specifically to the intron size observed in grapevine (Supplementary Table 8).

In eukaryotes with large genomes, the coding and repeated elements are distributed over the chromosomes and may be more or less interlaced, hence defining gene-poor and gene-rich regions. It has previously been noticed that the distribution of the genes along the chromosomes of rice and *Arabidopsis thaliana* is fairly homogeneous[2,3]. In contrast, we observe large regions that alternate between high and low gene density in *V. vinifera* (Supplementary Figs 2 and 3). As expected, the density of TEs reflects a pattern substantially complementary to gene density. We observe a similar characteristic in the genome sequence of poplar, therefore indicating a dynamic for the invasion of TEs that is shared with the grapevine (Supplementary Fig. 3).

A striking feature of the grapevine proteome lies in the existence of large families related to wine characteristics, which have a higher gene copy number than in the other sequenced plants. Stilbene synthases (STSs) drive the synthesis of resveratrol, the grapevine phytoalexin that has been associated with the health benefits associated with moderate consumption of red wine[15,16]. The family of genes encoding STSs has a noticeable expansion: 43 genes have been identified. Of these, 20 have previously been shown to be expressed after infection by *Plasmopara viticola*, thus confirming that they are likely to be functional. The terpene synthases (TPSs) drive the synthesis of terpenoids; these secondary metabolites are major components of resins, essential oils and aromas (their relative abundance is directly correlated with the aromatic features of wines[17]) and are involved in plant–environment interactions. In comparison with the 30–40 genes of this family in *Arabidopsis*, rice and poplar, the grapevine TPS family is more than twice as large, with 89 functional genes and 27 pseudogenes. Classification based on known plant homologues reveals that the subclass of putative monoterpene synthases represents only 15% of the *Arabidopsis* TPS family[18] whereas this subclass represents 40% of the grapevine TPS family. This result suggests a high diversification of grapevine monoterpene synthases that specifically produce C_{10} terpenoids present in aroma (such as geraniol, linalool, cineole and α-terpineol). Furthermore, the grapevine genome annotation has also revealed genes encoding homologues to the two forms of geranyl diphosphate synthases (GPPSs), the enzymes that produce the substrate for monoterpene synthases: both the homodimeric GPPS and the heterodimeric form are present; the latter is present only in plants such as *Mentha piperita* and *Clarkia breweri*, which produce large quantities of monoterpenes[19]. Most of the STS and TPS genes occur as 20 clusters, including up to 33 paralogous genes located in a 680-kb stretch.

Because global duplication events seem to be a frequent event in plant evolution[20], we searched the genome of *V. vinifera* for paralogous regions by using protein sequence similarity. Paralogous regions are defined as chromosome fragments in which homologous genes are present in clusters. Statistical analysis[21] of these clusters reveals that 94.5% have high probability of being paralogous ($P < 10^{-4}$; Supplementary Table 11). Most *Vitis* gene regions have two different paralogous regions, which we have grouped together as triplets (Supplementary Fig. 5; coverage details in Supplementary Table 10). We conclude that the present-day grapevine haploid genome originated from the contribution of three ancestral genomes. It is yet to be demonstrated whether this content came from a true hexaploidization event or through successive genome duplications. The resulting plant had a diploid content that corresponds to the three full diploid contents of the three ancestors; it may therefore be described as a 'palaeo-hexaploid' organism. A number of rearrangements have affected the original three complements after the formation of the palaeo-hexaploid state. However, the gene order has been sufficiently conserved to permit the alignment of most regions with their two siblings.

We explored the time of formation of the palaeo-hexaploid arrangement by comparing grapevine gene regions with those of other completely sequenced plant genomes. If the palaeo-hexaploid complement is present in another species, it should result in a one-for-one pairing of gene regions between the two species considered. In contrast, if another species's genome evolved before palaeo-hexaploid formation, it should result in a one-to-three relationship between the other species and the grapevine genome. The available genome sequences were those of poplar[1], *Arabidopsis*[3] and rice (*Oryza sativa*[2]), of which poplar is considered to be most closely related to grapevine. All clusters constructed between the orthologues in the three comparisons have $P < 10^{-4}$ (Table 1c). When the gene order in poplar is compared with that in grapevine, there are two clear distributions. First, the grapevine regions align with two poplar segments, as would be expected from a recent whole-genome duplication (WGD) in the poplar lineage[1]. Second, each of the three grapevine regions that form a homologous triplet recognizes different pairs of poplar segments (Fig. 1a and Supplementary Fig. 6). This shows that the palaeo-hexaploidy observed in grapevine was already present in its common ancestor with poplar.

Poplar belongs to the Eurosid I clade. The sister clade to Eurosid I is that of Eurosid II, which contains the model species *Arabidopsis*. Its gene order was compared with that in the grapevine genome. Two distributions appear: first, most grapevine regions correspond to four *Arabidopsis* segments (Supplementary Fig. 7); second, each component of a triplicated group in grapevine recognizes four different regions in *Arabidopsis* (Fig. 1b). This shows that the grapevine palaeo-hexaploidy was present in the common ancestor to *Arabidopsis* and grapevine, and therefore that it is a trait common to all Eurosids. This is confirmed by the homology level distribution between paralogues of the grapevine, indicating a lower conservation than between *Vitis*/*Arabidopsis* orthologues (Supplementary Fig. 4). The Eurosid group contains many economically important flowering plants such as legumes, cotton and Brassicaceae. Our present results establish these species as having a palaeo-hexaploid common ancestor. The grapevine/*Arabidopsis* comparison also reveals that the *Arabidopsis* lineage underwent two WGDs after its separation from the Eurosid I clade[21–24]. This contradicts some models based on more indirect evidence that placed the most ancient of these two duplications at the base of the Eurosid group, or even earlier[4,20–22]. Some studies had also suggested a possible third duplication event in the distant past of the *Arabidopsis* lineage, potentially at the base of

the angiosperm radiation. The controversy about this third event is now resolved by the *Vitis* genome comparisons: this event corresponds to the palaeo-hexaploidy formation that remains evident in the grapevine genome but has been difficult to characterize in *Arabidopsis* and poplar because of the more recent WGDs. In particular, the *Arabidopsis* genome lineage has undergone many rearrangements and chromosome fusions such that the ancestral gene order is particularly difficult to deduce from this species (Fig. 2).

Grapevines, like *Arabidopsis* and poplar, are dicotyledonous plants that diverged from monocotyledons about 130–240 Myr ago[25,26]. Because rice is a monocotyledon, we assessed the presence or absence of palaeo-hexaploidy in its genome sequence. The observed pattern is the opposite of that seen for *Arabidopsis* and poplar: constituents of a grapevine triplet are generally orthologous to the same group of rice regions (Fig. 1c and Supplementary Fig. 11). Because rice and grapevine are phylogenetically distant, it is more difficult to detect relations of orthology across the two whole genomes: rearrangements, duplication and gene loss have affected the gene orders differently in the two lineages (Supplementary Fig. 10). Even with this limitation, we observed numerous cases of one-to-three relationships between

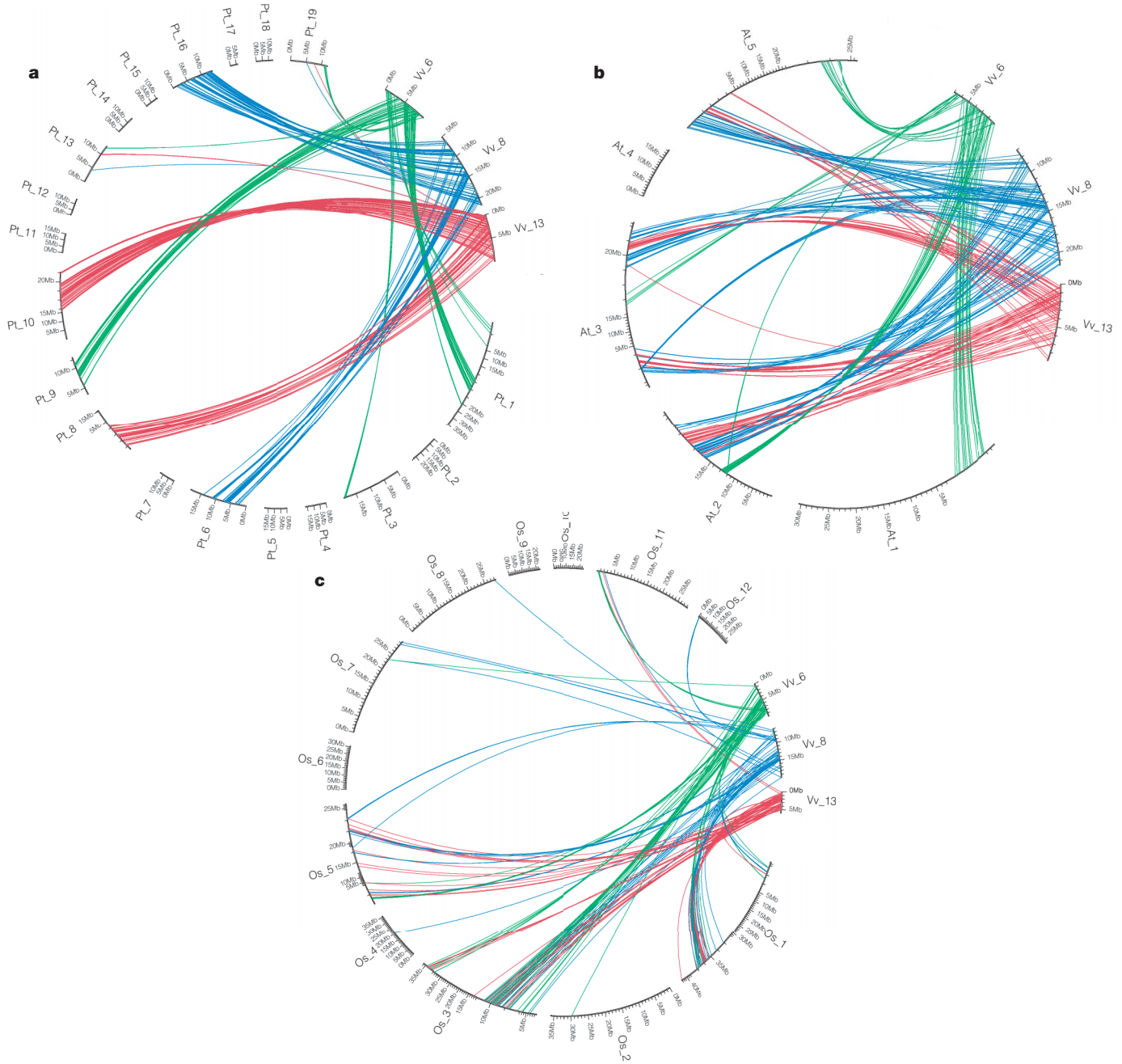

Figure 1 | Comparison between three paralogous *Vitis* genomic regions and their orthologues in *P. trichocarpa*, *A. thaliana* and *O. sativa*. Orthologous gene pairs are joined with a different colour for each of the three paralogous grapevine chromosomes 6 (green), 8 (blue) and 13 (red). **a**, Orthologous regions in the poplar genome are different for each of the three *Vitis* chromosomes, showing that the triplication predates the poplar/*Vitis* separation. One *Vitis* region recognizes two poplar segments because of a WGD in the poplar lineage after the separation. **b**, Orthologous regions with *Arabidopsis* are different for each of the three *Vitis* chromosomes. This shows that the *Arabidopsis*/*Vitis* ancestor had the same palaeo-hexaploid content. One *Vitis* region corresponds to four *Arabidopsis* segments, indicating the presence of two WGDs in the *Arabidopsis* lineage after separation from the *Vitis* lineage. **c**, Orthologous regions in rice are the same for the three paralogous chromosomes. This indicates that the triplication was not present in the common ancestor of monocotyledons and dicotyledons. The presence in rice of different homologous blocks is due to global duplications in the rice lineage after divergence from dicotyledons.

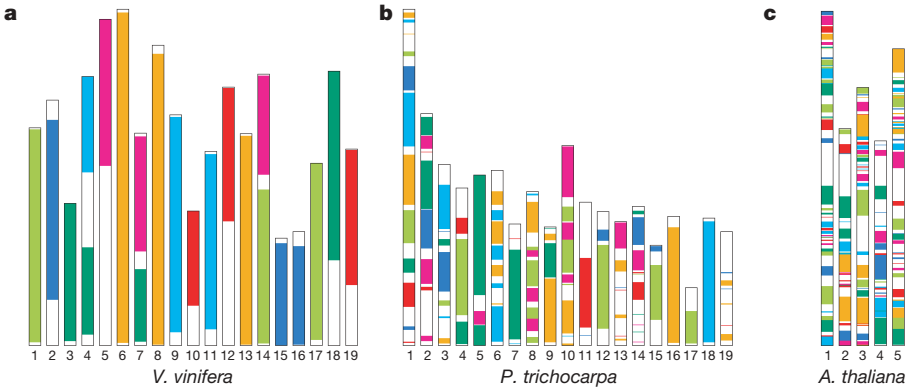

Figure 2 | Schematic representation of paralogous regions derived from the three ancestral genomes in the karyotypes of *V. vinifera*, *P. trichocarpa* and *A. thaliana*. Each colour corresponds to a syntenic region between the three ancestral genomes that were defined by their occurrence as linked clusters in grapevine, independently of intrachromosomal rearrangements. The *V. vinifera* genome (**a**) is by far the closest to the ancestral arrangement, whereas that of *Arabidopsis* (**c**) is thoroughly rearranged, and *P. trichocarpa* (**b**) presents an intermediate situation. The seven colours probably correspond to linkage groups at the time of the palaeo-hexaploid ancestor.

rice and grapevine (Supplementary Figs 8, 9 and 11); 23% of orthologous blocks include the paralogous regions that originate from the grapevine palaeo-hexaploidy. For *Arabidopsis*, this number is as low as 1.4% (this difference is significant at 5%: $\chi^2 = 8.9$; Supplementary Table 12), despite the fact that the *Arabidopsis* genome has suffered many gene losses since its two WGDs. These gene losses would be expected to obscure the orthologous relations with the grapevine genome, but they are clearly insufficient to explain the high number of one-to-three relationships observed in the rice–grapevine comparison. The most probable explanation for this excess is that the rice ancestor did not exhibit the palaeo-hexaploidy observed in the grapevine, poplar and *Arabidopsis*.

These findings are summarized in Fig. 3: the triplicated arrangement is apparent after the separation of the monocotyledons and dicotyledons and before the spread of the Eurosid clade. Future genome sequencing projects for other clades of dicotyledons, such as Solanaceae or basal eudicots, will help in situating the triplication event more precisely, and eventually in establishing its precise nature (hexaploidization or genome duplications at distant times).

Public access to the grapevine genome sequence will help in the identification of genes underlying the agricultural characteristics of this species, including domestication traits. A selective amplification of genes belonging to the metabolic pathways of terpenes and tannins has occurred in the grapevine genome, in contrast with other plant genomes. This suggests that it may become possible to trace the diversity of wine flavours down to the genome level. Grapevine is also a crop that is highly susceptible to a large diversity of pathogens including powdery mildew, oidium and Pierce disease. Other *Vitis* species such as *V. riparia* or *V. cinerea*, which are known to be resistant to several of these pathogens, are interfertile with *V. vinifera* and can be used for the introduction of resistance traits by advanced backcrosses[27] or by gene transfer. Access to the *Vitis* sequence and the exploitation of synteny will speed up this process of introgression of pathogen resistance traits. As a consequence of this, it is hoped that it will also prompt a strong decrease in pesticide use.

The high quality of the assembly, due mainly to the highly homozygous nature of the PN40024 line, enables the discovery of three ancestral genomes constituting the diploid content of grapevine. The Greek historian Thucydides wrote that Mediterranean people began to emerge from ignorance when they learnt to cultivate olives and grapes. This first characterization of the grapevine genome, with its indication of a palaeo-hexaploid ancestral genome for many dicotyledonous plants, addresses fundamental questions related to the origin and importance of this event in the history of flowering plants. Future work may help in correlating the differential fates of the three gene complements with phenotypic traits of dicotyledonous species.

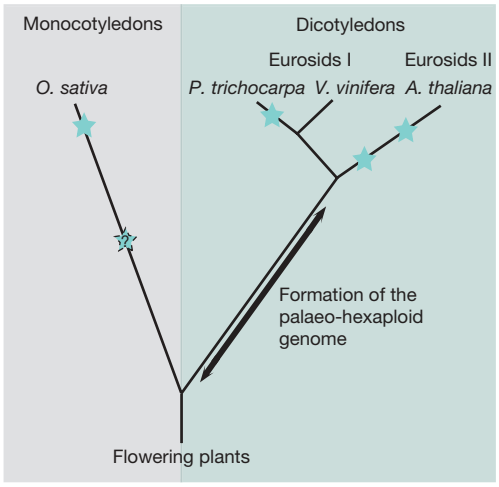

Figure 3 | Positions of the polyploidization events in the evolution of plants with a sequenced genome. Each star indicates a WGD (tetraploidization) event on that branch. The question mark indicates that ancient events are visible in the rice genome that would require other monocotyledon genome sequences to be resolved. The formation of the palaeo-hexaploid ancestral genome occurred after divergence from monocotyledons and before the radiation of the Eurosids.

METHODS SUMMARY

Gene annotation. Protein-coding genes were predicted by combining *ab initio* models, *V. vinifera* complementary DNA alignments, and alignments of proteins and genomic DNA from other species. The integration of the data was performed with GAZE[28]. Details are given in Supplementary Information.

Paralogous and orthologous gene sets. Statistical testing of homologous regions was performed as described in ref. 21.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 5 April; accepted 7 August 2007.
Published online 26 August 2007.

1. Tuskan, G. A. *et al.* The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* **313,** 1596–1604 (2006).
2. International Rice Genome Sequencing Project. The map-based sequence of the rice genome. *Nature* **436,** 793–800 (2005).
3. Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408,** 796–815 (2000).
4. De Bodt, S., Maere, S. & Van de Peer, Y. Genome duplication and the origin of angiosperms. *Trends Ecol. Evol.* **20,** 591–597 (2005).
5. Scannell, D. R., Byrne, K. P., Gordon, J. L., Wong, S. & Wolfe, K. H. Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. *Nature* **440,** 341–345 (2006).

6. Jaillon, O. *et al.* Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. *Nature* **431,** 946–957 (2004).
7. Aury, J. M. *et al.* Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia*. *Nature* **444,** 171–178 (2006).
8. Maere, S. *et al.* Modeling gene and genome duplications in eukaryotes. *Proc. Natl Acad. Sci. USA* **102,** 5454–5459 (2005).
9. Blanc, G. & Wolfe, K. H. Functional divergence of duplicated genes formed by polyploidy during *Arabidopsis* evolution. *Plant Cell* **16,** 1679–1691 (2004).
10. Seoighe, C. & Gehring, C. Genome duplication led to highly selective expansion of the *Arabidopsis thaliana* proteome. *Trends Genet.* **20,** 461–464 (2004).
11. McGovern, P. E., Hartung, U., Badler, V., Glusker, D. L. & Exner, L. J. The beginnings of wine making and viniculture in the anciant Near East and Egypt. *Expedition* **39,** 3–21 (1997).
12. Jaffe, D. B. *et al.* Whole-genome sequence assembly for mammalian genomes: Arachne 2. *Genome Res.* **13,** 91–96 (2003).
13. Lodhi, M. A., Daly, M. J., Ye, G. N., Weeden, N. F. & Reisch, B. I. A molecular marker based linkage map of *Vitis*. *Genome* **38,** 786–794 (1995).
14. Doligez, A. *et al.* An integrated SSR map of grapevine based on five mapping populations. *Theor. Appl. Genet.* **113,** 369–382 (2006).
15. Baur, J. A. *et al.* Resveratrol improves health and survival of mice on a high-calorie diet. *Nature* **444,** 337–342 (2006).
16. Baur, J. A. & Sinclair, D. A. Therapeutic potential of resveratrol: the *in vivo* evidence. *Nature Rev. Drug Discov.* **5,** 493–506 (2006).
17. Mateo, J. J. & Jimenez, M. Monoterpenes in grape juice and wines. *J. Chromatogr. A* **881,** 557–567 (2000).
18. Aubourg, S., Lecharny, A. & Bohlmann, J. Genomic analysis of the terpenoid synthase (AtTPS) gene family of *Arabidopsis thaliana*. *Mol. Genet. Genomics* **267,** 730–745 (2002).
19. Tholl, D. *et al.* Formation of monoterpenes in *Antirrhinum majus* and *Clarkia breweri* flowers involves heterodimeric geranyl diphosphate synthases. *Plant Cell* **16,** 977–992 (2004).
20. Adams, K. L. & Wendel, J. F. Polyploidy and genome evolution in plants. *Curr. Opin. Plant Biol.* **8,** 135–141 (2005).
21. Simillion, C., Vandepoele, K., Van Montagu, M. C., Zabeau, M. & Van de Peer, Y. The hidden duplication past of *Arabidopsis thaliana*. *Proc. Natl Acad. Sci. USA* **99,** 13627–13632 (2002).
22. Bowers, J. E., Chapman, B. A., Rong, J. & Paterson, A. H. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* **422,** 433–438 (2003).
23. Vision, T. J., Brown, D. G. & Tanksley, S. D. The origins of genomic duplications in *Arabidopsis*. *Science* **290,** 2114–2117 (2000).
24. Blanc, G., Hokamp, K. & Wolfe, K. H. A recent polyploidy superimposed on older large-scale duplications in the *Arabidopsis* genome. *Genome Res.* **13,** 137–144 (2003).
25. Wolfe, K. H., Gouy, M., Yang, Y. W., Sharp, P. M. & Li, W. H. Date of the monocot–dicot divergence estimated from chloroplast DNA sequence data. *Proc. Natl Acad. Sci. USA* **86,** 6201–6205 (1989).
26. Crane, P. R., Friis, E. M. & Pedersen, K. R. The origin and early diversification of angiosperms. *Nature* **374,** 27–33 (1995).
27. Eshed, Y. & Zamir, D. An introgression line population of *Lycopersicon pennellii* in the cultivated tomato enables the identification and fine mapping of yield-associated QTL. *Genetics* **141,** 1147–1162 (1995).
28. Howe, K. L., Chothia, T. & Durbin, R. GAZE: a generic framework for the integration of gene-prediction data by dynamic programming. *Genome Res.* **12,** 1418–1427 (2002).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements The sequencing of the grapevine genome was launched and carried out after a scientific cooperation agreement between the Ministry of Agriculture in France and the Ministry of Agriculture in Italy, involving l'Institut National de la Recherche Agronomique (INRA), Consiglio per la Ricerca e Sperimentazione in Agricoltura (CRA) and Friuli Venezia Giulia Region. This work was financially supported by Consortium National de Recherche en Génomique, Agence Nationale de la Recherche, INRA, and by MiPAF (VIGNA-CRA), Friuli Innovazione, Università di Udine, Federazione BCC, Fondazione CRUP, Fondazione Carigo, Fondazione CRT, Vivai Cooperativi Rauscedo, Eurotech, Livio Felluga, Marco Felluga, Venica e Venica, Le Vigne di Zamò (IGA). We thank S. Cure for correcting the manuscript; F. Câmara and R. Guigo for the calibration of the GeneID gene prediction software, and the Centre Informatique National de l'Enseignement Supérieur for computing resources.

Author Information The final assembly and annotation are deposited in the EMBL/Genbank/DDBJ databases under accession numbers CU459218–CU462737 (for all scaffolds) and CU462738–CU462772 (for chromosome reconstitutions and unanchored scaffolds). An annotation browser and further information on the project are available from http://www.genoscope.cns.fr/vitis, http://www.vitisgenome.it/ and http://www.appliedgenomics.org/. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to P.W. (pwincker@genoscope.cns.fr).

The French-Italian Public Consortium for Grapevine Genome Characterization
Olivier Jaillon[1]*, Jean-Marc Aury[1]*, Benjamin Noel[1], Alberto Policriti[2,3], Christian Clepet[4], Alberto Casagrande[2,5], Nathalie Choisne[1,4], Sébastien Aubourg[4], Nicola Vitulo[6,15], Claire Jubin[1], Alessandro Vezzi[6,15], Fabrice Legeai[7], Philippe Hugueney[8], Corinne Dasilva[1], David Horner[9,15], Erica Mica[9,15], Delphine Jublot[4], Julie Poulain[1], Clémence Bruyère[4], Alain Billault[1], Béatrice Segurens[1], Michel Gouyvenoux[1], Edgardo Ugarte[1], Federica Cattonaro[2], Véronique Anthouard[1], Virginie Vico[1], Cristian Del Fabbro[2,3], Michaël Alaux[7], Gabriele Di Gaspero[2,5], Vincent Dumas[8], Nicoletta Felice[2,5], Sophie Paillard[4], Irena Juman[2,5], Marco Moroldo[4], Simone Scalabrin[2,3], Aurélie Canaguier[4], Isabelle Le Clainche[4], Giorgio Malacrida[6,15], Eléonore Durand[7], Graziano Pesole[10,11,15], Valérie Laucou[12], Philippe Chatelet[13], Didier Merdinoglu[8], Massimo Delledonne[14,15], Mario Pezzotti[15,16], Alain Lecharny[4], Claude Scarpelli[1], François Artiguenave[1], M. Enrico Pè[9,15], Giorgio Valle[6,15], Michele Morgante[2,5], Michel Caboche[4], Anne-Françoise Adam-Blondon[4], Jean Weissenbach[1], Francis Quétier[1] & Patrick Wincker[1]
*These authors contributed equally to this work.

Affiliations for participants: [1]Genoscope (CEA) and UMR 8030 CNRS-Genoscope-Université d'Evry, 2 rue Gaston Crémieux, BP5706, 91057 Evry, France. [2]Istituto di Genomica Applicata, Parco Scientifico e Tecnologico di Udine, Via Linussio 51, 33100 Udine, Italy. [3]Dipartimento di Matematica ed Informatica, Università degli Studi di Udine, via delle Scienze 208, 33100 Udine, Italy. [4]URGV, UMR INRA 1165, CNRS-Université d'Evry Genomique Végétale, 2 rue Gaston Crémieux, BP5708, 91057 Evry cedex, France. [5]Dipartimento di Scienze Agrarie ed Ambientali, Università degli Studi di Udine, via delle Scienze 208, 33100 Udine, Italy. [6]CRIBI, Università degli Studi di Padova, viale G. Colombo 3, 35121 Padova, Italy. [7]URGI, UR1164 Génomique Info, 523, Place des Terrasses, 91034 Evry Cedex, France. [8]UMR INRA 1131, Université de Strasbourg, Santé de la Vigne et Qualité du Vin, 28 rue de Herrlisheim, BP20507, 68021 Colmar, France. [9]Dipartimento di Scienze Biomolecolari e Biotecnologie, Università degli Studi di Milano, via Celoria 26, 20133 Milano, Italy. [10]Dipartimento di Biochimica e Biologia Molecolare, Università degli Studi di Bari, via Orabona 4, 70125 Bari, Italy. [11]Istituto Tecnologie Biomediche, Consiglio Nazionale delle Ricerche, via Amendola 122/D, 70125 Bari, Italy. [12]UMR INRA 1097, IRD-Montpellier SupAgro-Univ. Montpellier II, Diversité et Adaptation des Plantes Cultivées, 2 Place Pierre Viala, 34060 Montpellier Cedex 1, France. [13]UMR INRA 1098, IRD-Montpellier SupAgro-CIRAD, Développement et Amélioration des Plantes, 2 Place Pierre Viala, 34060 Montpellier Cedex 1, France. [14]Dipartimento Scientifico e Tecnologico, Università degli Studi di Verona Strada Le Grazie 15 – Ca' Vignal, 37134 Verona, Italy. [15]Dipartimento di Scienze, Tecnologie e Mercati della Vite e del Vino, Università degli Studi di Verona, via della Pieve, 70 37029 S. Floriano (VR), Italy. [16]VIGNA-CRA Initiative; Consorzio Interuniversitario Nazionale per la Biologia Molecolare delle Piante, c/o Università degli Studi di Siena, via Banchi di Sotto 55, 53100 Siena, Italy.

METHODS

Genome sequencing. The *V. vinifera* PN40024 genome was sequenced with the use of a whole-genome shotgun strategy. All data were generated by paired-end sequencing of cloned inserts using Sanger technology on ABI3730xl sequencers. Supplementary Table 2 gives the number of reads obtained per library.

Genome assembly and chromosome anchoring. All reads were assembled with Arachne[12]. We obtained 20,784 contigs that were linked into 3,830 supercontigs of more than 2 kb. The contig N_{50} was 64 kb, and the supercontig N_{50} was 1.9 Mb. The total supercontig size was 498 Mb, remarkably close to the expected size of 475 Mb. This indicates that the PN40024 has retained few heterozygous regions. Remaining heterozygosity was assessed by aligning all supercontigs with each other. We first selected the supercontigs more than 30 kb in size that were covered over more than 40% of their length by another supercontig with more than 95% identity. After visual inspection of the alignments, we added to this list the supercontigs more than 10 kb in size that aligned at more than 40% of their length with supercontigs identified previously. All potential cases were then inspected visually to discard potential heterozygous regions (aligning relatively homogeneously across their complete length) and retained repeated regions (with more heterogeneous alignments). This treatment identified 11 Mb of potentially allelic supercontigs. We confirmed that in most cases their coverage was about half the average of the homozygous supercontigs. Only one supercontig of each allelic pair was therefore conserved in the final assembly, which consists of 3,514 supercontigs ($N_{50} = 2$ Mb) containing 19,577 contigs ($N_{50} = 66$ kb), totalling 487 Mb. If the haploid genome size of 475 Mb is considered correct, then our final assembly contains only about 12 Mb of remaining heterozygosity, or 2.6%.

A set of 30,151 bacterial artificial chromosome (BAC) fingerprints of the BAC clones of a Cabernet–Sauvignon library[29] were assembled into 1,763 contigs with FPC[30], v. 8. In parallel, 1,981 markers were anchored on a subset of BAC clones[31], among which 388 markers mapped onto the genetic map, and 77,237 BAC end sequences were obtained[31]. Blat[32] alignments (90% identity on 80% of the length, fewer than five hits) were performed with BAC end sequences on the 3,830 supercontigs of sequences with lengths over 2 kb. The results were then filtered with homemade Perl scripts to keep only the occurrences in which two paired ends were matching at a distance of less than 300 kb and with a consistent orientation. Two supercontigs were considered linked to each other if two BAC links could be found or one BAC link and a BAC contig link. A total number of 111 ultracontigs were constructed with this procedure.

Genome annotation. Several resources were used to build *V. vinifera* gene models automatically with GAZE[28]. We used predictions of repetitive regions by repeatscout[33], conserved coding regions predicted by the exofish method[34,35], genewise[36] alignments of proteins from Uniprot[37], Geneid[38] and Snap[39] *ab initio* gene predictions, and alignments of several cDNA resources (Supplementary Information).

A weight was assigned to each resource to further reflect its reliability and accuracy in predicting gene models. This weight acts as a multiplier for the score of each information source, before being processed by GAZE. When applied to the entire assembled sequence, GAZE predicted 30,434 gene models.

Paralogous and orthologous gene sets. We identified orthologous genes in six pairs of genomes from four species: *A. thaliana*, *O. sativa*, *P. trichocarpa* and *V. vinifera*. Each pair of predicted gene sets was aligned with the Smith–Waterman algorithm, and alignments with a score higher than 300 (BLOSUM62; gapo = 10, gape = 1) were retained. Two genes, A from genome GA and B from genome GB, were considered orthologues if B was the best match for gene A in GB and A was the best match for B in GA.

For each orthologous gene set with *V. vinifera*, clusters of orthologous genes were generated. A single linkage clustering with a euclidean distance was used to group genes. The distances were calculated with the gene index in each chromosome rather than the genomic position. The minimal distance between two orthologous genes was adapted in accordance with the selected genomes. Finally, we retained only clusters that were composed of at least six genes for *Arabidopsis* and *O. sativa*, and eight genes for *P. trichocarpa* (Supplementary Table 10).

To validate the clustering quality we used a method described previously[21]. For each cluster we computed the probability of finding this cluster in the gene homology matrix (Supplementary Table 11). This matrix was constructed from two compared chromosomes with genes numbered according to their position on each chromosome, with no reference to physical distances.

Paralogous genes were computed by comparing all-against-all of *V. vinifera* proteins by using blastp, and alignments with an expected value of less than 0.1 were retained and realigned with the Smith–Waterman algorithm[40]. Two genes A and B were considered paralogues if B was the best match for gene A and A was the best match for B. Moreover, clusters of paralogous genes were constructed in the same fashion as orthologous clusters (Supplementary Table 10).

29. Adam-Blondon, A. F. *et al.* Construction and characterization of BAC libraries from major grapevine cultivars. *Theor. Appl. Genet.* **110,** 1363–1371 (2005).
30. Soderlund, C., Humphray, S., Dunham, A. & French, L. Contigs built with fingerprints, markers, and FPC V4.7. *Genome Res.* **10,** 1772–1787 (2000).
31. Lamoureux, D. *et al.* Anchoring of 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 (2006).
32. Kent, W. J. BLAT—the BLAST-like alignment tool. *Genome Res.* **12,** 656–664 (2002).
33. Price, A. L., Jones, N. C. & Pevzner, P. A. *De novo* identification of repeat families in large genomes. *Bioinformatics* **21** (Suppl. 1), i351–i358 (2005).
34. Roest Crollius, H. *et al.* Estimate of human gene number provided by genome-wide analysis using *Tetraodon nigroviridis* DNA sequence. *Nature Genet.* **25,** 235–238 (2000).
35. Jaillon, O. *et al.* Genome-wide analyses based on comparative genomics. *Cold Spring Harb. Symp. Quant. Biol.* **68,** 275–282 (2003).
36. Birney, E., Clamp, M. & Durbin, R. GeneWise and Genomewise. *Genome Res.* **14,** 988–995 (2004).
37. Bairoch, A. *et al.* The Universal Protein Resource (UniProt). *Nucleic Acids Res.* **33,** D154–D159 (2005).
38. Parra, G., Blanco, E. & Guigo, R. GeneID in *Drosophila. Genome Res.* **10,** 511–515 (2000).
39. Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* **5,** 59 (2004).
40. Smith, T. F. & Waterman, M. S. Identification of common molecular subsequences. *J. Mol. Biol.* **147,** 195–197 (1981).