Supplementary data

References

Supplementary Figures

Supplementary Tables

1 Origin of the PN40024 near homozygous line

Near homozygous lines were derived from Pinot Noir at the INRA station of Colmar[1] by 9 successive selfing steps. Later on, this material was analysed with 36 SSR markers known to be heterozygous in Pinot Noir, based on the work of Hocquigny *et al*[2]. In one line, only one of the SSR markers tested was heterozygous in PN40024 (97% homozygous). The level of homozygosity varied in the other lines between 75% and 94%. However, for some loci (8 in PN40024), the size of the observed allele did not fit with the size of the alleles present in Pinot Noir (data not shown). It was thus suspected that an outcross might have occurred in the former generations. Plants were still available from the 4th to the 8th generations of selfing and could be checked for the same SSR. The non-Pinot alleles were consistently present in all these generations (data not shown), leading to the conclusion that the outcross event occurred between the first, the second or the third generation of selfing. Even though the number of selfing generations the PN40024 was derived from may be lower than anticipated, its level of homozygosity was further inspected and shown to be quite good. Sixty-six additional SSR markers were genotyped: seven out of 102 loci were heterozygous in PN40024.

A paternity search was done by comparing the PN40024 genotype at 20 SSR markers with the genotype of 2,234 previously scored accessions to the germ plasm collection of Vassal (Laucou, V. unpublished results). PN40024 was homozygous for the 20 SSR and presented alleles that were not present in Pinot Noir for 6 markers out of the 20 (Supplementary Table S1). Twelve accessions out of 2,234 could be the donors of these 6 alleles. Eleven of these accessions were discarded as possible parents since they were not present in Colmar when the crosses were made (some of them are recent introductions and were not present in France at all) : the remaining possibility was Helfensteiner. Helfensteiner was obtained in Germany in the early 20th century from a cross between Pinot Noir and Frankenthal. It should be noted that PN40024 could also be a selfing of Helfensteiner as it shares its alleles at all SSR loci with Helfensteiner.

High molecular weight DNA was prepared from 5 g of PN40024 young leaves using the procedure described in Adam-Blondon *et al*[3] and was used for the construction of the BAC library. A second DNA extraction was performed from the same quantity of material, following the same protocol, except that the nuclei were purified using several cycles of differential centrifugation and one ultracentrifugation purification through a 2M sucrose gradient. All steps were performed in H buffer containing 0.5% (v/v) Triton X-100. DNA was purified by Bet-CsCl ultracentrifugation. This preparation was used for the development of the plasmid and fosmid libraries (Supplementary Table S2).

2 Genomic sequencing

The *Vitis vinifera* PN40024 genome was sequenced using a Whole Genome Shotgun strategy. All data were generated by paired-end sequencing of cloned inserts using Sanger technology on ABI3730xl sequencers. Supplementary Table S2 gives the number of reads obtained per library

3 cDNA sequencing

Full-length-enriched cDNA libraries have been constructed from *Vitis vinifera* leaves, flower buds, and a cell line under various stress conditions. For assessing the quality of these libraries, 1,920 clones were sequenced on both ends, producing 1,785 useful reads on the 5'-end and 1,859 on the 3'-end, plus 54 reads corresponding to short, poor-quality or no-insert sequences. A third internal read was performed on a hundred of the biggest cDNAs; in total 1,494 full inserts plus 262 partial sequences were characterized, 37% of the full-inserts are 1.2 to 2.4 kb long and 49% are between 0.9 and 1.2 kb (Supplementary Table S4a). The 1,785 5'-reads assemble with a mean redundancy of ~1.9 clones/ gene (Supplementary Table S5). Blast analysis against *Arabidopsis* shows significant matches for 1,672 5'-reads which overall correspond to 782 different proteins. From a preliminary analysis, 81.6% of these 1,672 cDNAs are long enough to encompass the beginning of the homologous coding sequence and therefore are likely to contain a complete ORF. In total, 5'-end sequences were sequenced on 48,239 clones from the four libraries corresponding to 5,038 different loci (Supplementary Table S4b).

Material and Methods

Four full-length cDNA libraries were constructed from various *Vitis vinifera* tissue pools. For library A, Cabernet Sauvignon CS2 cells were pooled from cell cultures produced under normal conditions or by applying one of the following 24-hour stress strategies: "anaerobic stress" (N_2 atmosphere), "heat stress" (incubation at 31°C), "cold stress" (3h at 5°C then 19h at 17°C), "osmotic stress" (polyethylene glycol 6000, 202 g/l), "salt stress" (NaCl, 0.1 *M*), "antibiotic stress" (hygromycine, 5mg/l). Pinot noir (PN162) leaves and petioles were collected for library B, Pinot Noir (PN177) flower buds at various developmental stages for library C and Pinot Noir PN40024 leaves and petioles for library D.

Total RNAs were extracted according to the method of Chang[4] followed by a purification on an Rneasy spin column (Qiagen) according to the manufacturer's recommendations. Poly $(A)^+$ RNAs were extracted with a poly AT tract mRNA isolation system from Promega.

Full-length cDNAs were prepared from 5 µg poly$(A)^+$ RNA as described previously[5], by replacing the original Gateway adapters with Sfi1 DNA oligos P1: ATCCAGGGCCAAATCGGCCT, P2: NNNNNAGGCCGATTTG and P3: TTGTGGCCCTTATGGCCTTTTTTTTTTTTTTTTTTTTTTVN (purchased at Sigma, N stands for dA, dG, dC or dT; V: dA, dG or dC; a: 3'NH2). Double-strand cDNAs above 1kb were

fractionated on agarose gel, *Sfi*1-digested (New England Biolabs) and ligated in the corresponding *Sfi*1 sites of a plasmid derived from the Promega pGEMT-easy. Electroporation of *E. coli* DH10B T1^r strain (Invitrogen) generated >10^7 transformants per µg of ligated DNA and a 10^{-4} vector background.

4 Assembly and chromosome anchoring

4.1 Assembly

All reads were assembled with Arachne[6]. We obtained 20,784 contigs that were linked into 3,830 supercontigs of more than 2 kb. The contig N50 was 64 kb, and the supercontig N50 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 against each other. We first selected the supercontigs of more than 30 kb in size that are covered at more than 40% of their length by another supercontig with more than 95% identity. After visual inspection of the alignments, we add to this list the supercontigs of more than 10 kb in size that aligned at more than 40 % of their length to supercontigs identified previously. All potential cases were then visually inspected 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. These "allelic" supercontigs were discarded from the final assembly, that consists of 3,514 supercontigs (N50=2 Mb) containing 19,577 contigs (N50=66 kb), totalling 487 Mb. If the haploid genome size of 475 Mb is considered the correct value, then our final assembly contains only about 12 Mb of remaining heterozygosity, or 2.6%.

4.2 Chromosome anchoring

The anchorage of the sequence supercontigs along the grapevine genome was performed in two steps:

- when possible the supercontigs were joined together into ultracontigs using paired BAC end sequences (BES) from Cabernet-Sauvignon and BAC contigs from the same BACs from our Cabernet-Sauvignon physical map
- the ultracontigs and remaining supercontigs were then aligned along a genetic map of the *Vitis vinifera* genome. All the results were stored in a CMap database[7] for graphical display[8]

4.2.1 Construction of the ultracontigs

A set of 30,151 BAC fingerprints of the BAC clones of a Cabernet-Sauvignon library[3] were assembled into 1,763 contigs using FPC[9] v8. In parallel, 1,981 markers have been anchored on a subset of BAC clones[10], among which 388 markers mapped on the genetic map and 77,237 BAC end sequences were obtained[10]. Blat[11] alignments (90% of identity on 80% of the length, less than 5 hits) were performed with the BES on the 3,830 supercontigs of sequences, with lengths over 2kb. The results were then filtered using homemade Perl scripts to keep only the occurrences in which two paired ends were matching at a distance inferior to 300kb 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 could be constructed using this procedure.

4.2.2 Ordering and orientating ultracontigs and supercontigs along the *Vitis vinifera* genetic map

The map published by Doligez et al[12] was used as a reference map (all information about the map and its markers is accessible at URGI[13]). Blat[11] (90% of identity on 80% of the length, less than 5 hits) and e-PCR[14] (with running parameters W = 4, N = 2, M = 250 and a product default size of 400 bp) were performed for 409 monolocus genetic markers on the supercontig sequences. A total of 401 of these markers could be anchored on the genome sequence. For 8 markers no hit was found on the sequence supercontigs; however, for 6 of them we had access to the primer sequences and they were thus tested only by e-PCR. All the results were manually inspected using CMap[7] resulting in 142 supercontigs (120 supercontigs arranged into 37 ultracontigs as described above and 22 single supercontigs) anchored and oriented representing a path with a total length of 303,085,820 bp (62% of the genome size) and 49 supercontigs (20 supercontigs arranged into 8 ultracontigs and 29 single supercontigs) anchored but not oriented representing a total length of 39,539,237 bp (which have been placed in random linkage groups). The non-anchored ultracontigs were not further considered. The Supplementary Table S3 and Supplementary Fig. S1 describe their distribution along the grapevine chromosomes. The N50 of the orientated supercontigs was high, ranging from 1.3 (linkage group 2) up to 12.7 megabases (linkage group 18), showing the high quality of the assembly.

5 Genome Annotation

5.1 Construction of the training set

Non-redundant *Vitis* full-length cDNAs and EST contigs from the TIGR Gene Index were aligned against the genomic sequences. The intron-exon structures obtained have been carefully annotated to check each splicing site, translation initiation codon choice and CDS integrity. A

clean set of 301 complete genes was obtained and used to train gene prediction algorithms and optimize their parameters.

5.2 Repeat Masking

Most of the genome comparisons were performed with repeat masked sequences. For this purpose, we searched and masked sequentially several kinds of repeats:

- known repeats and transposons available in Repbase with the Repeat masker program[15]
- tandem repeats with the TRF program[16]
- *ab initio* detection : RepeatScout[17]

5.3 Identification of repetitive and transposable elements

We analysed the repetitive sequence composition of the grape genome, for which very little information is available. Microsatellites (also termed simple sequence repeats) were identified using a modified version of Sputnik[18] (Supplementary Table S6). We used a total of 600,000 sequences that were not assembled by Arachne (subdivided in three separate sets, two consisting of 100,000 sequences and one of 400,000 sequences) to reconstruct the ancestral sequences of repetitive and transposable elements by means of the ReAS software[19]. We took all the consensus sequences produced by ReAS on the three sets and created a library of repeats for genome annotation that were matched to the sequence assembly using RepeatMasker[15]. In order to better characterize the autonomous transposable element component we assembled a curated set of plant transposable element encoded proteins (including Class I, Class II and Helitrons) derived from the TREP database[20] and from GenBank. BlastX searches of the assembled sequence against this set of proteins were performed to identify regions of homology to known TEs. We then performed manual annotation of transposable elements in approximately 4 Mbp of assembled sequence using a combination of approaches (ReAS annotation, BlastX results, dot plot analysis to detect direct and inverted repeats) to identify a set of 79 putative complete transposons of different classes that were then searched on the genome assembly using RepeatMasker. The combination of all three approaches was used to estimate the total fraction of the genome corresponding to repetitive/transposable elements. We also used all three approaches previously described to examine the distribution of repeats and transposable elements in introns, using experimentally verified introns only (derived from gene predictions obtained from cDNA sequences produced in this project, see above), in order not to be influenced by possible inaccuracies in intron-exon boundary predictions.

5.4 Exofish comparisons

Exofish[21] comparisons were performed at the CINES (Centre Informatique National de l'Enseignement Supérieur), with the Biofacet software package from Gene-IT[22]. When ecores (Evolutionarily COnserved REgions) were contiguous in the two genomes, they were included in the same ecotig[23] (contig of ecores). Exofish comparisons were performed between *Vitis vinifera* and three other plant genomes: *Arabidopsis thaliana*, *Oryza sativa* and *Populus trichocarpa*. HSPs were filtered according to their length and percent identity.

5.5 Genewise

The Uniprot[24] database was used to detect conserved genes between *Vitis vinifera* and other species. As Genewise[25] is time greedy, the Uniprot database was first aligned with the *Vitis vinifera* genome assembly using Blat[11]. Each significant match was chosen for a Genewise alignment.

5.6 Geneid and SNAP

Geneid[26] and SNAP[27] *ab inito* gene prediction software were trained on 301 *Vitis vinifera* genes from the training set.

5.7 *Vitis vinifera* cDNAs

A two-step strategy was used to align the *Vitis vinifera* cDNA clones on the genomic reference sequence[28,29]. Preliminary transcript models were created based on the alignments of the 5' and 3' repeat-masked EST sequence reads derived from the cDNA clones and the *Vitis vinifera* genome assembly. The repeats taken into account by the masking procedure were limited to microsatellites. The HSPs obtained by the BLAST[30] comparisons were combined in a coherent manner, consistent with their position on the reference genomic sequence. In this way, one or several models were built for each transcript, composed of one or several tentative exons based on the alignment with the genome sequence. The model with the highest total score defined by the sum of the scores of each HSP (total score = 800) was selected as the preliminary transcript model that underwent further analysis. cDNA clones with discrepant alignments of their 5' and 3' sequences on the genome were considered to be putative chimeras and were excluded from the analysis.

The unmasked regions of such preliminary transcript models were extended by 5 kb of genomic sequence on each end, and realigned with the cDNA clones using the Est2genome[31]. This procedure defined transcript models with a high fraction of *bona fide* intron-exon boundaries.

These transcript models were fused in gene models by a single linkage clustering approach, in which transcript models from the same genomic region and same strand sharing at least 100 bp are merged in a single model.

5.8 Dicotyledon ESTs

A collection of 2,181,790 public ESTs (from the Eudicotyledon clade) was first aligned with the *Vitis vinifera* genome assembly using Blat[11]. This database was composed of public mRNAs downloaded from the NCBI[32] and clusters of ESTs from the TIGR Plant Transcript Assemblies database[33]. To refine Blat alignment, we used Est2genome[31]. Each significant match was chosen for an alignment with Est2genome. Blat alignments were made using default parameters between translated genomic and translated ESTs.

5.9 Integration of resources using GAZE

All the resources described here were used to automatically build *Vitis vinifera* gene models using GAZE[34]. Individual predictions from each of the programs (Geneid, SNAP, Exofish, Genewise and Est2genome) were broken down into segments (coding, intron, intergenic) and signals (start codon, stop codon, splice acceptor, splice donor, transcript start, transcript stop).

Exons predicted by *ab initio* software, Exofish, Genewise, and Est2genome were used as coding segments. Introns predicted by Genewise and Est2genome were used as intron segments. Intergenic segments created from the span of each mRNA, with a negative score (coercing GAZE not to split genes). Predicted repeats were used as intron and intergenic segments, and non-coding RNAs as intergenic segments, to avoid prediction of genes coding proteins in such regions.

The whole genome was scanned to find signals (splice sites, start and stop codons), and two signals, transcript start and stop, were extracted from the ends of mRNAs.

Each segment extracted from a software output which predicts exon boundaries (like Genewise, Est2genome or *ab initio* predictors), was used by GAZE only if GAZE chose the same boundaries. Each segment or signal from a given program was given a value reflecting our confidence in the data, and these values were used as scores for the arcs of the GAZE automaton. All signals were given a fixed score, but segment scores were context sensitive: coding segment scores were linked to the percentage identity (%ID) of the alignment; intronic segment scores were linked to the %ID of the flanking exons. 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 processing by GAZE. When applied to the entire assembled sequence, GAZE predicted 30,434 gene models.

5.10 Non-coding RNA

A complete search of the assembly was performed with tRNAscan-SE[35] with relaxed settings applied in both tRNAscan and EufindtRNA. A number of 600 tRNA genes including 1 selenocysteine tRNA (as well as 133 potential tRNA pseudogenes) were predicted. The program

srpSCAN[36] yielded 8 high confidence predictions for 7SLRNA genes. Four of these genes were clustered on a single contig while the other 4 copies were distributed in 2 clusters of 2 genes. A number of 257 C/D box SnoRNAs were identified using SnoScan[37]. 5S ribosomal RNA sequences were identified with sequence similarity searches using the available *Vitis vinifera* 5S sequence (AJ972877.1), Contigs with significant hits were examined with INFERNAL[38] using the RFAM00001 model (5SrRNA). 5S rRNA genes were found to be distributed in two principal clusters. The numbers and distributions of these genes are similar to those observed in both *Arabidopsis* and poplar[39].

MicroHarvester[40] was used to search for members of all characterized plant microRNA families[41] (present in release 9.1 of MiRBase[41]) yielding 164 high confidence predictions (Table1). As in other higher plants, the miR169 family appears to be the largest of the currently known micro RNA families (27 genes distributed for the most part in two genomic clusters). Twenty-one families appear to be present in grapevine, *Arabidopsis*, poplar and rice; 1 family (miR403) is present in grapevine (6 members), *Arabidopsis* (1 member) and poplar (3 members) - to the exclusion of rice; 3 families (miR477, miR479, miR482) are present in grapevine and poplar (to the exclusion of *Arabidopsis*), while 4 families (miR828, miR838, miR845, miR858) are found in grapevine and *Arabidopsis* (but not yet characterized in poplar). Strikingly, the miR845 family is apparently greatly expanded in grapevine (9 members) compared to *Arabidopsis* (2 members). Interestingly, we found 5 candidate members of the miR535 family in grape. This family is present in *Phsycomitrella patens* and in rice - suggesting its ancestral nature - however, the homologs identified here are the first such sequences in dicots. Analogously, we found a member of the miR1213 family, previously only identified in *Phsycomitrella*. Finally, the miR395 family is expanded in grapevine (14 members, 13 of which constitute a single positional cluster) compared to *Arabidopsis* (6) and poplar (10). Interestingly, this family is thought to be involved in the regulation of sulfate metabolism through the targeting of messages encoding ATP Sulphurylases and further investigations into the role of this microRNA family in grapevine may thus be of particular agricultural relevance.

6 Identification of orthologous genes

We identified orthologous genes in 6 pairs of genomes from 4 species: *Arabidopsis thaliana*, *Oryza sativa*, *Populus trichocarpa* and *Vitis 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 orthologs if B is the best match for gene A in GB and A is the best match for B in GA.

For each orthologous gene set with *Vitis vinifera*, clusters of orthologous genes have been generated. A single linkage clustering with a euclidian distance was used to group genes. The distances were calculated using 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 only retained clusters that were composed of at least 6 genes for *Arabidopsis* and rice, and 8 genes for poplar (Supplementary Table S10).

To validate the clustering quality, we used the method described by Simillion et al[42]. For each cluster, we computed the probability of finding this cluster in the Gene Homology Matrix (Supplementary Table S11). This matrix was constructed from 2 compared chromosomes with genes numbered according to their position on each chromosome, with no reference to physical distances.

7 Identification of paralogous genes

Initially an all-against-all comparison of *Vitis vinifera* proteins was performed using the Smith-Waterman algorithm and alignments with an e-value lower than 0.1 were retained. Two genes, A and B were considered paralogs if B is the best match for gene A and A was the best match of B. Moreover, clusters of paralogous genes were constructed, in the same fashion as orthologous clusters, section 5 (Supplementary Table S10).

8 Protein domain analysis

InterProScan was run against all *Arabidopsis thaliana*, *Populus trichocarpa*, *Oryza sativa* and *Vitis vinifera* proteins as described earlier[43]. Matches which fulfilled the following criteria were retained :

- match is tagged as “True Positive” by InterProScan (status=T) ;
- match with an e-value less or equal to 10^{-1}.

A total of 3,931 InterPro domains (with IPR number) were found in *Vitis vinifera*, and correspond to 21,649 V*itis vinifera* proteins (Supplementary Table S9).

Targeting peptides, signals and transmembrane segments have been predicted on the grape proteome using an optimized pipeline merging the Predotar, ChloroP, Psort and TMhmm tools. The results indicate that 13%, 3%, 12% and 3% of the grape proteins are localized in endoplasmic reticulum, mitochondria, chloroplast and nucleus respectively. Furthermore, 24% of the predicted proteins have at least one transmembrane hydrophobic domain. All these values are similar to the subcellular localizations predicted for the *Arabidopsis* and rice proteomes.

9 Functional annotation

9.1 Enzyme annotation

Enzyme detection in predicted *Vitis vinifera* proteins was performed with PRIAM[44], using the PRIAM July 2004 ENZYME release. A total of 935 different EC numbers, corresponding to enzyme domains, are associated with 7,593 *Vitis vinifera* proteins. Therefore, about 25% of *Vitis vinifera* proteins contain at least one enzymatic domain.

9.2 Association of metabolic pathways with enzymes and *Vitis vinifera* proteins

From EC numbers, potential metabolic pathways were deduced using the KEGG pathway database[45]. Links between EC numbers and metabolic pathways were obtained from the KEGG website. Using this file and the PRIAM results, the 7,593 *Vitis vinifera* proteins which have an EC number were assigned to 200 pathways.

Following the KEGG pathway hierarchy, pathways from the same family were grouped together. For instance, glycolysis and TCA cycle belong to carbohydrate metabolism. Using this method, the different pathways found in *Vitis vinifera* define 16 pathway families.

References

1. Bronner, A. & Oliveira, J. Création et étude de lignées chez Pinot Noir (Vitis vinifera L.). Journal International de la Vigne et du Vin 25, 133-148 (1991).
2. Hocquigny, S. et al. Diversification within grapevine cultivars goes through chimeric states. Genome 47, 579-589 (2004).
3. Adam-Blondon, A. F. et al. Construction and characterization of BAC libraries from major grapevine cultivars. Theor Appl Genet 110, 1363-71 (2005).
4. Chang, S. A simple and efficient method for isolating RNA from pine trees. Plant Molecular Biology Reporter 11(2), 113-116 (1993).
5. Clepet, C., Le Clainche, I. & Caboche, M. Improved full-length cDNA production based on RNA tagging by T4 DNA ligase. Nucleic Acids Res 32, e6 (2004).
6. Jaffe, D. B. et al. Whole-genome sequence assembly for mammalian genomes: Arachne 2. Genome Res 13, 91-6 (2003).
7. CMap. http://www.gmod.org/cmap/.
8. URGI. http://urgi.versailles.inra.fr/cmap.
9. Soderlund, C., Humphray, S., Dunham, A. & French, L. Contigs built with fingerprints, markers, and FPC V4.7. Genome Res 10, 1772-87 (2000).
10. 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, 344-56 (2006).
11. Kent, W. J. BLAT--the BLAST-like alignment tool. Genome Res 12, 656-64 (2002).
12. Doligez, A. et al. An integrated SSR map of grapevine based on five mapping populations. Theor Appl Genet 113, 369-82 (2006).
13. URGI. http://urgi.versailles.inra.fr/GnpMap/mapping/welcome.do.
14. Schuler, G. D. Sequence mapping by electronic PCR. Genome Res 7, 541-50 (1997).
15. Smit, A., Hubley, R & Green, P. *RepeatMasker Open-3.0* 1996-2004 http://www.repeatmasker.org.
16. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res 27, 573-80 (1999).
17. Price, A. L., Jones, N. C. & Pevzner, P. A. De novo identification of repeat families in large genomes. Bioinformatics 21 Suppl 1, i351-8 (2005).
18. Morgante, M., Hanafey, M. & Powell, W. Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. Nat Genet 30, 194-200 (2002).
19. Li, R. et al. ReAS: Recovery of ancestral sequences for transposable elements from the unassembled reads of a whole genome shotgun. PLoS Comput Biol 1, e43 (2005).
20. TREP. http://wheat.pw.usda.gov/ITMI/Repeats/.
21. Roest Crollius, H. et al. Estimate of human gene number provided by genome-wide analysis using Tetraodon nigroviridis DNA sequence. Nat Genet 25, 235-8 (2000).
22. Gene-It. www.gene-it.com.
23. Jaillon, O. et al. Genome-wide analyses based on comparative genomics. Cold Spring Harb Symp Quant Biol 68, 275-82 (2003).
24. Bairoch, A. et al. The Universal Protein Resource (UniProt). Nucleic Acids Res 33, D154-9 (2005).
25. Birney, E., Clamp, M. & Durbin, R. GeneWise and Genomewise. Genome Res 14, 988-95 (2004).
26. Parra, G., Blanco, E. & Guigo, R. GeneID in Drosophila. Genome Res 10, 511-5 (2000).
27. Korf, I. Gene finding in novel genomes. BMC Bioinformatics 5, 59 (2004).
28. Porcel, B. M. et al. Numerous novel annotations of the human genome sequence supported by a 5'-end-enriched cDNA collection. Genome Res 14, 463-71 (2004).

29. Castelli, V. et al. Whole genome sequence comparisons and "full-length" cDNA sequences: a combined approach to evaluate and improve Arabidopsis genome annotation. Genome Res 14, 406-13 (2004).
30. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. J Mol Biol 215, 403-10 (1990).
31. Mott, R. EST_GENOME: a program to align spliced DNA sequences to unspliced genomic DNA. Comput Appl Biosci 13, 477-8 (1997).
32. NCBI. http://www.ncbi.nlm.nih.gov/.
33. Childs, K. L. et al. The TIGR Plant Transcript Assemblies database. Nucleic Acids Res 35, D846-51 (2007).
34. Howe, K. L., Chothia, T. & Durbin, R. GAZE: a generic framework for the integration of gene-prediction data by dynamic programming. Genome Res 12, 1418-27 (2002).
35. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res 25, 955-64 (1997).
36. Regalia, M., Rosenblad, M. A. & Samuelsson, T. Prediction of signal recognition particle RNA genes. Nucleic Acids Res 30, 3368-77 (2002).
37. Lowe, T. M. & Eddy, S. R. A computational screen for methylation guide snoRNAs in yeast. Science 283, 1168-71 (1999).
38. Griffiths-Jones, S. et al. Rfam: annotating non-coding RNAs in complete genomes. Nucleic Acids Res 33, D121-4 (2005).
39. Tuskan, G. A. et al. The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). Science 313, 1596-604 (2006).
40. Dezulian, T., Remmert, M., Palatnik, J. F., Weigel, D. & Huson, D. H. Identification of plant microRNA homologs. Bioinformatics 22, 359-60 (2006).
41. MicroRNA. http://microrna.sanger.ac.uk/sequences/.
42. Simillion, C., Vandepoele, K., Saeys, Y. & Van de Peer, Y. Building genomic profiles for uncovering segmental homology in the twilight zone. Genome Res 14, 1095-106 (2004).
43. Zdobnov, E. M. & Apweiler, R. InterProScan--an integration platform for the signature-recognition methods in InterPro. Bioinformatics 17, 847-8 (2001).
44. Claudel-Renard, C., Chevalet, C., Faraut, T. & Kahn, D. Enzyme-specific profiles for genome annotation: PRIAM. Nucleic Acids Res 31, 6633-9 (2003).
45. Kanehisa, M., Goto, S., Kawashima, S. & Nakaya, A. The KEGG databases at GenomeNet. Nucleic Acids Res 30, 42-6 (2002).

Supplementary Figures

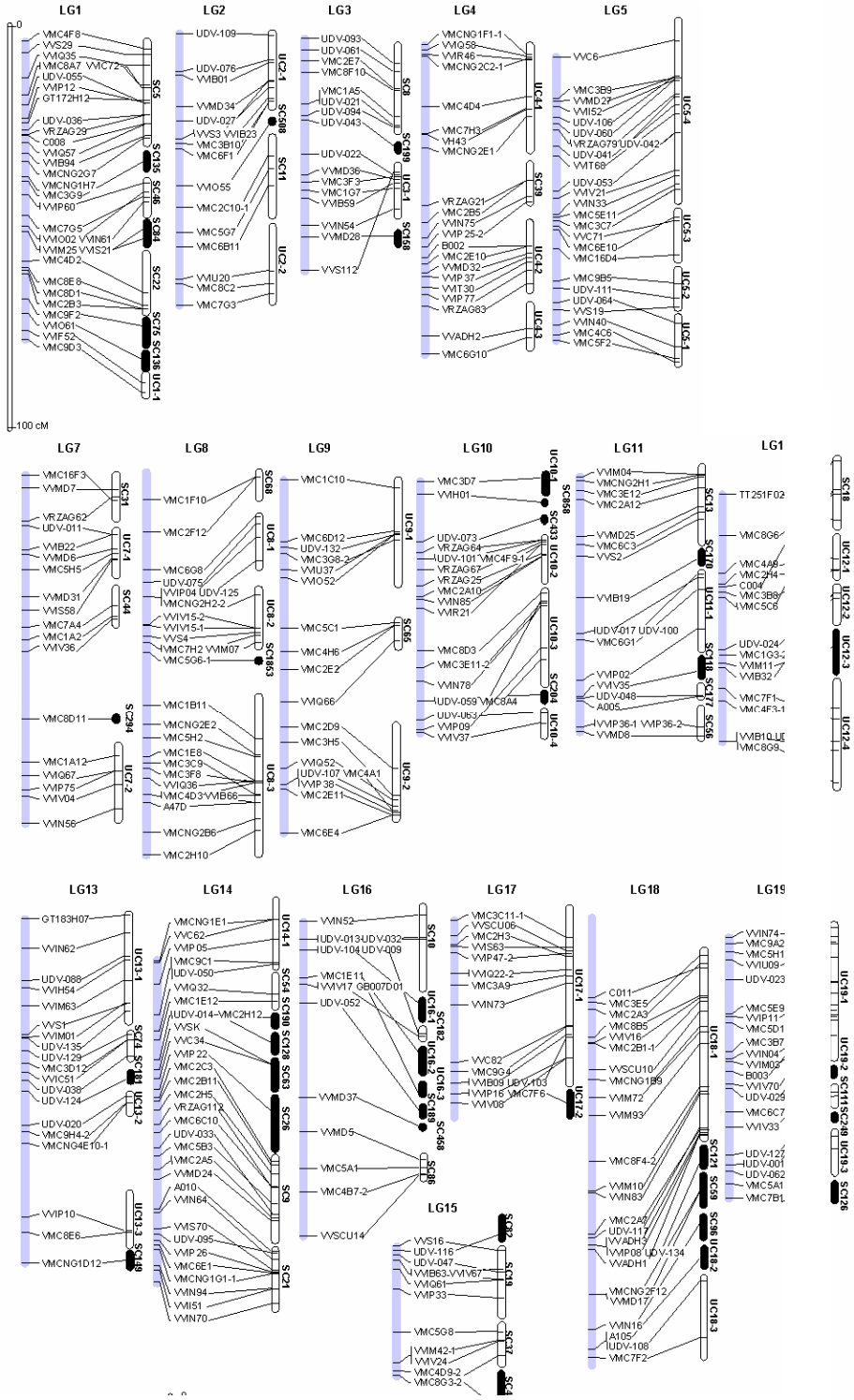

Figure S1. Map of the sequence supercontigs (SC) and ultracontigs (UC) along the linkage groups (LG) of the grapevine genetic map. The linkage groups are represented as grey bars on the left. Only the informative markers are represented. The sequence supercontigs and ultracontigs are represented on the right as white bars (orientated) and black bars (random orientation).

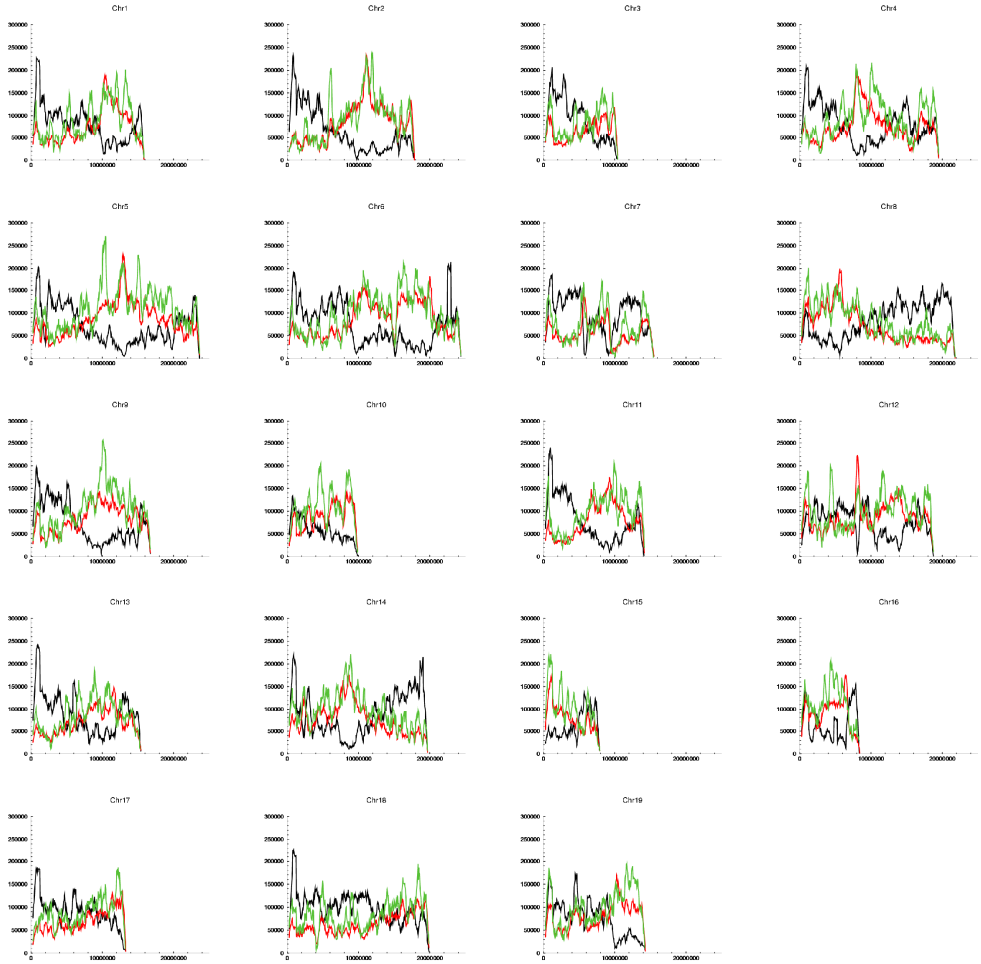

Figure S2. Density profiles on the chromosomes of *Vitis vinifera* of the number of coding bases (black), number of bases in repeat regions (red) and number of bases in transposable elements (green).

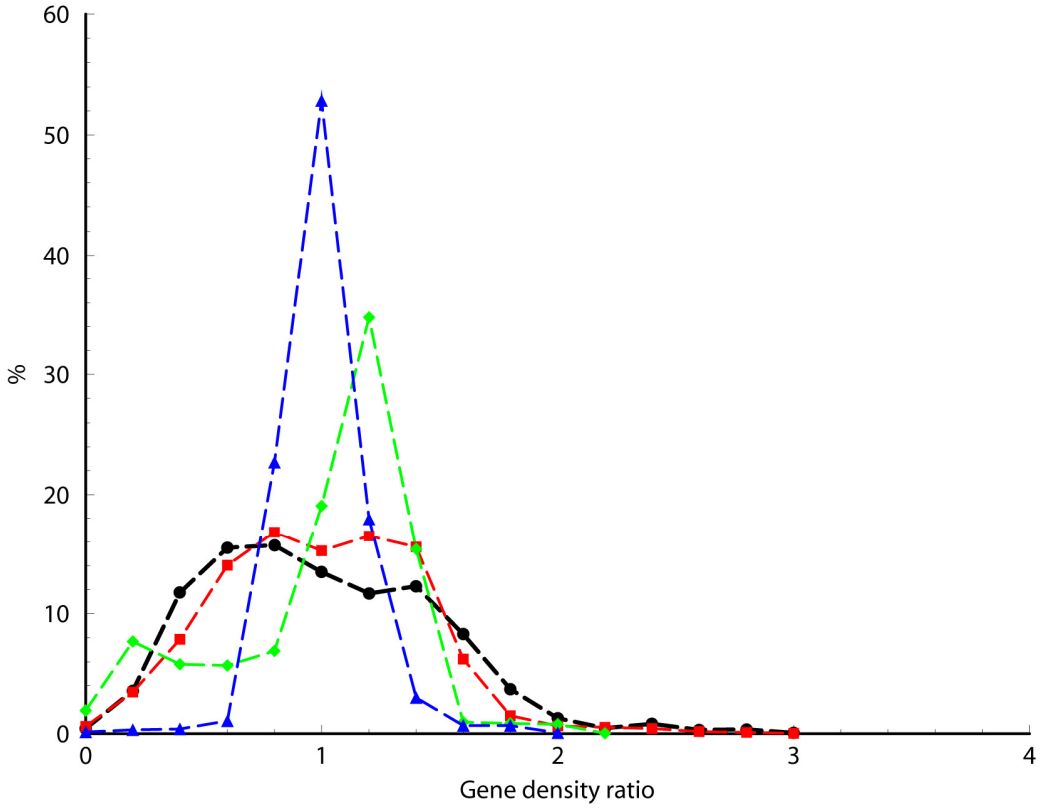

Figure S3. Distribution of the gene density homogeneity in the 4 plants (black : *V. vinifera*, red : *P. trichocarpa*, green : *A. thaliana*, blue : *O. sativa*). Ratio of gene density is measured in the 4 plants as follows. For each sliding window of 500 Kb, we report the ratio of the gene density in this window over the gene density average. On the X axis the gene density ratio classes are reported, on the Y axis the percent of windows falling into each class.

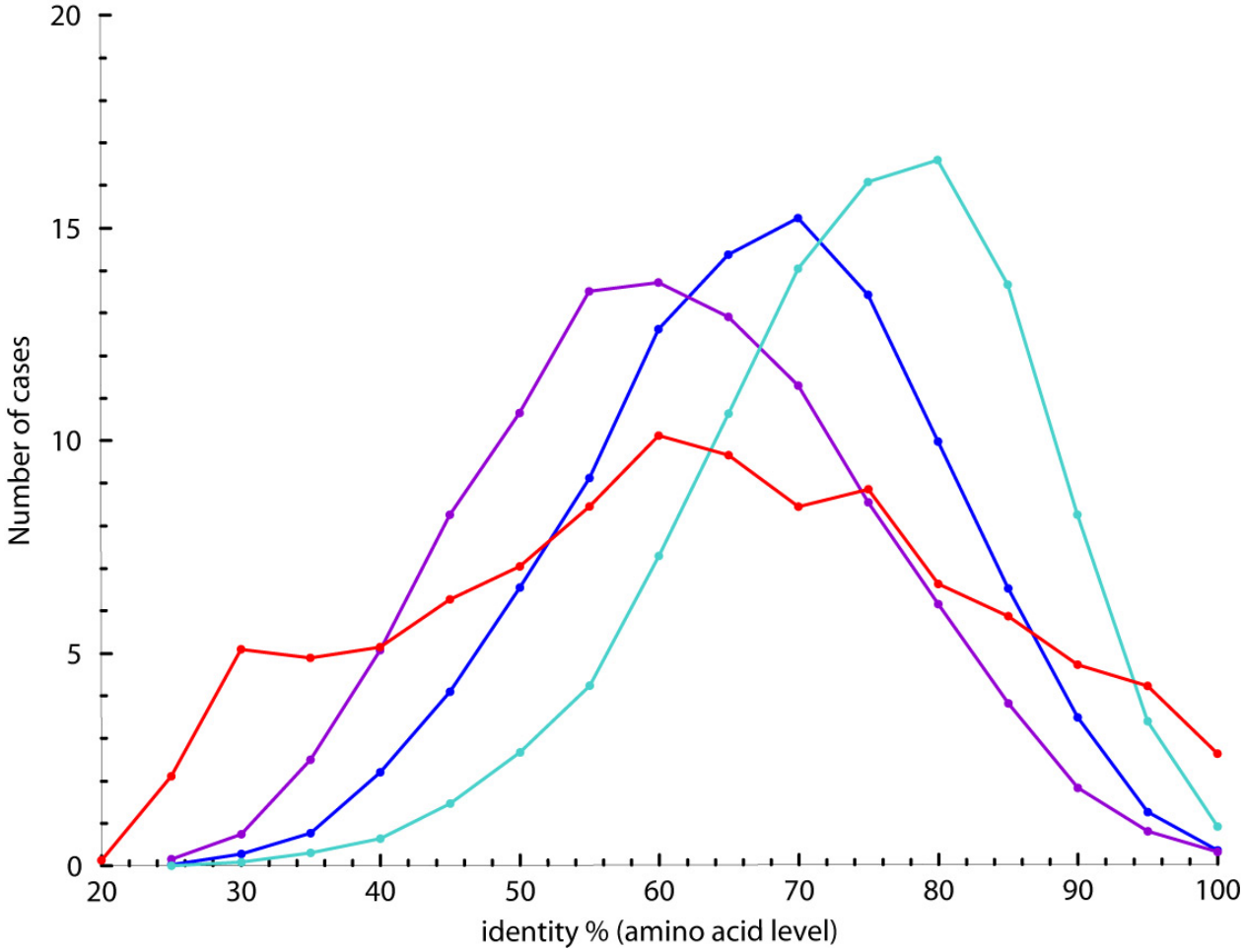

Figure S4. Distribution of the percent identity between pairs of orthologous protein sets (light blue : *Vitis* vs poplar; dark blue : *Vitis* vs *Arabidopsis*; purple : *Vitis* vs rice). Red : distribution of the percent identity between *Vitis* paralogous proteins, excluding paralogs linked on the same chromosome.

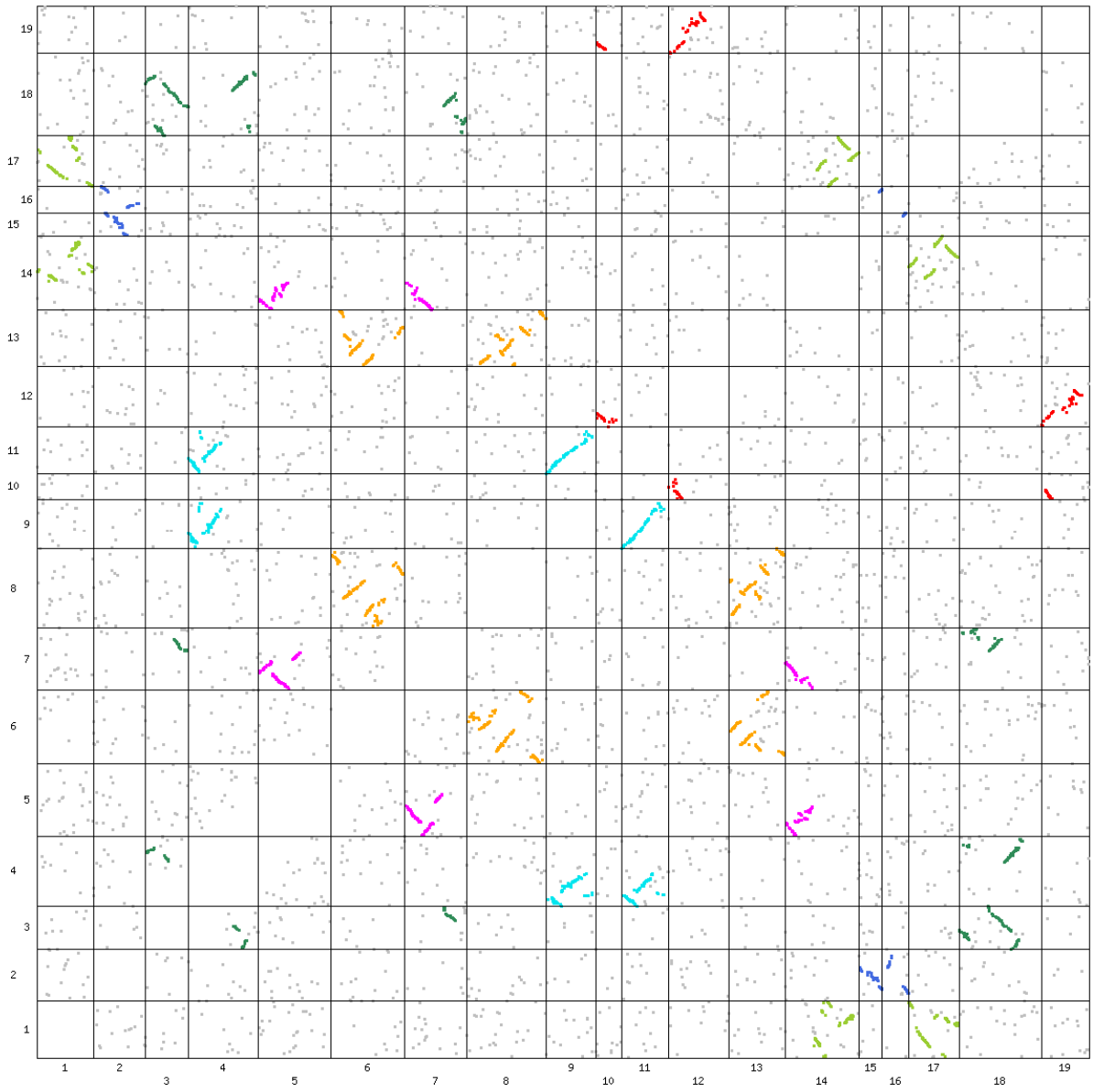

Figure S5. The grape genome originated from a polyploidy event that joined three ancestral genomes. The nineteen chromosomes of grape are represented on both the x and y axis. Dots represent the positions of paralogous pairs of genes. For clarity, intrachromosomal paralogs are not shown. Clusters of paralogs form a succession of dots, that indicate that the gene order of the ancestral genome was locally maintained. These clusters are painted in seven colours. Each colour marks paralogous blocks, that were colinear in the ancestors of the three constituents of the grape genome. Some regions are not painted in triplicate in this grid, either because a whole region is not visible in synteny with two others in the present-day grape genome (too many rearrangements or gene loss), or because one or two syntenic regions lie in supercontigs which are still not anchored.

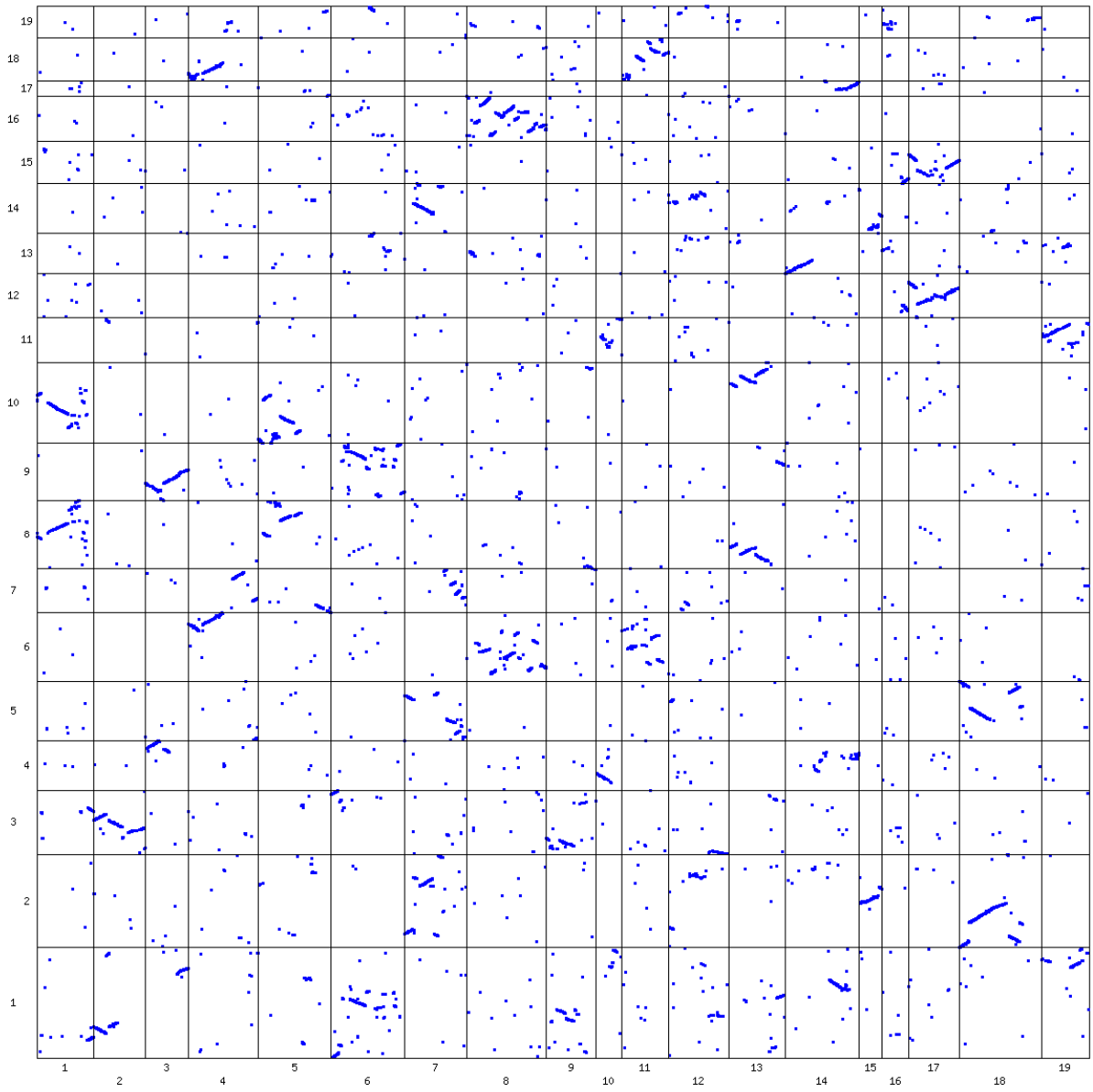

Figure S6. The distribution of 8,604 orthologous genes between *Vitis vinifera* (x axis) and *Populus trichocarpa* (y axis) chromosomes.

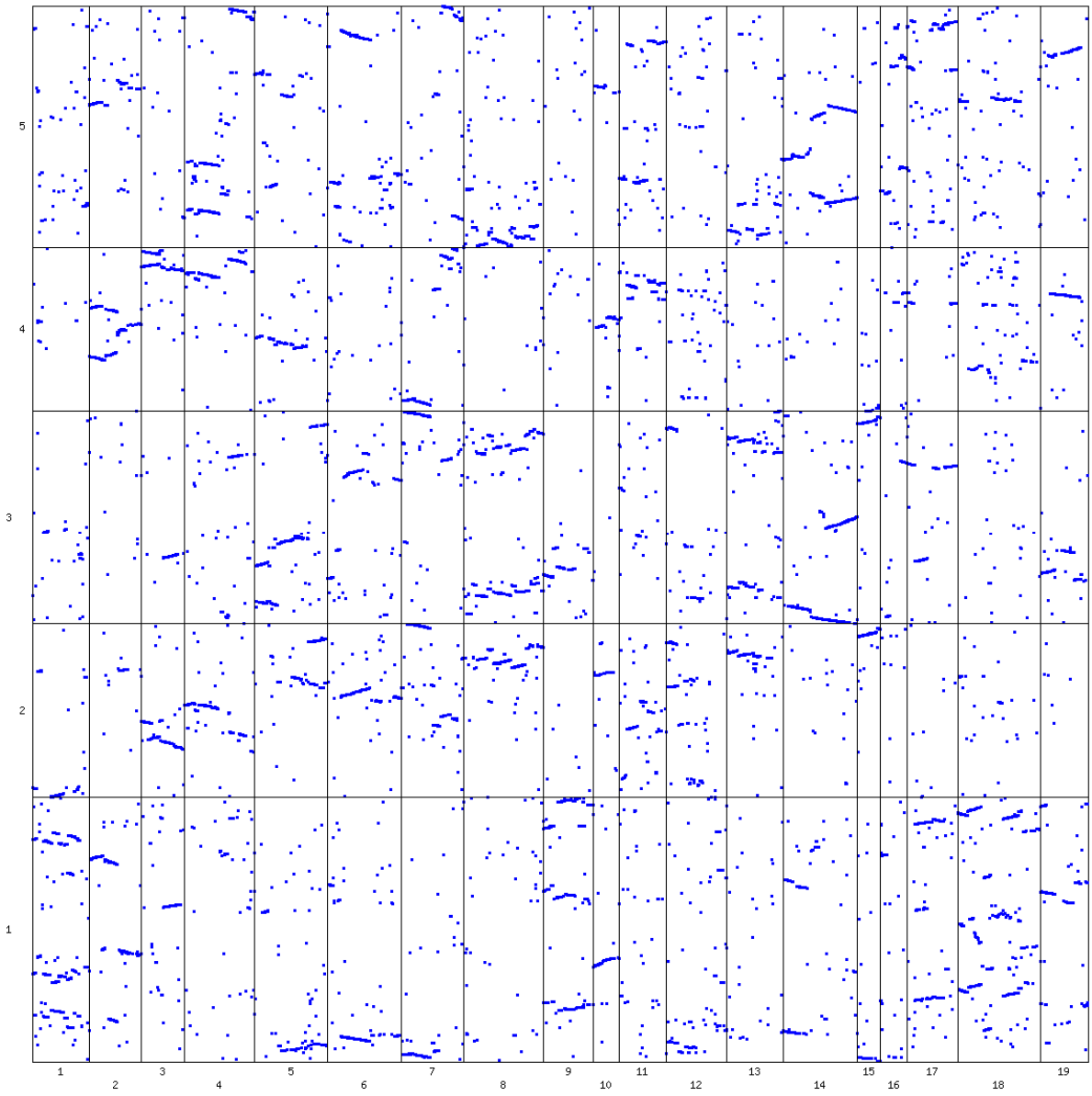

Figure S7. The distribution of 9,225 orthologous genes between *Vitis vinifera* (x axis) and *Arabidopsis thaliana* (y axis) chromosomes.

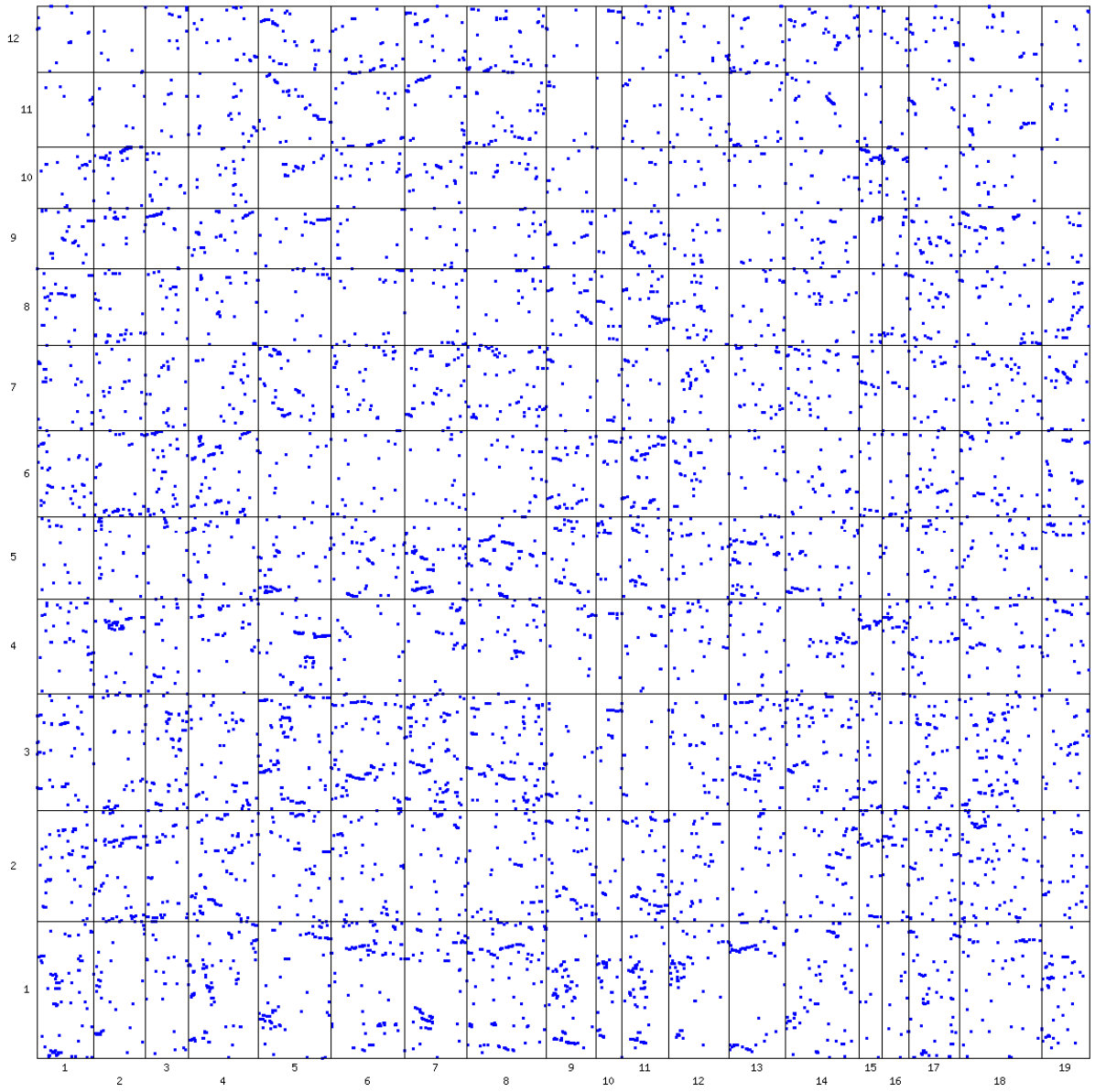

Figure S8. The distribution of 7,952 orthologous genes between *Vitis vinifera* (x axis) and *Oryza sativa* (y axis) chromosomes.

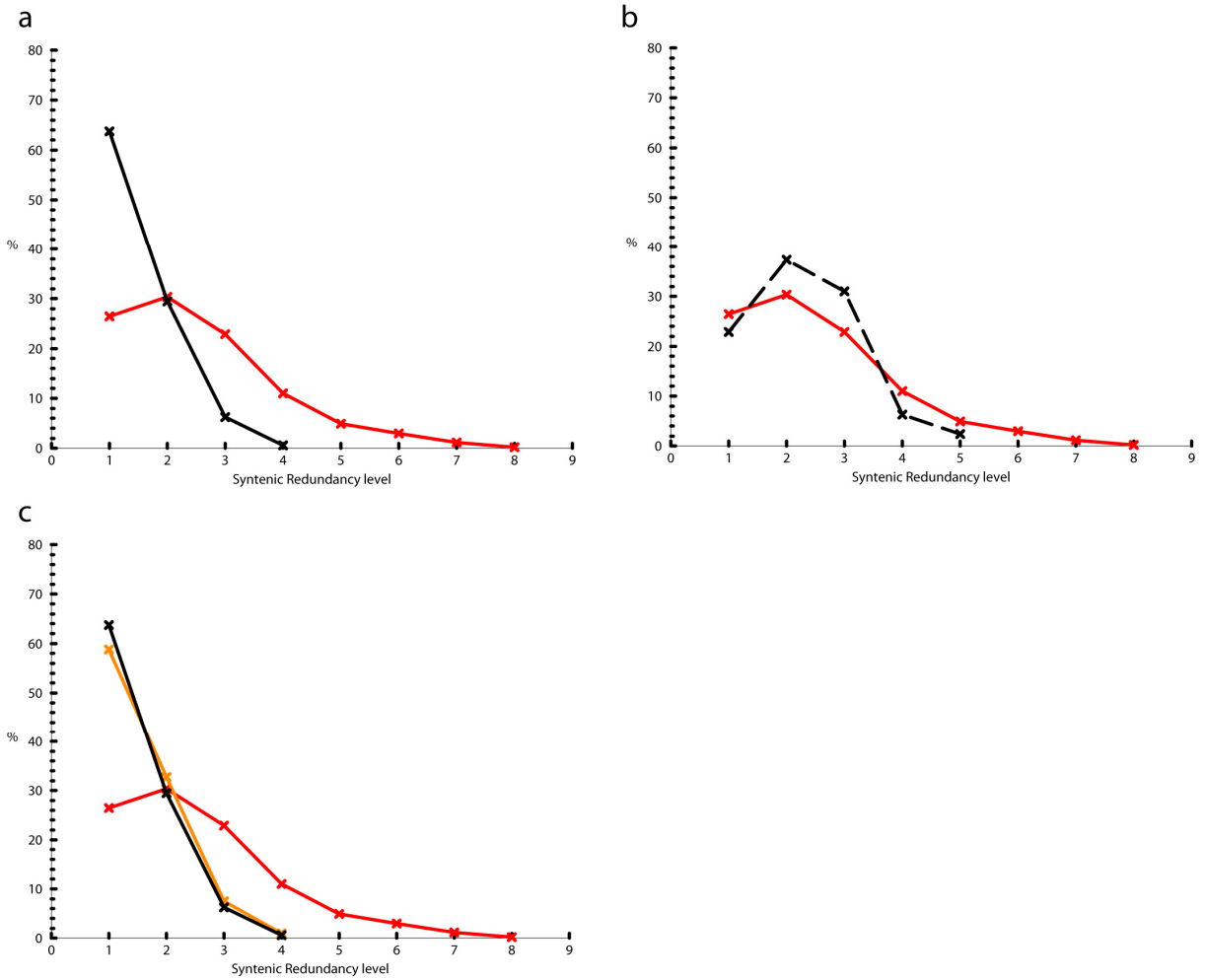

Figure S9. The rice genome shows no evidence of the paleo-hexaploidy content of dicotyledons.

a. Number of grape regions orthologous to one region of *Arabidopsis* (black solid line), and orthologous to one region of rice (red solid line). The black line corresponds in majority to a 1 to 1 situation, compatible to an absence of polyploidization event in grape since the last common ancestor with *Arabidopsis*. The shape of the red line indicates the presence in grape of a polyploidization event that is absent in rice.

b. Number of *Arabidopsis* regions orthologous to one region of grape (black dashed line), and number of grape regions orthologous to one rice region (red solid line). Events of polyploidisation in *Arabidopsis* lineage since the last common ancestor with grape, cause the different shape of the curve compared to the solid black line in a. Here, the shapes of the two curves are similar.

c. The red solid line is the number of grape regions orthologous to one rice region. The paralogous relationships in grape are then eliminated by re-calculating the syntenic redundancy level considering as a single block in grape each doublet or triplet corresponding to a known paralogous region in the paleo-hexaploid (orange curve). This distribution now fits that detected in grape with a genome bearing the triplication (black solid line, comparison *Arabidopsis*-grape), indicating that the shape of the red curve is probably due to the absence of the ancient triplication in the rice genome.

X axis : syntenic redundancy level (number of blocks detected orthologous in one genome with a single block in another genome).

Y axis : percentage of cases.

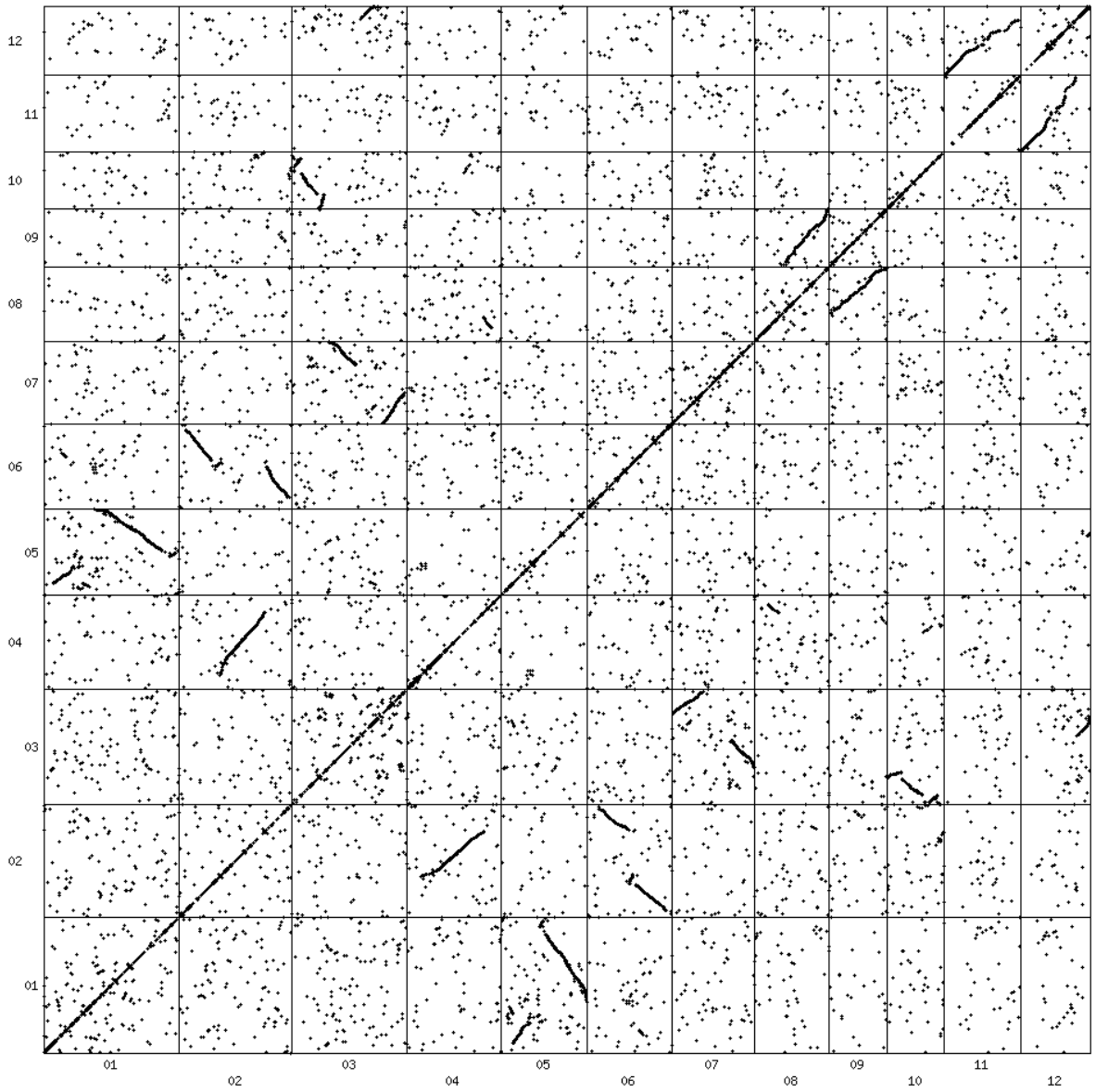

Figure S10. The distribution of 14,613 paralogous genes of the rice *Oryza sativa*. Each column, corresponding to a rice chromosome, can be grouped with at least one other column. The diagonal line displays paralogous links between genes that are close and on the same chromosome (recent segmental duplications).

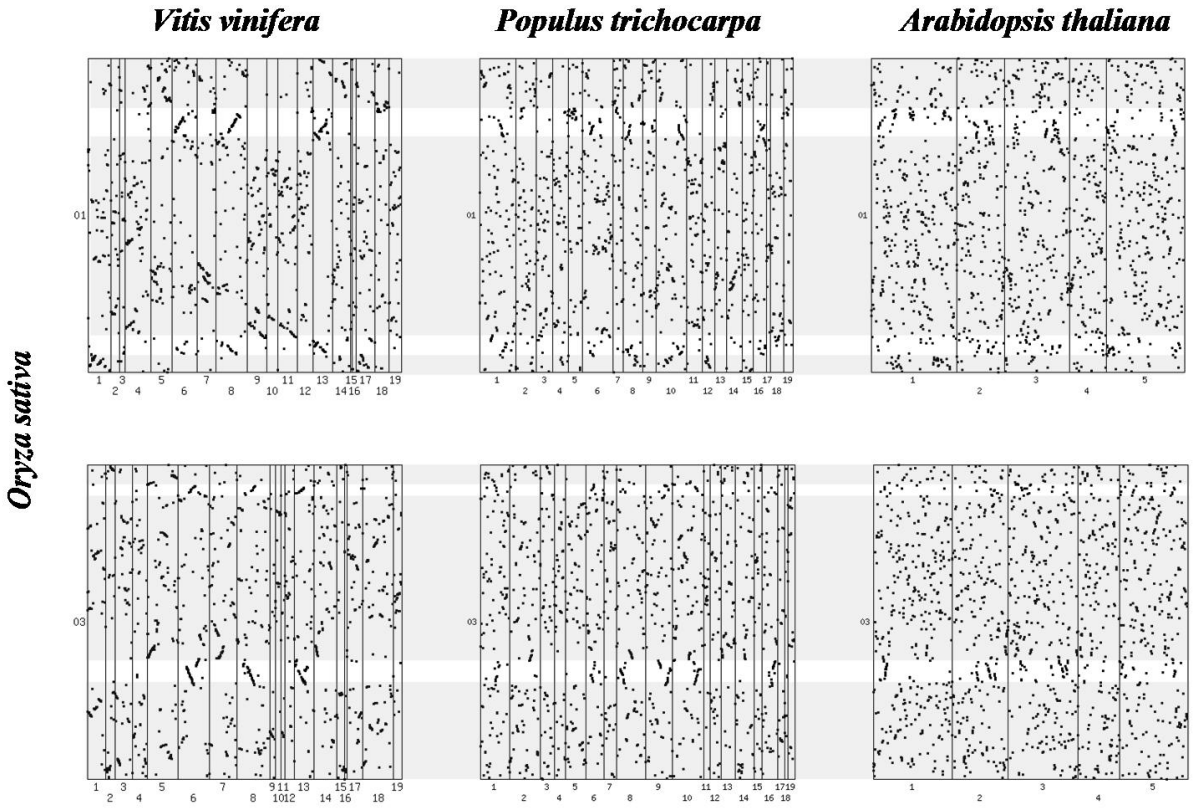

Figure S11. Distribution of paralogous genes of chromosomes 1 and 3 of rice and all the chromosomes of *V. vinifera*, *P. trichocarpa* and *A. thaliana*. White zones highlight highly conserved syntenic regions : each region in rice corresponds to 3 regions in grape, 6 in poplar and more than 8 in *Arabidopsis.*

Supplementary Tables

Table S1. Genotypes at 20 SSR markers of Pinot Noir, PN40024 and its possible parents. Alleles found in PN40024 are in bold. Alleles present in PN40024 but not in Pinot Noir are in grey boxes. The SSR markers have been extensively described in Doligez *et al*[12].

SSR id	**Pinot Noir**	**PN40024**	Jean Macé	La Guitte	Bouquettraube B	**Frankenthal rouge foncé**	Fredericton	Shirazi	Inconnu blanc	**Helfensteiner**	*Vitis vinifera* not identified	*Vitis vinifera* not identified	*Vitis vinifera* not identified	*Vitis vinifera subsp. Silvestris*
VMC1B11	**165**	**165**	**165**	**165**	171	171	171	**165**	**165**	**165**	171	169	**165**	169
	171		173	171	184	171	184	173	173	171	171	182	167	182
VMC4F3-1	171	**181**	164	171	169	171	171	**181**	**181**	171	171	**181**	164	181
	177		**181**	**181**	**181**	**181**	**181**	187	**181**	**181**	**181**	200	**181**	**181**
VVIB01	**288**	**288**	294	294	**288**	294	294	298	290	**288**	294	290	294	290
	294		294	294	294	294	294	298	298	294	294	290	294	290
VVIH54	163	**165**	**165**	**165**	163	**165**	**165**	**165**	**165**	**165**	**165**	139	139	139
	167		167	167	**165**	**165**	**165**	**165**	177	167	**165**	**165**	**165**	**165**
VVIN16	149	**157**	149	155	149	149	149	151	147	149	149	147	149	147
	157		155	**157**	149	155	**157**	151	151	**157**	155	151	151	151
VVIN73	**263**	**263**	**263**	**263**	256	**263**	**263**	**263**	**263**	**263**	**263**	**263**	256	**263**
	265		**263**	**263**	**263**	**263**	**263**	**263**	**263**	**263**	**263**	**263**	**263**	**263**
VVIP31	**178**	**178**	176	**178**	174	**178**	**178**	182	182	**178**	**178**	190	**178**	190
	182		182	182	**178**	190	194	184	184	182	**178**	190	194	190
VVIP60	**315**	**315**	**315**	**315**	**315**	**315**	**315**	**315**	**315**	**315**	**315**	**315**	311	**315**
	317		319	319	319	319	**315**	317	319	317	**315**	319	328	319
VVIQ52	83	**79**	**79**	**79**	**79**	**79**	77	**79**	77	**79**	**79**	**79**	77	**79**
	83		83	83	83	81	**79**	**79**	**79**	83	81	83	**79**	83
VVIV37	149	**167**	165	159	159	155	**167**	155	159	149	155	155	163	155
	159		**167**	**167**	**167**	**167**	**167**	**167**	**167**	**167**	**167**	**167**	**167**	**167**
VVIV67	360	**368**	353	**368**	360	**368**	357	353	343	**368**	**368**	347	353	347
	368		**368**	**368**	**368**	**368**	**368**	353	353	**368**	**368**	355	353	355
VVMD21	**246**	**246**	247	247	247	247	247	247	247	**246**	247	241	241	241
	247		255	247	247	247	247	247	247	247	247	**246**	253	**246**
VVMD24	**212**	**212**	0	210	206	206	206	206	210	206	206	206	206	206
	214		0	**212**	206	210	206	214	214	**212**	210	210	210	210
VVMD25	**238**	**238**	**238**	240	248	240	254	**238**	240	**238**	240	254	254	254
	248		240	248	254	254	254	240	248	254	254	266	254	266
VVMD27	182	**178**	176	**178**	**178**	**178**	176	**178**	176	**178**	**178**	**178**	172	**178**
	186		**178**	186	191	182	**178**	191	**178**	182	**178**	182	**178**	182
VVMD28	216	**235**	0	216	227	**235**	243	**235**	**235**	**235**	**235**	233	233	233
	235		0	243	**235**	243	263	243	257	**235**	**235**	243	234	243
VVMD32	239	**271**	0	251	**271**	251	0	249	**271**	**271**	**271**	251	249	251
	271		0	**271**	**271**	**271**	0	**271**	**271**	**271**	**271**	**271**	261	**271**
VVMD5	**225**	**225**	**225**	**225**	223	234	**225**	232	236	**225**	234	**225**	232	**225**
	236		244	234	236	236	234	236	244	236	234	232	238	232
VVMD7	239	**247**	239	243	243	**247**	239	**247**	239	243	**247**	**247**	**247**	**247**
	243		**247**	**247**	**247**	**247**	**247**	253	**247**	**247**	**247**	249	253	249
VVS2	**135**	**135**	131	149	133	133	133	139	139	133	133	133	139	133
	149		153	153	149	153	149	143	149	**135**	153	145	141	145

Table S2. Sequencing overview.

Library type	Insert sizes	Reads (millions)	Coverage
Plasmid, high copy number	3 kb	3.4	4.6 x
Plasmid, low copy number	10 kb	2.7	3.6 x
Fosmids	40 kb	0.03	0.04 x
BACs	100 kb	0.1	0.16 x
total		6.23	8.4 x

Table S3. Overview of the anchoring of the assembly on the grapevine chromosomes.

Linkage group	Size (bp)	Number of supercontigs	Number of markers	N50 of supercontigs (bp)	Number of ultracontigs (number of supercontigs in ultracontigs)
LG1	15,630,816	5	24	7,473,787	1 (1)
LG1_random	5,496,190	4	4	1,567,357	0
LG2	17,603,400	15	15	1,359,084	2 (14)
LG2_random	60,809	1	1	60,809	0
LG3	10,186,927	5	15	6,238,017	1 (4)
LG3_random	1,343,266	2	2	867,932	0
LG4	19,293,076	14	21	3,066,225	3 (13)
LG5	23,428,299	10	26	2,071,933	4 (10)
LG6	24,148,918	10	26	5,371,753	3 (10)
LG7	15,233,747	11	17	3,189,795	2 (9)
LG7_random	176,143	1	1	176,143	0
LG8	21,557,227	10	26	2,700,301	3 (9)
LG8_random	12,125	1	1	12,125	0
LG9	16,532,244	6	20	2,980,855	2 (5)
LG10	9,647,040	6	15	2,296,208	3 (6)
LG10_random	2,206,354	8	5	437,620	1 (5)
LG11	13,936,303	5	12	5,465,665	1 (2)
LG11_random	1,958,407	2	2	1,210,238	0
LG12	18,540,817	10	17	2,817,145	4 (9)
LG12_random	2,826,407	2	1	1,464,313	1 (2)
LG13	15,191,948	9	19	2,542,976	3 (9)

LG13_random	1,580,403	2	2	932,749	0
LG14	19,480,434	6	29	4,315,032	1 (3)
LG14_random	5,432,426	4	3	3,690,152	0
LG15	7,693,613	2	11	4,849,857	0
LG15_random	4,297,576	2	2	2,711,818	0
LG16	8,158,851	3	10	5,958,581	0
LG16_random	4,524,411	9	7	1,275,354	3 (7)
LG17	13,059,092	5	14	5,345,817	1 (6)
LG17_random	1,763,011	2	1	1,567,215	1 (2)
LG18	19,691,255	5	22	12,675,388	2 (5)
LG18_random	5,949,186	5	7	1,429,425	1 (2)
LG19	14,071,813	5	20	7,851,008	2 (4)
LG19_random	1,912,523	4	3	1,160,223	1 (2)
Total	342,625,057	191	401	3,827,944 (non-random) 1,429,425 (random)	

Table S4. a. Distribution of the insert size of cDNA libraries. **b.** Sequencing overview of the full-length cDNA libraries.

Insert size	Number
1.8 - 2.4 kb	22
1.2 - 1.8 kb	528
0.9 - 1.2 kb	728
0.6 - 0.9 kb	193
0.3 - 0.6 kb	23
Full-inserts	1,494
Incomplete	262
Total	1,756

libraries	**clones**	**raw reads**	**useful reads (clones)**
A	19,504	20,194	18,819 (18,163)
B	12,832	13,748	13,640 (12,729)
C	15,524	15,898	15,164 (14,803)
D	379	742	734 (377)
Total	48,239	50,582	48,357 (46,072)

Table S5. Redundancy of cDNA libraries.

Useful 5' reads	1,785
Clusters	247
Singlets	648
Different cDNAs	895
Clones / cDNA	1.9

Table S6. Frequency of microsatellites in the grape genome.

Repeat type	Counts	Counts per Mbp	Average length[b]
Mono[a]	68,216	136.6	16.3
Di[a]	47,021	94.3	23.5
Tri[a]	42,018	84.2	17.7
Tetra[a]	54,899	110.1	14.1
Penta[a]	27,480	55.1	16.1
Total/mean	239,634	480.3	17.4

[a]Mono, mononucleotide repeats; Di, dinucleotide repeats; Tri, trinucleotide repeats; Tetra, tetranucleotide repeats; Penta, pentanucleotide repeats. [b]Average length is expressed in bp.

Table S7. Frequency of transposable elements in the grape genome.

Type	No. of occurrences	Coverage (kb)	Genome fraction (%)
Repeated sequences (ReAS derived)	n.d.	185,346.7	38.81
Transposable elements proteins (BlastX)	35,024	52,898.0	11.08
Class I	33,118	50,863.3	10.65
Non-LTR: LINEs	5,504	6792.9	1.42
LTR: Ty1/copia	17,293	24,640.8	5.16
LTR: Ty3/gypsy	9,632	17,659.6	3.70
Other LTR	88	103.6	0.02
Other class I	601	166.6	0.35
Class II	1,797	1,975.9	0.41
Helitrons	109	58.9	0.01
Manually annotated Transposable elements	111,876	83,404.7	17.47
Class I	105,532	81,363.7	17.04
Non-LTR: LINEs	15,216	12,131.1	2.54
LTR: Ty1/copia	56,890	39,848.3	8.35
LTR: Ty3/gypsy	14,093	15,339.8	3.21
Other LTR	18,688	13,191.5	2.76
Other class I	645	853.0	0.18
Class II	6,344	2,040.9	0.43
Helitrons	0	0.0	0.00

Table S8. Frequency of transposable elements in experimentally verified introns.

	Coverage (kb)	Intron fraction (%)
Manually annotated Transposable elements	1,793.7	12.37
Class I	1,770.7	12.21
Non-LTR: LINEs	1,175.4	8.10
LTR: Ty1/copia	506.0	3.49
LTR: Ty3/gypsy	37.5	0.26
Other LTR	34.7	0.24
Other class I	17.0	0.12
Class II	23.0	0.16
Helitrons	0.0	0.00

Table S9. Top 50 Interpro domains in *Vitis vinifera* genome.

InterPro domain	Proteins	InterPro family description
IPR011009	1,485	Protein kinase-like
IPR000719	1,470	Protein kinase
IPR001245	1,312	Tyrosine protein kinase
IPR002290	1,271	Serine/threonine protein kinase
IPR008271	955	Serine/threonine protein kinase, active site
IPR001611	908	Leucine-rich repeat
IPR002885	605	Pentatricopeptide repeat
IPR002182	504	NB-ARC
IPR008940	501	Protein prenyltransferase
IPR001128	440	Cytochrome P450
IPR009057	416	Homeodomain-like
IPR000767	403	Disease resistance protein
IPR003593	347	AAA ATPase
IPR009007	309	Peptidase aspartic, catalytic
IPR012287	288	Homeodomain-related
IPR013210	283	Leucine rich repeat, N-terminal
IPR002401	266	E-class P450, group I
IPR001005	264	Myb, DNA-binding
IPR001841	261	Zinc finger, RING-type
IPR001680	259	WD-40 repeat
IPR012336	245	Thioredoxin-like fold
IPR002213	240	UDP-glucuronosyl/UDP-glucosyltransferase
IPR011046	236	WD40-like
IPR012677	223	Nucleotide-binding, alpha-beta plait
IPR012335	222	Thioredoxin fold
IPR013781	219	Glycoside hydrolase, catalytic core
IPR000504	209	RNA-binding region RNP-1 (RNA recognition motif)
IPR005162	206	Retrotransposon gag protein
IPR003439	199	ABC transporter related
IPR002110	181	Ankyrin
IPR001480	180	Curculin-like (mannose-binding) lectin
IPR011989	178	Armadillo-like helical
IPR008972	173	Cupredoxin
IPR011990	171	Tetratricopeptide-like helical
IPR005123	160	2OG-Fe(II) oxygenase
IPR008930	160	Terpenoid cylases/protein prenyltransferase alpha-alpha toroid
IPR011050	153	Virulence factor, pectin lyase fold
IPR001810	151	Cyclin-like F-box
IPR000157	148	Toll-Interleukin receptor
IPR012334	148	Pectolytic enzyme, Pectin lyase fold
IPR002048	138	Calcium-binding EF-hand
IPR008949	138	Terpenoid synthase
IPR011992	134	EF-Hand type
IPR001650	131	Helicase, C-terminal
IPR011598	131	Helix-loop-helix DNA-binding
IPR008994	130	Nucleic acid-binding, OB-fold
IPR001878	129	Zinc finger, CCHC-type
IPR001471	128	Pathogenesis-related transcriptional factor and ERF
IPR011051	127	Cupin, RmlC-type
IPR014001	125	DEAD-like helicases, N-terminal

Table S10. Description of clusters of orthologous (or paralogous) genes obtained after applying SLCs.

Couple of species	Number of clusters	Number of orthologous or paralogous genes in clusters	Average (max) number of genes in clusters	Genomic coverage on *Vitis vinifera* (% of the anchored and oriented sequence)
Vitis vinifera – Populus trichocarpa	197	7,155	36.3 (210)	261Mb (88%)
Vitis vinifera – Arabidopsis thaliana	267	7,087	26.6 (97)	284Mb (96%)
Vitis vinifera – Oryza sativa	286	3,470	12.1 (40)	266Mb (89%)
Vitis vinifera – Vitis vinifera	146	2,948	20.2 (80)	211Mb (71%)

Table S11. Statistical description of the validity of paralogous (or orthologous) clusters.

Couple of species	Percentage of valid clusters (pvalue = 10^{-4})	Average of p-value	Min. p-value	Max. p-value
Vitis vinifera – Populus trichocarpa	100	$1.98E^{-16}$	0	$3.8E^{-14}$
Vitis vinifera – Arabidopsis thaliana	100	$6.22E^{-13}$	0	$9.67E^{-11}$
Vitis vinifera – Oryza sativa	100	$1.87E^{-8}$	0	$2.15E^{-6}$
Vitis vinifera – Vitis vinifera	94.5	$2.22E^{-5}$	0	$7.09E^{-4}$

Table S12. Orthologous regions between poplar, Arabidopsis or rice versus grape, and their relations with grape paralogous regions.

	Number of orthologous blocks	Number of orthologous blocks containing 2 or 3 paralogous grape regions	Number of orthologous blocks containing 3 paralogous grape regions
Vitis vinifera – Oryza sativa	507	288 (57%)	118 (23%)
Vitis vinifera – Arabidopsis thaliana	350	53 (15%)	5 (1.4%)
Vitis vinifera – Populus trichocarpa	184	12 (6.5%)	0 (0%)